Demystifying Classifier-Free Guidance for Auto-Regressive Image Generation

Zhiling Zhou [1*] **Jiachun Pan** [2*] **Fengzhuo Zhang** [1*†] **Dirk Bergemann** [1] **Zhuoran Yang** [1]

Abstract

Classifier-free guidance (CFG) has been widely adopted in autoregressive (AR) models for high-quality image generation. Despite its strong empirical performance, its mechanism in AR models remains unclear. This paper demystifies the mechanism of CFG through both empirical and theoretical studies. By examining the top-ranked semantics of different components in CFG, we show that texture information is primarily encoded in the difference between the conditional and unconditional logits, whereas both logits share previous-token repetition as their leading semantics. This strong repetition semantics obscures the desired texture information, causing greedy decoding from the conditional logits alone to produce nearly pure-color images. Through a training-dynamics analysis of shallow transformers, we prove that this shared repetition semantics does not arise from limitations or failures of pretraining, but instead originates from the texture sparsity of images. We further show that CFG improves generation quality by rectifying this repetition bias during inference. Motivated by the shared semantics between conditional and unconditional logits, we propose Attention Weight Reuse (AttnReuse), which reuses intermediate attention computations from the conditional-logit forward pass to accelerate the unconditional-logit computation. AttnReuse reduces about 25% of attention computation with little performance degradation across different models.

1. Introduction

Auto-Regressive (AR) models have recently emerged as a promising paradigm for unifying text and image generation (Wang et al., 2026b; Liu et al., 2024; Chen et al., 2025), offering a competitive alternative to diffusion models (Rombach et al., 2022; Esser et al., 2024). Analogous to text generation, they discretize images into visual tokens and cast image generation as next-token prediction. Recent advances have pushed the performance frontier of them through high-quality data curation (Xin et al., 2025), improved discrete latent representations (Tian et al., 2024), and flexible inference frameworks (Cui et al., 2025). During inference, most AR models adopt Classifier-Free Guidance (CFG), which improves visual-token prediction by extrapolating the conditional logits away from the unconditional logits. Empirically, CFG substantially improves generation quality and roughly doubles inference computation.

Despite the wide adoption of CFG, its mechanism in AR models remains poorly understood. Most existing studies analyze CFG in diffusion models, where the explanations rely heavily on properties of the reverse diffusion process. However, this process has no direct analogue in AR models, making existing understandings of CFG not directly applicable. This paper takes a first step toward understanding the mechanism underlying CFG for AR models. Specifically, we ask:

1. *Why is classifier-free guidance effective for AR image generation?*
2. *Given that AR models succeed in text generation without CFG, does the necessity of CFG in image generation rely on a limitation or failure of AR pretraining?*
3. *Is doubled inference computation necessary for the effectiveness of CFG in AR models?*

To answer the first question, we first conduct empirical studies to reveal the roles of different components in CFG. Through both qualitative and quantitative analyses, we find that image texture information is primarily captured by the difference between the conditional and unconditional logits, whereas the top-ranked decoded semantics of the conditional logits alone correspond to nearly pure-color images. To understand why this occurs, we compare the conditional and unconditional logits in both logit space and pixel space. The results show that their top-ranked decoded semantics are highly similar. Through causal interventions during

[*]Equal contribution [†]Project Lead. [1]Yale University [2]National University of Singapore. Correspondence to: Fengzhuo Zhang <fengzhuo.zhang@yale.edu>, Zhuoran Yang <zhuoran.yang@yale.edu>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

image generation, we further show that these shared top-ranked semantics mainly correspond to repeating previous tokens, explaining why decoding from the conditional logits alone tends to produce pure-color images. Overall, our experiments suggest that CFG amplifies the texture information encoded in the logit difference, thereby correcting the repetition bias present in the conditional logits.

To answer the second question, we theoretically study a shallow transformer with a single attention layer to explain why conditional and unconditional logits share repetition as their top-ranked semantics. Through training-dynamics analysis, we prove that this shared repetition semantics is not caused by insufficient or failed pretraining. Instead, it can arise from the sparsity of image textures, namely the prevalence of nearly pure-color blocks in training images. This sparsity makes the conditional and unconditional logits of well-pretrained models provably favor previous-token repetition, while CFG effectively rectifies this bias to generate the desired images.

To answer the third question, we reduce the computational burden of CFG by reusing intermediate states from the conditional-logit computation. Motivated by the shared semantics of conditional and unconditional logits, we propose Attention Weight Reuse (AttnReuse). The algorithm decomposes the attention output of conditional logits into weighted text-token values and weighted image-token values, and directly reuses the weighted image-token values as the attention output for unconditional-logit computation. It reduces about 25% of attention computation with little performance degradation across different models. In addition, the efficiency of AttnReuse generalizes to diffusion models.

Summarizing these results, we identify a clear mechanism underlying the effectiveness of CFG: CFG **rectifies the repetition bias** in the conditional logits by **amplifying texture information in the conditional–unconditional logit difference**. This bias arises from the texture sparsity of images, and can be reused to reduce the computation of CFG.

2. Preliminaries

Auto-regressive models provide a promising paradigm for jointly modeling images and text within a unified architecture (Liu et al., 2024; Chen et al., 2025). These AR models represent images as visual tokens in $\mathcal{X}_{\text{vis}}$ and text as textual tokens in $\mathcal{X}_{\text{txt}}$ (Wang et al., 2026b). Let $\mathcal{X} = \mathcal{X}_{\text{vis}} \cup \mathcal{X}_{\text{txt}}$ denote the token vocabulary, $\Delta(\mathcal{X})$ the set of distributions over $\mathcal{X}$, and $\mathcal{X}^*$ the set of all finite token sequences over $\mathcal{X}$. A model $P_\theta : \mathcal{X}^* \to \Delta(\mathcal{X})$ with parameters $\theta \in \Theta$ maps each token sequence to a distribution over the next token. It is trained by next-token prediction, where the total cross-entropy loss decomposes into a textual loss and a visual loss (Wang et al., 2026b; Chen et al., 2025). The textual loss is the same as in standard LLM training. For the visual loss, to enable Text-to-Image (T2I) generation, the model predicts the next visual token $x_{t+1} \in \mathcal{X}_{\text{vis}}$ conditioned on the text prompt $z_1, \ldots, z_{T_{\text{txt}}} \in \mathcal{X}_{\text{txt}}$ and the preceding visual tokens $x_1, \ldots, x_t \in \mathcal{X}_{\text{vis}}$, for $1 \le t < T_{\text{vis}}$. Formally,

$$\mathcal{L}_{\text{vis}}(\theta) = \mathbb{E}_{t, x_{1:t+1}}\big[-\log P_\theta(x_{t+1} | z_1, \cdots, z_{T_{\text{txt}}}, x_1, \cdots, x_t)\big] \tag{1}$$

where $T_{\text{txt}}, T_{\text{vis}} > 0$ denote the numbers of textual and visual tokens, $x_{1:t} \in \mathcal{X}^t$ denotes the prefix, and the expectation is taken over t sampled uniformly from the visual-token positions $1 \le t < T_{\text{vis}}$, and over the corresponding prefix $x_{1:t+1}$ drawn from the training-data distribution.

Classifier-Free Guidance is a decoding algorithm originally designed for diffusion models (Ho & Salimans, 2022), which is found to be also effective in the AR models. During the decoding, CFG enhances the logits $\log P_\theta$ learned in the pretraining by adding the difference between the conditional logit and the unconditional logit. Concretely, CFG samples the visual token at temperature $\tau \ge 0$ as follows

$$\begin{aligned} x_{t+1} &\sim \texttt{sm}\big(\tau^{-1}\text{logit}^{\lambda}(\cdot \,|\, z_{1:T_{\text{txt}}}, x_1, \cdots, x_t)\big), \\ \text{logit}^{\lambda} &= \text{logit}^{\text{cond}} + \lambda(\text{logit}^{\text{cond}} - \text{logit}^{\text{uncond}}). \end{aligned} \tag{2}$$

The conditional logit is the logit learned in the pretraining, i.e., $\text{logit}^{\text{cond}}(\cdot \,|\, z_{1:T_{\text{txt}}}, x_1, \cdots, x_t) = \log P_\theta(\cdot \,|\, z_{1:T_{\text{txt}}}, x_1, \cdots, x_t) \in \mathbb{R}^{|\mathcal{X}|}$. The unconditional logit is the logit with empty or negative prompts, i.e., $\text{logit}^{\text{uncond}}(\cdot \,|\, \tilde{z}_{1:T_{\text{txt}}}, x_1, \cdots, x_t) = \log P_\theta(\cdot \,|\, \tilde{z}_{1:T_{\text{txt}}}, x_1, \cdots, x_t)$, where $\tilde{z}_{1:T_{\text{txt}}}$ is an empty prompt, i.e., $\tilde{z}_i = \varnothing$, or a negative prompt, e.g., "Generate an image with low quality". For diffusion models, existing works explain the effectiveness of CFG from a sampling perspective by exploiting the reverse-time dynamics of the diffusion process (Wu et al., 2024; Bradley & Nakkiran, 2024). However, these explanations do not directly apply to AR models, where no analogous reverse-time diffusion equation is available. We refer readers to Appendix A for further related work.

Transformer is the backbone of modern AR sequence modeling (Vaswani et al., 2017). Its attention modules play a central role in capturing token dependencies by aggregating hidden states through data-dependent weights. Specifically, given the hidden states $H = [h_1, \ldots, h_T]^\top \in \mathbb{R}^{T \times D}$ for T tokens, the attention module is defined as $\texttt{attn}(H; W_Q, W_K, W_V) = \texttt{sm}(\texttt{MASK}(HW_Q^\top W_K H^\top))HW_V^\top \in \mathbb{R}^{T \times D}$, where $W_Q, W_K, W_V \in \mathbb{R}^{D \times D}$ are the query, key, and value weight matrices, $\texttt{sm}(\cdot)$ denotes the row-wise softmax, and $\texttt{MASK}(\cdot)$ denotes the causal mask operator, i.e., $\texttt{MASK}(M)_{t,l} = M_{t,l}$ if $l \le t$ and $\texttt{MASK}(M)_{t,l} = -\infty$ otherwise. Thus, each token can attend only to itself and previous tokens.

3. Empirical Understanding of Classifier-Free Guidance in AR Models

In this section, we empirically investigate why and how CFG contributes to the high-quality image generation in AR models. We conduct experiments on three representative models of different scales: LlamaGen (0.7B) (Sun et al., 2024), Lumina-mGPT 2.0 (7B) (Xin et al., 2025), and Emu3 (8B) (Wang et al., 2026b). Experimental details are provided in Appendix B.

3.1. Top-ranked Semantics of Components in Classifier-Free Guidance

We first conduct experiments to understand the top-ranked semantics of $\mathrm{logit}^{\mathrm{cond}}$ and $\mathrm{logit}^{\mathrm{cond}} - \mathrm{logit}^{\mathrm{uncond}}$ in Eqn. (2), i.e., the semantics induced by their highest-logit visual tokens. Given a text prompt $z_{1:T_{\mathrm{txt}}}$, we generate images by greedy decoding from $\mathrm{logit}^{\mathrm{cond}}$ and from $\mathrm{logit}^{\mathrm{cond}} - \mathrm{logit}^{\mathrm{uncond}}$, respectively, using temperature $\tau = 0$. The former corresponds to CFG with $\lambda = 0$, while the latter captures the limiting guidance direction as $\lambda \to \infty$. Since zero-temperature decoding always selects the highest-logit token, these experiments reveal the dominant semantics encoded by each logit. As a baseline, we also generate images using a standard finite guidance scale λ. Experimental details are provided in Appendix B, and representative examples are shown in Figure 1. The results show that the dominant behavior of $\mathrm{logit}^{\mathrm{cond}}$ is to generate images with nearly pure colors. In contrast, compared with images generated from logit^{λ}, images generated from $\mathrm{logit}^{\mathrm{cond}} - \mathrm{logit}^{\mathrm{uncond}}$ exhibit richer color variation, more complex foreground and background structures, and more detailed main objects. In summary, $\mathrm{logit}^{\mathrm{cond}} - \mathrm{logit}^{\mathrm{uncond}}$ emphasizes visual details and textures, but can also induce excessively complex appearances compared with standard CFG-generated images. To quantify this observation, we measure the texture complexity of an image token sequence $x_{1:T_{\mathrm{vis}}}$ as follows.

We first decode the visual token sequence $x_{1:T_{\mathrm{vis}}}$ into a pixel-space image, denoted by $I \in \{0, \ldots, 255\}^{C \times H \times W}$, where $C = 3$ in our experiments. We quantify texture complexity using three metrics: Pixel Entropy $\mathrm{Entropy}(I)$ (Shannon, 1948), Total Variation $\mathrm{TV}(I)$ (Rudin et al., 1992), and Orientation Entropy $\mathrm{OriEntropy}(I)$ (Redies et al., 2017), defined as $\mathrm{Entropy}(I) = C^{-1} \sum_{c=1}^{C} H(p_c)$, $\mathrm{TV}(I) = C^{-1} \sum_{c=1}^{C} \mathbb{E}_{i,j} \|\nabla_{i,j} I_{c,i,j}\|_1$, and $\mathrm{OriEntropy}(I) = C^{-1} \sum_{c=1}^{C} H(q_c)$, respectively. Here, $\nabla_{i,j} I_{c,i,j} = [I_{c,i+1,j} - I_{c,i,j},\ I_{c,i,j+1} - I_{c,i,j}]$ is the spatial gradient at pixel (i, j); $p_c(v) = (HW)^{-1} \sum_{h=1}^{H} \sum_{w=1}^{W} \mathbb{I}\{I_{c,h,w} = v\}$, for $0 \le v \le 255$, is the empirical intensity distribution; and $q_c(\theta) \propto \sum_{i,j} \|\nabla_{i,j} I_{c,i,j}\|_2 \cdot \mathbb{I}\{\mathrm{angle}(\nabla_{i,j} I_{c,i,j}) \in \theta\}$ is the magnitude-weighted orientation distribution over orientation bins, capturing structural regularity. Pixel entropy has been used to quantify the diversity of intensity distributions (Haralick et al., 2007; Yuan et al., 2025); Total Variation has been used to capture the overall magnitude of local spatial variations (Rudin et al., 1992; Chambolle et al., 2010); and edge-orientation entropy has been used to describe the diversity of edge directions (Redies et al., 2017; Grebenkina et al., 2018; McAdams et al., 2025). Figure 2 reports the average values of these metrics over 500 images generated from different logits. The results show that images generated from $\mathrm{logit}^{\mathrm{cond}} - \mathrm{logit}^{\mathrm{uncond}}$ exhibit the highest texture complexity for most models. In contrast, images generated from $\mathrm{logit}^{\mathrm{cond}}$ exhibit the lowest texture complexity, consistent with the qualitative results in Figure 1. Thus, we reach the following conclusion.

Finding 1: The top-ranked semantics of $\mathrm{logit}^{\mathrm{cond}}$ correspond to pure-color images, whereas those of $\mathrm{logit}^{\mathrm{cond}} - \mathrm{logit}^{\mathrm{uncond}}$ emphasize overly complex textures.

3.2. The Dominant Role of Unconditional Logits in Conditional Logits

Using the decomposition $\mathrm{logit}^{\mathrm{cond}} = \mathrm{logit}^{\mathrm{uncond}} + (\mathrm{logit}^{\mathrm{cond}} - \mathrm{logit}^{\mathrm{uncond}})$, Finding 1 suggests that the semantics encoded in $\mathrm{logit}^{\mathrm{cond}}$ can be approximately interpreted as

$$\mathrm{logit}^{\mathrm{cond}} \approx \mathrm{logit}^{\mathrm{uncond}} + \text{Complex Textures Info.} \quad (3)$$

However, zero-temperature decoding from $\mathrm{logit}^{\mathrm{cond}}$ produces images with nearly pure colors, as shown in Figure 1. This indicates that the top-ranked semantics of $\mathrm{logit}^{\mathrm{cond}}$ are dominated by $\mathrm{logit}^{\mathrm{uncond}}$ rather than by the complex texture information. In this section, we verify this interpretation by showing that, when the model is prefilled with images generated under a proper guidance scale λ^*, $\mathrm{logit}^{\mathrm{cond}}$ and $\mathrm{logit}^{\mathrm{uncond}}$ are aligned among high-logit entries. In other words, the top-ranked semantics of $\mathrm{logit}^{\mathrm{cond}}$ are largely shared with those of $\mathrm{logit}^{\mathrm{uncond}}$, so the unconditional logits explain the dominant semantics in conditional generation.

We measure the similarity between $\mathrm{logit}_t^{\mathrm{cond}}$ and $\mathrm{logit}_t^{\mathrm{uncond}}$ for $t \in [T_{\mathrm{vis}}]$ in logit and pixel spaces. For the logit-space comparison, let $\mathcal{K}_t^{\mathrm{cond}} = \mathrm{TopK}(\mathrm{logit}_t^{\mathrm{cond}}, k)$ and $\mathcal{K}_t^{\mathrm{uncond}} = \mathrm{TopK}(\mathrm{logit}_t^{\mathrm{uncond}}, k)$ denote the top-k token indices of the conditional and unconditional logits at position t, respectively, and define $\mathcal{U}_t = \mathcal{K}_t^{\mathrm{cond}} \cup \mathcal{K}_t^{\mathrm{uncond}}$. We measure their normalized ℓ_1 discrepancy on these high-logit coordinates by $\alpha^{\mathrm{diff}}((\mathrm{logit}_t^{\mathrm{cond}})_{t=1}^{T_{\mathrm{vis}}}, (\mathrm{logit}_t^{\mathrm{uncond}})_{t=1}^{T_{\mathrm{vis}}}) = T_{\mathrm{vis}}^{-1} \sum_{t=1}^{T_{\mathrm{vis}}} |\mathcal{U}_t|^{-1} \|\mathrm{logit}_{t,\mathcal{U}_t}^{\mathrm{cond}} - \mathrm{logit}_{t,\mathcal{U}_t}^{\mathrm{uncond}}\|_1$, where $\mathrm{logit}_{t,\mathcal{U}_t}$ denotes the restriction of logit_t to coordinates in $\mathcal{U}_t$. As a scale baseline, for any logit sequence $(\mathrm{logit}_t)_{t=1}^{T_{\mathrm{vis}}}$ with top-k set $\mathcal{K}_t$, we define $\alpha((\mathrm{logit}_t)_{t=1}^{T_{\mathrm{vis}}}) = (kT_{\mathrm{vis}})^{-1} \sum_{t=1}^{T_{\mathrm{vis}}} \|\mathrm{logit}_{t,\mathcal{K}_t}\|_1$. We denote

Figure 1. Generated images from $\text{logit}^{\text{cond}}$, logit^{λ}, and $\text{logit}^{\text{diff}} = \text{logit}^{\text{cond}} - \text{logit}^{\text{uncond}}$, shown from left to right in each grid. Prompts: **Left:** A glass bowl filled with oranges on a table. **Right:** An orange-reddish rose in a vase filled with water on top of a table.

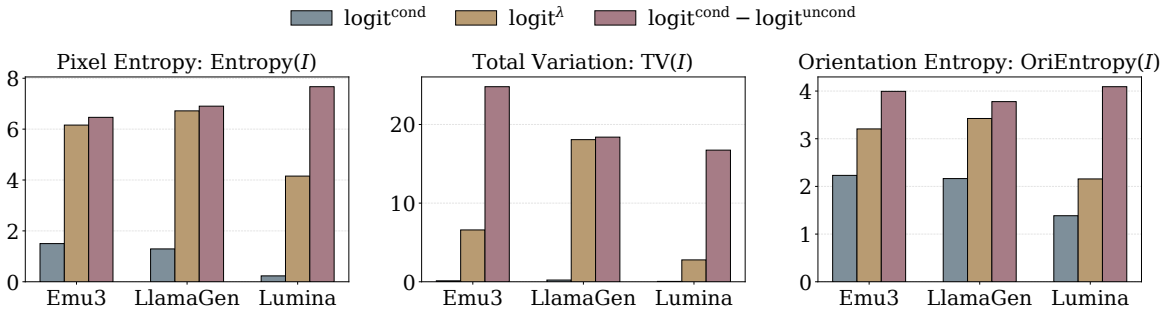

Figure 2. Texture complexity of images generated from $\text{logit}^{\text{cond}}$, logit^{λ}, and $\text{logit}^{\text{cond}} - \text{logit}^{\text{uncond}}$ under different metrics. Images generated from $\text{logit}^{\text{cond}} - \text{logit}^{\text{uncond}}$ exhibit the highest texture complexity, whereas images generated from $\text{logit}^{\text{cond}}$ exhibit the lowest.

the corresponding values for $\text{logit}^{\text{cond}}$ and $\text{logit}^{\text{uncond}}$ by α^{cond} and α^{uncond}, respectively.

Second, to measure the similarity between $(\text{logit}_t^{\text{cond}})_{t=1}^{T_{\text{vis}}}$ and $(\text{logit}_t^{\text{uncond}})_{t=1}^{T_{\text{vis}}}$ in pixel space, we first obtain token sequences $(x_t^{\text{cond}})_{t=1}^{T_{\text{vis}}}$ and $(x_t^{\text{uncond}})_{t=1}^{T_{\text{vis}}}$ by greedy decoding at each position. We then decode these sequences with the VQ decoder into normalized images $I^{\text{cond}}, I^{\text{uncond}} \in [0,1]^{C\times H\times W}$, and define their normalized ℓ_2 discrepancy as $\beta^{\text{diff}}(I^{\text{cond}}, I^{\text{uncond}}) = (CHW)^{-1/2}\|I^{\text{cond}} - I^{\text{uncond}}\|_2$, where $\|\cdot\|_2$ treats each image as a vector. As scale baselines, we define $\beta^{\text{cond}} = (CHW)^{-1/2}\|I^{\text{cond}}\|_2$ and $\beta^{\text{uncond}} = (CHW)^{-1/2}\|I^{\text{uncond}}\|_2$.

Figure 3 plots the average values of these metrics across all samples for different models. In both logit and pixel spaces, α^{diff} and β^{diff} are substantially smaller than the corresponding scale baselines, namely α^{cond} and α^{uncond} in logit space, and β^{cond} and β^{uncond} in pixel space. This indicates that the difference between $\text{logit}^{\text{cond}}$ and $\text{logit}^{\text{uncond}}$ is small relative to the magnitude of $\text{logit}^{\text{cond}}$ itself. In other words, most high-logit coordinates of $\text{logit}^{\text{cond}}$ are largely dominated by their unconditional counterparts in $\text{logit}^{\text{uncond}}$. This is consistent with our previous observation that zero-temperature decoding from $\text{logit}^{\text{cond}}$ produces images with nearly pure colors and few texture details.

Finding 2: The high-logit coordinates of the conditional logits are largely driven by the unconditional logits, indicating that texture-related information is weakly represented in the conditional logits.

3.3. Conditional and Unconditional Logits Both Tend to Repeat Previous Tokens

Findings 1 and 2 indicate that CFG is crucial for high-quality generation: texture-related information has limited influence on the top-ranked conditional predictions, while the unconditional logits dominate their high-logit coordinates. However, the training objective in Eqn. (1) suggests that the conditional logits should capture text-conditioned image content. This raises a natural question: why does the com-

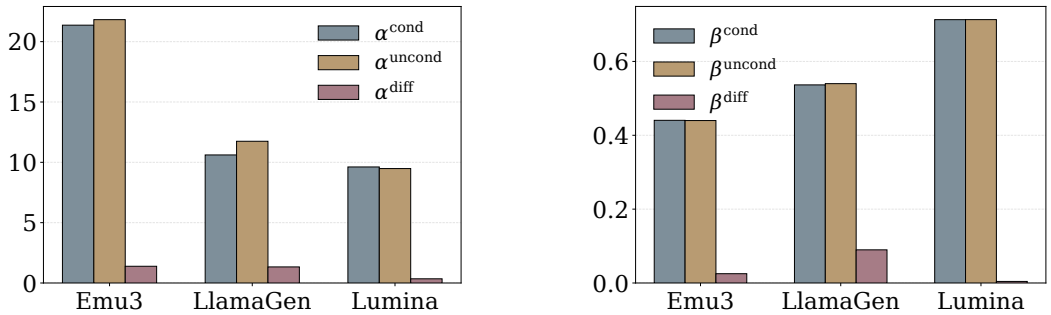

Figure 3. Similarity measurements between $(\text{logit}_t^{\text{cond}})_{t=1}^{T_{\text{vis}}}$ and $(\text{logit}_t^{\text{uncond}})_{t=1}^{T_{\text{vis}}}$ in logit space (**Left**) and pixel space (**Right**). Both results show that their discrepancy is small, indicating that the top-ranked semantics of $\text{logit}^{\text{cond}}$ are largely dominated by $\text{logit}^{\text{uncond}}$.

ponent shared by the conditional and unconditional logits give rise to pure-color images?

There are two possible explanations for why $\text{logit}^{\text{cond}}$ and $\text{logit}^{\text{uncond}}$ produce pure-color images under zero-temperature decoding. One is that the model assigns consistently high logits to a small set of special tokens, causing the same token to be selected at nearly all image-token positions. The other is that the model tends to repeat previously generated tokens, which would also yield nearly uniform images. We show that the second explanation is correct via causal intervention: conditioned on a prefix sequence, both $\text{logit}^{\text{cond}}$ and $\text{logit}^{\text{uncond}}$ favor tokens that repeat the prefix.

To verify the second explanation, we generate images from prescribed visual prefixes. Specifically, we fix a text condition $z_{1:T_{\text{txt}}}$ and a visual prefix $\hat{x}_{1:L}$, where $\hat{x}_{1:L}$ is a repeated sequence of a chosen visual token and L is the prefix length. We then continue generation from the conditional logits using greedy decoding: $x_{t+1}^{\text{cond}} = \arg\max_{x \in \mathcal{X}_{\text{vis}}} P_\theta(x \mid z_{1:T_{\text{txt}}}, \hat{x}_{1:L}, x_{L+1:t}^{\text{cond}})$, for $t = L, \ldots, T_{\text{vis}} - 1$. The unconditional generation is performed analogously, with the text condition replaced by an empty prompt $\tilde{z}_{1:T_{\text{txt}}}$. This yields two image-token sequences: $(\hat{x}_{1:L}, x_{L+1:T_{\text{vis}}}^{\text{cond}})$ and $(\hat{x}_{1:L}, x_{L+1:T_{\text{vis}}}^{\text{uncond}})$.

To quantify repetition behavior, we measure how often the generated tokens $x_{L+1:T_{\text{vis}}}$ repeat the pre-assigned prefix token. Specifically, let $\hat{x}_{1:L} = (s, \ldots, s)$ be a visual prefix consisting of a single token $s \in \mathcal{X}_{\text{vis}}$ repeated for the first L positions. For each generated position $L < t \leq T_{\text{vis}}$, we define $r_t^{\text{rep}}((\hat{x}_{1:L}, x_{L+1:T_{\text{vis}}})) = \mathbb{I}\{x_t = s\}$, where $\mathbb{I}\{\cdot\}$ is the indicator function. The repetition rate of the whole generated sequence is then $r^{\text{rep}}((\hat{x}_{1:L}, x_{L+1:T_{\text{vis}}})) = (T_{\text{vis}} - L)^{-1} \sum_{t=L+1}^{T_{\text{vis}}} r_t^{\text{rep}}((\hat{x}_{1:L}, x_{L+1:T_{\text{vis}}}))$.

Table 1. The average repeatition ratio r^{rep} after prefix tokens.

Model	$\text{logit}^{\text{uncond}}$	$\text{logit}^{\text{cond}}$
Emu3	0.5168	0.5479
LlamaGen	0.7457	0.7517
Lumina-mGPT	0.6804	0.7333

In our experiments, we set the prefix length L to the first 20% of visual-token positions and sample s uniformly from the set of common visual tokens. Experimental details are provided in Appendix B.4. Figure 4 reports the average r_t^{rep} for Lumina-mGPT across images. The results show that tokens generated from both $\text{logit}^{\text{cond}}$ and $\text{logit}^{\text{uncond}}$ largely repeat the prescribed token s. Specifically, among the first generated 10% of visual tokens, the repetition rate r_t^{rep} is around 0.8, indicating strong repetition behavior. Results for LlamaGen and Emu3 are deferred to Appendix B.4 and show the same trend. We also report the sequence-average repetition rate r^{rep} in Table 1, which shows that at least 50% of generated tokens repeat the assigned token on average. This empirical evidence supports that both $\text{logit}^{\text{cond}}$ and $\text{logit}^{\text{uncond}}$ strongly favor tokens that repeat the existing prefix, validating the hypothesis that local repetition is the primary driver of uniform, pure-color images.

Finding 3: The top-ranked semantics of $\text{logit}^{\text{cond}}$ and $\text{logit}^{\text{uncond}}$ are dominated by the repetition of previous tokens, causing zero-temperature decoding from these logits to generate pure-color images.

4. Theoretical Understanding of CFG in Shallow Transformers

Section 3 empirically shows that CFG is important for high-quality generation: texture-related information has limited influence on the top-ranked predictions of $\text{logit}^{\text{cond}}$, while the top-ranked semantics of $\text{logit}^{\text{cond}}$ are dominated by token repetition. However, these observations do not explain why the training objective in Eqn. (1) leads to conditional logits with these properties. In this section, we demystify this phenomenon by studying the training dynamics of shallow transformers on simplified data models.

We simplify the data model by partitioning visual tokens into visual feature tokens and visual background tokens: $\mathcal{X}_{\text{vis}} = \mathcal{X}_{\text{vis}}^f \cup \mathcal{X}_{\text{vis}}^b$. The feature-token set is $\mathcal{X}_{\text{vis}}^f = \{1, \ldots, K\}$, where each $k \in [K]$ corresponds to a unique texture and semantic concept. Each image contains a single

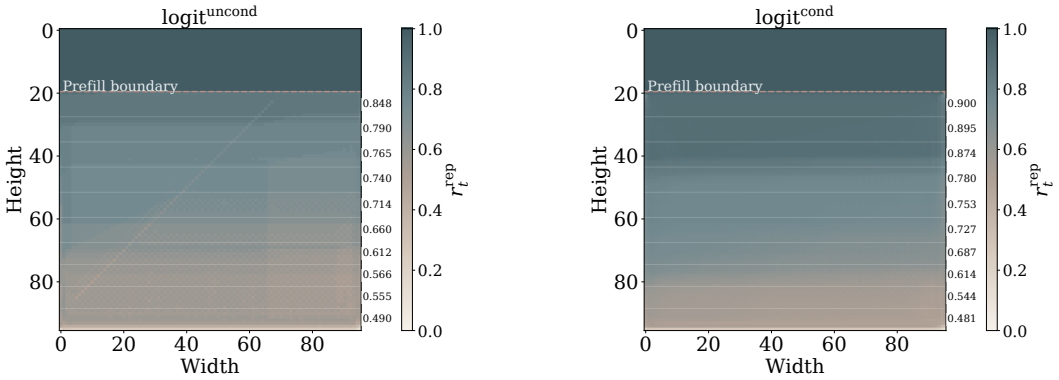

Figure 4. Position-wise repetition ratio r_t^{rep} of Lumina-mGPT, averaged over samples. Both $\text{logit}^{\text{cond}}$ and $\text{logit}^{\text{uncond}}$ show strong repetition of the prescribed prefix token.

feature-token type, with token k selected with probability p_k and $\sum_{k=1}^{K} p_k = 1$. The background-token set is $\mathcal{X}_{\text{vis}}^{b} = \{K+1, \ldots, K+J\}$, where each $K+j$ represents one background type. Each image contains a single background-token type, with token $K+j$ selected with probability q_j and $\sum_{j=1}^{J} q_j = 1$. Finally, the text-token set $\mathcal{X}_{\text{txt}} = \{K+J+1, \ldots, 2K+J\}$ represents the semantics of feature tokens, and we assume without loss of generality that feature token k is paired with text token $K+J+k$.

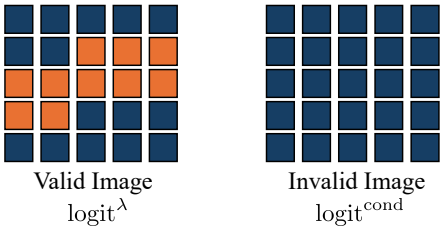

Figure 5. **Left:** Image $I_{k,j}$ for $k =$Orange, $j =$Blue with $\tau_k = 8$, $L = 10$, and $T_{\text{vis}} = 25$. **Right:** The invalid image generated by only a conditional model, which will simply repeat the background.

An image corresponding to text token $K+J+k$ and background token $K+j$, for $k \in [K]$ and $j \in [J]$, is represented as a visual-token sequence $I_{k,j} = (x_1, \ldots, x_{T_{\text{vis}}}) \in \mathcal{X}_{\text{vis}}^{T_{\text{vis}}}$ of length T_{vis}. In $I_{k,j}$, the tokens in the feature block $\mathcal{T}_k = \{\tau_k, \ldots, \tau_k + L - 1\}$ are all equal to the feature token k, while all other tokens are equal to the background token $K+j$. Here, $\tau_k \geq 3$ denotes the starting position of the feature block for the feature k, and all features share the same feature-block length L. See Figure 5 for an example.

We assume that the length of text token sequence is 1, i.e., $T_{\text{txt}} = 1$. Then we define the pretraining data distribution P_{data} over $(z_1, x_{1:T_{\text{vis}}})$ as follows. A feature class $k \in [K]$ and a background class $j \in [J]$ are drawn independently according to distributions $(p_k)_{k=1}^{K}$ and $(q_j)_{j=1}^{J}$, respectively. Given (k, j), we set $z_1 = K+J+k \in \mathcal{X}_{\text{txt}}$, and $x_{1:T_{\text{vis}}}$ are the tokens of $I_{k,j}$. Further, we assume the feature tokens are sparse in the whole image.

Assumption 1 (Texture Sparsity)**.** *The feature block satisfies $L = \Theta(\log T_{vis})$ as T_{vis} increases .*

This assumption states that texture occupies only a small portion of the image. This is intuitive, since texture often appears as localized high-frequency patches. To verify this empirically, we extract texture and flat regions from the ImageNet validation set (Deng et al., 2009) using a single-level 2D wavelet decomposition (Mallat, 1999). The approximation coefficients capture smooth structures, while the horizontal, vertical, and diagonal detail coefficients encode high-frequency texture and edges (Arivazhagan & Ganesan, 2003; Lu et al., 1997). We aggregate the detail components into a high-frequency response map, compute its magnitude, and apply Otsu thresholding (Otsu, 1975) to identify texture pixels. The texture-area ratio is the fraction of pixels above the threshold, and the flat-area ratio is its complement. As shown in Figure 6, non-texture regions occupy most of the image area on average, supporting our assumption.

We then introduce the shallow transformer used to learn the image distribution. The token x_t at position t is encoded as $h_t = [f_{x_t}^\top, e_t^\top]^\top \in \mathbb{R}^D$, where $f_{x_t} \in \mathbb{R}^d$ is the token embedding, $e_t \in \mathbb{R}^{d_e}$ is the positional embedding, and $D = d + d_e$. The input to the transformer is the concatenation of these embeddings, $H = [h_0, h_1, \ldots, h_t]^\top \in \mathbb{R}^{(t+1)\times D}$, where h_0 denotes the embedding of the text token. For ease of analysis, we adopt the standard orthonormality assumption on embeddings (Allen-Zhu & Li, 2020; Huang et al., 2024; Nichani et al.).

Assumption 2 (Orthonormality of Embeddings)**.** *The token embeddings and positional embeddings are orthonormal, i.e., $f_i^\top f_{i'} = \mathbb{I}\{i = i'\}$, and $e_t^\top e_{t'} = \mathbb{I}\{t = t'\}$.*

We consider a single-layer transformer. Given input $H \in \mathbb{R}^{(t+1)\times D}$, the parameter-merged attention operator is defined as $\texttt{attn}(H; \widetilde{W}_{QK}, \widetilde{W}_V) = \texttt{sm}(\texttt{MASK}(H\widetilde{W}_{QK}H^\top))H\widetilde{W}_V^\top \in \mathbb{R}^{(t+1)\times D}$, where $\widetilde{W}_{QK}, \widetilde{W}_V \in \mathbb{R}^{D\times D}$ are the merged query-key parameter and value parameter, respectively. They are specified

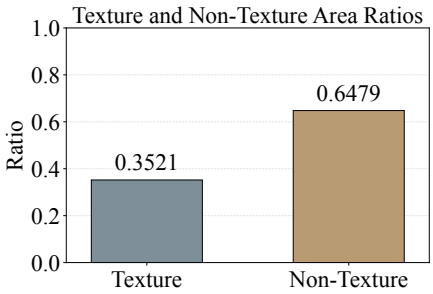

Figure 6. **Left:** Visualization of the texture region (white) in the images. **Right:** The texture and non-texture ratios averaged on ImageNet Validation Set.

as

$$\widetilde{W}_{QK} = \begin{bmatrix} 0 & 0 \\ W_{QK} & 0 \end{bmatrix}, \quad \widetilde{W}_V = \begin{bmatrix} W_V & 0 \\ 0 & 0 \end{bmatrix},$$

where $W_{QK} \in \mathbb{R}^{d_e \times d}$ captures the correlation between token embedding and positions, while $W_V \in \mathbb{R}^{d \times d}$ captures the correlation among token embeddings. The attention output is then mapped to logits by the Language-Model (LM) head: $\mathrm{logit} = \mathtt{attn}(H; \widetilde{W}_{QK}, \widetilde{W}_V)\widetilde{W}_{\mathrm{lm}} \in \mathbb{R}^{(t+1)\times(K+J)}$. Here the parameter of the LM head $\widetilde{W}_{\mathrm{lm}} \in \mathbb{R}^{D\times(K+J)}$ is fixed to $\widetilde{W}_{\mathrm{lm}} = [W_{\mathrm{lm}}^\top, 0]^\top$, where $W_{\mathrm{lm}} = [f_1, \ldots, f_{K+J}] \in \mathbb{R}^{d\times(K+J)}$ ties the visual token embeddings to the LM head. Let logit_t denote the logit of the t-th visual token, i.e., the $(t+1)$-th row of logit. The next-token distribution is $P_\theta(x_{t+1} \mid z_1, x_1, \cdots, x_t) = \mathtt{sm}(\mathrm{logit}_t)$. We train $\theta = (W_{QK}, W_V)$ using the cross-entropy loss $\mathcal{L}_{\mathrm{vis}}(\cdot)$ in Eqn. (1), under the population distribution P_{data} over $(z_1, x_{1:T_{\mathrm{vis}}})$. Following prior theoretical analyses of transformer training dynamics (Nichani et al.; Wen et al., 2024), we initialize W_{QK} and W_V at zero and use a two-stage protocol: first updating W_V with W_{QK} fixed, and then updating W_{QK} with W_V fixed. Details are deferred to Appendix C. We denote the resulting pretrained parameter by $\hat{\theta}$.

Given the pretrained model $\hat{\theta}$, we generate the image-token sequence $\hat{I}^\lambda_{k,j}(\hat{\theta}) = (\hat{x}_1, \ldots, \hat{x}_T) \in \mathcal{X}_{\mathrm{vis}}^{T_{\mathrm{vis}}}$ corresponding to text token $K+J+k$ and background token $K+j$, for $k \in [K]$ and $j \in [J]$, as follows. We first fix the text token and the first visual token as $z_1 = K+J+k$ and $\hat{x}_1 = K+j$, respectively. We then generate the remaining visual tokens by greedy decoding: $\hat{x}_{t+1} = \arg\max_{x\in\mathcal{X}_{\mathrm{vis}}} \mathrm{logit}^\lambda_t(x \mid z_1, \hat{x}_{1:t})$, for $t = 1, \ldots, T_{\mathrm{vis}} - 1$, where logit^λ_t is induced by $\hat{\theta}$ through Eqn. (2). We next show that CFG is essential for shallow transformers to recover the desired image-token sequence, i.e., $\hat{I}^\lambda_{k,j}(\hat{\theta}) = I_{k,j}$. To highlight the main intuition, we first present an informal version of the result and defer the formal statement to Appendix E. Throughout this section, we write $a = \mathsf{poly}(b)$ to indicate that a is asymptotically upper bounded by a polynomial in b.

Theorem 1 (Informal Statement)**.** *Consider the asymptotic regime $T_{vis} \to \infty$ where $\tau_k = \tau = \mathsf{poly}(T_{vis})$ for all $k \in [K]$, $J = \mathsf{poly}(T_{vis})$, $K = \Theta(1)$, and $p_k = \Theta(1)$, $q_j = 1/J$. With Assumptions 1 and 2 and suitable learning rates, the pretrained model $\hat{\theta}$ satisfies the following properties.*

1. *(Minimal Training loss.) The population loss is small, i.e., $\mathcal{L}_{vis}(\hat{\theta}) = O(T_{vis}^{-1}(\log T_{vis})^2)$.*
2. *(Alignment of $\mathrm{logit}^{\mathrm{cond}}$ and $\mathrm{logit}^{\mathrm{uncond}}$.) When prefilled with any target image $I_{k,j}$, the conditional and unconditional models make identical predictions at every visual position, i.e., $\arg\max_{x\in\mathcal{X}_{vis}} \mathrm{logit}^{\mathrm{cond}}_{t,x} = \arg\max_{x\in\mathcal{X}_{vis}} \mathrm{logit}^{\mathrm{uncond}}_{t,x}$, for all $t \in [T_{vis}]$, $k \in [K]$, $j \in [J]$.*
3. *(Conditional generation collapses to repeated background tokens.) With guidance scale $\lambda = 0$ (pure conditional generation), the model does not generate a valid image: for every $k \in [K]$ and $j \in [J]$, $\hat{I}^0_{k,j}(\hat{\theta}) = (K+j, \cdots, K+j) \neq I_{k,j}$.*
4. *(CFG recovers valid images.) There exists a guidance scale $\lambda^* = \Theta(1)$ such that generation with logits $\mathrm{logit}^{\lambda^*}$ is exact: $\hat{I}^{\lambda^*}_{k,j}(\hat{\theta}) = I_{k,j}$, for all $k \in [K], j \in [J]$.*

The small population loss shows that the transformer is well trained, so the phenomena below are not artifacts of failed optimization. The alignment between $\mathrm{logit}^{\mathrm{cond}}$ and $\mathrm{logit}^{\mathrm{uncond}}$ theoretically justifies Finding 2: the top-ranked semantics of the conditional and unconditional logits coincide. Consequently, generation with $\mathrm{logit}^{\mathrm{cond}}$ collapses to repeated background tokens, producing a pure-color image, which is consistent with Findings 1 and 3. Finally, CFG with a constant guidance scale λ^* extracts the correct signal from the logit difference $\mathrm{logit}^{\mathrm{cond}} - \mathrm{logit}^{\mathrm{uncond}}$ and recovers the target image. Therefore, all the empirical findings in Section 3 provably arise in shallow transformers.

5. Efficiency Improvement Via Attention Reuse

Building on the insights from Sections 3 and 4, we optimize AR inference efficiency by analyzing attention decomposition. Since Section 4 shows that a single-layer Transformer can recover CFG dynamics, we decompose the attention

Type	Model	Global	Entity	Attribute	Relation	Other	**Overall**↑	Attn. FLOPs↓
Diffusion	SD3-Medium	87.9	91.0	88.8	80.7	88.7	84.1	100%
	+ AttnReuse	90.1	86.5	88.8	86.4	85.7	82.5	**87%**
AR ($\tau = 0$)	Emu3	87.5	83.5	84.9	85.0	85.3	76.7	100%
	+ AttnReuse	84.3	83.5	83.4	83.5	82.9	74.7	**75%**
	Janus-Pro 7B	79.3	90.2	89.0	89.7	84.5	83.3	100%
	+ AttnReuse	79.2	87.5	87.6	89.3	88.6	82.3	**75%**
	Lumina-GPT 2.0	69.9	65.9	71.7	69.4	70.9	52.3	100%
	+ AttnReuse	71.8	68.7	70.8	69.9	64.7	52.2	**75%**
AR ($\tau = 1$)	Emu3	85.2	86.6	86.8	90.2	83.1	80.6	100%
	+ AttnReuse	84.5	86.5	86.3	89.7	84.2	80.4	**75%**
	Janus-Pro 7B	86.9	88.9	89.4	89.3	89.4	84.1	100%
	+ AttnReuse	84.9	88.7	87.1	90.1	88.9	82.0	**75%**
	Lumina-GPT 2.0	88.3	88.9	88.0	91.7	88.5	84.3	100%
	+ AttnReuse	88.4	88.3	87.4	90.8	84.8	83.7	**75%**

Table 2. **DPG-Bench results.** Comparison of model backbones with and without AttnReuse (gray rows). Attention FLOPs are normalized to the standard CFG baseline (100%), encompassing both conditional and unconditional computations. To balance efficiency and fidelity, AttnReuse is applied to middle layers (e.g., 16/32 for AR models), preserving early layers for low-level representations and late layers for final predictions. For SD3-Medium, reuse is further restricted to the final 90% of denoising timesteps to maintain global layout. This strategy reduces relative attention FLOPs to 75% for AR models and 87% for SD3-Medium while maintaining competitive performance.

output for a query q_t into visual and textual contributions. Let $k_{1:t}$ and $v_{1:t}$ be the keys and values for text tokens $(1 \dots T_{\text{txt}})$ and visual tokens $(T_{\text{txt}} + 1 \dots t)$. The attention output is:

$$\frac{\sum_{l=1}^{t} \exp(q_t^\top k_l) v_l}{\sum_{l=1}^{t} \exp(q_t^\top k_l)} = \gamma_t \cdot \frac{\sum_{l=T_{\text{txt}}+1}^{t} \exp(q_t^\top k_l) v_l}{\sum_{l=T_{\text{txt}}+1}^{t} \exp(q_t^\top k_l)} + (1-\gamma_t) \cdot \frac{\sum_{l=1}^{T_{\text{txt}}} \exp(q_t^\top k_l) v_l}{\sum_{l=1}^{T_{\text{txt}}} \exp(q_t^\top k_l)},$$

where $\gamma_t = \sum_{l=T_{\text{txt}}+1}^{t} \exp(q_t^\top k_l) / \sum_{l=1}^{t} \exp(q_t^\top k_l)$ denotes the fraction of attention mass assigned to visual tokens. This separation clarifies the CFG mechanism: the first visual term captures the trajectory shared by both conditional and unconditional branches—consistent with our Section 3 observation that both branches tend to repeat previous tokens. In contrast, the second textual term drives the deviation from the unconditional prediction, effectively forming the guidance direction $\text{logit}^{\text{cond}} - \text{logit}^{\text{uncond}}$.

Motivated by this observation, we propose *Attention Weight Reuse* (AttnReuse), with the full procedure given in Algorithm 1 in Appendix B.5. Instead of performing a separate unconditional forward pass, AttnReuse reuses the precomputed conditional attention weights by masking out text-token entries and renormalizing each row over visual-token keys. The resulting distribution efficiently approximates the unconditional attention weights. In practice, we apply AttnReuse to about half of the layers for AR models, with implementation details deferred to Appendix B.5.

We evaluate AttnReuse on three representative AR image generation models: Emu3 (Wang et al., 2026b), Lumina-GPT 2.0 (Xin et al., 2025), and Janus-Pro (Chen et al., 2025). To test its generality beyond AR models, we also apply it to the diffusion-based Stable Diffusion 3 (SD3) (Esser et al., 2024). We use **DPG-Bench** (Hu et al., 2024), which contains over 1,000 dense and structurally complex prompts for evaluating fine-grained semantic alignment. Results are reported under $\tau = 0$ and $\tau = 1$, corresponding to greedy decoding and stochastic sampling, respectively. The $\tau = 0$ setting isolates the model's top-ranked semantics, while $\tau = 1$ better reflects practical generation with sampling diversity. Detailed evaluation settings are shown in Appendix Table 3.

Quantitative results are summarized in Table 2, with qualitative examples shown in Appendix Fig. 12. Across different AR models, AttnReuse provides a high-fidelity proxy for the standard unconditional pass: it preserves structural faithfulness with only marginal performance degradation, while reducing attention-related FLOPs by approximately 25%. Consistent with these findings, qualitative results show no observable degradation across all evaluated models.

6. Conclusion

Our work demystifies the mechanism underlying CFG in AR image generation. We show that texture information is primarily encoded in the difference between the conditional and unconditional logits, whereas both logits share previous-token repetition as their top-ranked semantics. We prove that this repetition semantics arises from texture sparsity rather than training failure, and that CFG improves generation quality by rectifying this bias during inference. Motivated by this shared semantics, we propose AttnReuse, which reduces about 25% of attention computation with little performance degradation. A limitation of our work is its focus on image generation; we leave extensions to audio and video generation for future work.

References

Allen-Zhu, Z. and Li, Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

Arivazhagan, S. and Ganesan, L. Texture classification using wavelet transform. *Pattern recognition letters*, 24 (9-10):1513–1521, 2003.

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International conference on machine learning*, pp. 322–332. PMLR, 2019a.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. R., and Wang, R. On exact computation with an infinitely wide neural net. *Advances in neural information processing systems*, 32, 2019b.

Bradley, A. and Nakkiran, P. Classifier-free guidance is a predictor-corrector. *arXiv preprint arXiv:2408.09000*, 2024.

Cai, J.-F., Liu, H., Su, Z., and Wang, C. Improving classifier-free guidance of flow matching via manifold projection. *arXiv preprint arXiv:2601.21892*, 2026.

Cao, Y., Chen, Z., Belkin, M., and Gu, Q. Benign overfitting in two-layer convolutional neural networks. *Advances in neural information processing systems*, 35:25237–25250, 2022.

Chambolle, A., Caselles, V., Cremers, D., Novaga, M., Pock, T., et al. An introduction to total variation for image analysis. *Theoretical foundations and numerical methods for sparse recovery*, 9(263-340):227, 2010.

Chen, X., Mishra, N., Rohaninejad, M., and Abbeel, P. Pixelsnail: An improved autoregressive generative model. In *International conference on machine learning*, pp. 864–872. PMLR, 2018.

Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., and Ruan, C. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Chidambaram, M., Gatmiry, K., Chen, S., Lee, H., and Lu, J. What does guidance do? a fine-grained analysis in a simple setting. *Advances in Neural Information Processing Systems*, 37:84968–85005, 2024.

Chung, H., Kim, J., Park, G. Y., Nam, H., and Ye, J. C. Cfg++: Manifold-constrained classifier free guidance for diffusion models. *arXiv preprint arXiv:2406.08070*, 2024.

Cui, Y., Chen, H., Deng, H., Huang, X., Li, X., Liu, J., Liu, Y., Luo, Z., Wang, J., Wang, W., et al. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., and Tang, J. Cogview: Mastering text-to-image generation via transformers. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 19822–19835. Curran Associates, Inc., 2021.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pp. 1675–1685. PMLR, 2019.

Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12873–12883, June 2021.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Fu, H., Yang, Z., Wang, M., and Chen, M. Unveil conditional diffusion models with classifier-free guidance: A sharp statistical theory. *arXiv preprint arXiv:2403.11968*, 2024.

Galashov, A., Pokle, A., Doucet, A., Gretton, A., Delbracio, M., and De Bortoli, V. Learn to guide your diffusion model. *arXiv preprint arXiv:2510.00815*, 2025.

Gao, J., Zheng, T., Zou, J., Yang, F., Liu, S., Fan, L., Zhang, Z., Zhang, H., Chen, J., Jiang, P.-T., et al. C2fg: Control classifier-free guidance via score discrepancy analysis. *arXiv preprint arXiv:2603.08155*, 2026.

Ghosh, D., Hajishirzi, H., and Schmidt, L. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023.

Golikov, E., Pokonechnyy, E., and Korviakov, V. Neural tangent kernel: A survey. *arXiv preprint arXiv:2208.13614*, 2022.

Grebenkina, M., Brachmann, A., Bertamini, M., Kaduhm, A., and Redies, C. Edge-orientation entropy predicts preference for diverse types of man-made images. *Frontiers in Neuroscience*, 12:678, 2018.

Haralick, R. M., Shanmugam, K., and Dinstein, I. H. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 2007.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Hu, X., Wang, R., Fang, Y., Fu, B., Cheng, P., and Yu, G. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Huang, Y., Wen, Z., Chi, Y., and Liang, Y. A theoretical analysis of self-supervised learning for vision transformers. *arXiv preprint arXiv:2403.02233*, 2024.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Jelassi, S., Sander, M., and Li, Y. Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems*, 35:37822–37836, 2022.

Jiang, K. and Huang, J. A survey on vision autoregressive model. *arXiv preprint arXiv:2411.08666*, 2024.

Jiao, Y., Chen, Y., and Li, G. Towards a unified framework for guided diffusion models. *arXiv preprint arXiv:2512.04985*, 2025.

Jin, C., Shi, Q., and Gu, Y. Stage-wise dynamics of classifier-free guidance in diffusion models. *arXiv preprint arXiv:2509.22007*, 2025.

Karras, T., Aittala, M., Kynkäänniemi, T., Lehtinen, J., Aila, T., and Laine, S. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2024.

Kynkäänniemi, T., Aittala, M., Karras, T., Laine, S., Aila, T., and Lehtinen, J. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *Advances in Neural Information Processing Systems*, 37:122458–122483, 2024.

Lehman Pavasovic, K., Verbeek, J., Biroli, G., and Mezard, M. Understanding classifier-free guidance: High-dimensional theory and non-linear generalizations. *arXiv e-prints*, pp. arXiv–2502, 2025.

Li, G. and Jiao, Y. Provable efficiency of guidance in diffusion models for general data distribution. *arXiv preprint arXiv:2505.01382*, 2025.

Li, X., Wang, R., and Qu, Q. Towards understanding the mechanisms of classifier-free guidance. *arXiv preprint arXiv:2505.19210*, 2025.

Liu, D., Zhao, S., Zhuo, L., Lin, W., Xin, Y., Li, X., Qin, Q., Qiao, Y., Li, H., and Gao, P. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*, 2024.

Lu, C. S., Chung, P. C., and Chen, C. F. Unsupervised texture segmentation via wavelet transform. *Pattern recognition*, 30(5):729–742, 1997.

Lu, Y., Ma, C., Lu, Y., Lu, J., and Ying, L. A mean field analysis of deep resnet and beyond: Towards provably optimization via overparameterization from depth. In *International Conference on Machine Learning*, pp. 6426–6436. PMLR, 2020.

Ma, X., Zhou, M., Liang, T., Bai, Y., Zhao, T., Chen, H., and Jin, Y. Star: Scale-wise text-to-image generation via auto-regressive representations. *arXiv preprint arXiv:2406.10797*, 3, 2024.

Malarz, D., Kasymov, A., Zięba, M., Tabor, J., and Spurek, P. Classifier-free guidance with adaptive scaling. 2025.

Mallat, S. *A wavelet tour of signal processing*. Elsevier, 1999.

McAdams, P., Svobodova, S., Newman, T.-J., Terry, K., Mather, G., Skelton, A. E., and Franklin, A. The edge orientation entropy of natural scenes is associated with infant visual preferences and adult aesthetic judgements. *Plos one*, 20(2):e0316555, 2025.

Mei, S., Misiakiewicz, T., and Montanari, A. Learning with invariances in random features and kernel models. In *Conference on Learning Theory*, pp. 3351–3418. PMLR, 2021.

Nichani, E., Damian, A., and Lee, J. D. How transformers learn causal structure with gradient descent. In *Forty-first International Conference on Machine Learning*.

Otsu, N. A threshold selection method from gray-level histograms. *Automatica*, 11:285–296, 1975.

Papalampidi, P., Wiles, O., Ktena, I., Shtedritski, A., Bugliarello, E., Kajic, I., Albuquerque, I., and Nematzadeh, A. Dynamic classifier-free diffusion guidance via online feedback. *arXiv preprint arXiv:2509.16131*, 2025.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., and Tran, D. Image transformer. In *International conference on machine learning*, pp. 4055–4064. PMLR, 2018.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8821–8831. PMLR, 18–24 Jul 2021.

Redies, C., Brachmann, A., and Wagemans, J. High entropy of edge orientations characterizes visual artworks from diverse cultural backgrounds. *Vision Research*, 133:130–144, 2017.

Reed, S., van den Oord, A., Kalchbrenner, N., Bapst, V., Botvinick, M., and De Freitas, N. Generating interpretable images with controllable structure. 2016.

Reed, S., Oord, A., Kalchbrenner, N., Colmenarejo, S. G., Wang, Z., Chen, Y., Belov, D., and Freitas, N. Parallel multiscale autoregressive density estimation. In *International conference on machine learning*, pp. 2912–2921. PMLR, 2017.

Rojas, K., He, Y., Lai, C.-H., Takida, Y., Mitsufuji, Y., and Tao, M. Improving classifier-free guidance in masked diffusion: Low-dim theoretical insights with high-dim impact. In *The Fourteenth International Conference on Learning Representations*, 2026.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Rudin, L. I., Osher, S., and Fatemi, E. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

Sadat, S., Vontobel, T., Salehi, F., and Weber, R. M. Guidance in the frequency domain enables high-fidelity sampling at low cfg scales. *arXiv preprint arXiv:2506.19713*, 2025.

Salimans, T., Karpathy, A., Chen, X., and Kingma, D. P. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017.

Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., and Yuan, Z. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Tang, H., Wu, Y., Yang, S., Xie, E., Chen, J., Chen, J., Zhang, Z., Cai, H., Lu, Y., and Han, S. Hart: Efficient visual generation with hybrid autoregressive transformer. *arXiv preprint arXiv:2410.10812*, 2024.

Tian, K., Jiang, Y., Yuan, Z., Peng, B., and Wang, L. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.

van den Oord, A., Vinyals, O., and kavukcuoglu, k. Neural discrete representation learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, H., Liu, Y., Chi, J., Liu, F., Xue, R., and Duan, Y. Cfg-ctrl: Control-based classifier-free diffusion guidance. *arXiv preprint arXiv:2603.03281*, 2026a.

Wang, X., Dufour, N., Andreou, N., Cani, M.-P., Abrevaya, V. F., Picard, D., and Kalogeiton, V. Analysis of classifier-free guidance weight schedulers. *arXiv preprint arXiv:2404.13040*, 2024.

Wang, X., Cui, Y., Wang, J., Zhang, F., Wang, Y., Zhang, X., Luo, Z., Sun, Q., Li, Z., Wang, Y., et al. Multimodal learning with next-token prediction for large multimodal models. *Nature*, pp. 1–7, 2026b.

Wen, K., Zhang, H., Lin, H., and Zhang, J. From sparse dependence to sparse attention: unveiling how chain-of-thought enhances transformer sample efficiency. *arXiv preprint arXiv:2410.05459*, 2024.

Wu, Y., Chen, M., Li, Z., Wang, M., and Wei, Y. Theoretical insights for diffusion guidance: A case study for gaussian mixture models. *arXiv preprint arXiv:2403.01639*, 2024.

Xia, M., Xue, N., Shen, Y., Yi, R., Gong, T., and Liu, Y.-J. Rectified diffusion guidance for conditional generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13371–13380, 2025.

Xiao, L., Bahri, Y., Sohl-Dickstein, J., Schoenholz, S., and Pennington, J. Dynamical isometry and a mean field theory of cnns: How to train 10,000-layer vanilla convolutional neural networks. In *International conference on machine learning*, pp. 5393–5402. PMLR, 2018.

Xin, Y., Yan, J., Qin, Q., Li, Z., Liu, D., Li, S., Huang, V. S.-J., Zhou, Y., Zhang, R., Zhuo, L., et al. Lumina-mgpt 2.0: Stand-alone autoregressive image modeling. *arXiv preprint arXiv:2507.17801*, 2025.

Xiong, J., Liu, G., Huang, L., Wu, C., Wu, T., Mu, Y., Yao, Y., Shen, H., Wan, Z., Huang, J., et al. Autoregressive models in vision: A survey. *arXiv preprint arXiv:2411.05902*, 2024.

Ye, H., Kevin, R., and Molei, T. What exactly does guidance do in masked discrete diffusion models. *arXiv preprint arXiv:2506.10971*, 2025.

Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Yuan, L., Li, X., Zhang, Y., Zhang, J., Li, H., and Gao, X. Mlep: Multi-granularity local entropy patterns for generalized ai-generated image detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Zhao, X. and Schwing, A. Studying classifier (-free) guidance from a classifier-centric perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13271–13279, 2026.

A. Related Work

Auto-Regressive Image Generation has recently re-emerged as a promising paradigm for unifying text and image generation, alongside diffusion-based models (Rombach et al., 2022) and flow-matching methods (Esser et al., 2024). Early work on autoregressive image generation dates back to PixelCNN (Van den Oord et al., 2016), followed by a series of improvements in architecture design and training methodology (Reed et al., 2016; Salimans et al., 2017; Reed et al., 2017; Parmar et al., 2018; Chen et al., 2018). However, these methods generate images directly in pixel space. With the development of latent representations and visual tokenizers, such as VQ-VAE (van den Oord et al., 2017) and VQGAN (Esser et al., 2021), subsequent work moved autoregressive modeling to the latent space (Ramesh et al., 2021; Ding et al., 2021; Yu et al., 2022), substantially easing learning and improving generation quality. Building on these advances, recent works since 2024 have further scaled AR models to achieve performance competitive with diffusion models (Sun et al., 2024; Tian et al., 2024; Ma et al., 2024; Tang et al., 2024). Beyond competitive image-generation performance, AR models also provide an effective framework for unified multimodal generation, including text (Chen et al., 2025), video (Wang et al., 2026b), and depth (Liu et al., 2024). See Xiong et al. (2024); Jiang & Huang (2024) for comprehensive surveys of AR models. Recent efforts extend AR modeling to jointly handle text and vision, such as Janus-Pro (Chen et al., 2025), which unify understanding and generation within a single transformer. Similarly, Lumina-GPT (Liu et al., 2024; Xin et al., 2025) explores AR modeling for multi-modal signals including depth and geometry, while the Emu3 (Wang et al., 2026b) series and its successor Emu3.5 (Cui et al., 2025) scale AR frameworks to support high-quality image and video generation with strong cross-modal alignment. These works highlight the flexibility of AR modeling in handling diverse modalities within a unified token space.

Classifier-Free Guidance is a simple test-time algorithm for diffusion models that substantially improves generation quality (Ho & Salimans, 2022). A large body of work has subsequently sought to further enhance CFG. For example, prior studies improve CFG by designing low-quality score estimates (Karras et al., 2024), imposing manifold constraints (Chung et al., 2024; Cai et al., 2026), incorporating frequency-domain information (Sadat et al., 2025), and designing adaptive guidance strategies (Papalampidi et al., 2025; Malarz et al., 2025; Galashov et al., 2025). Beyond these algorithmic improvements, another line of work aims to understand why the original CFG is effective. Chidambaram et al. (2024); Fu et al. (2024); Zhao & Schwing (2026); Wu et al. (2024); Ye et al. (2025) study simplified learning settings for diffusion models, deriving theoretical insights into guidance in large pretrained models. In contrast, Bradley & Nakkiran (2024); Xia et al. (2025); Li & Jiao (2025); Jiao et al. (2025) directly assume bounded pretraining error and analyze the sampling process itself. These works interpret CFG through the lens of sampling algorithms and establish convergence guarantees from different perspectives. Most of the above works treat the role of CFG as fixed throughout sampling. By contrast, Gao et al. (2026); Lehman Pavasovic et al. (2025); Li et al. (2025); Jin et al. (2025); Rojas et al. (2026); Wang et al. (2024); Kynkäänniemi et al. (2024) examine how the effect of CFG varies across different sampling stages, propose non-stationary views of guidance, and develop adaptive algorithms accordingly. From a different perspective, Wang et al. (2026a) studies CFG through the lens of control theory. However, all of these works focus on understanding CFG in diffusion models. In contrast, our work provides an understanding of CFG in AR models.

Theoretical studies of training dynamics have provided important insights into neural networks for computer vision. Existing works cover a broad range of architectures, including MLPs (Du et al., 2019), CNNs (Cao et al., 2022), and Vision Transformers (Jelassi et al., 2022), as well as different learning regimes, including lazy learning (Jacot et al., 2018) and feature learning (Lu et al., 2020). Following the seminal work of Jacot et al. (2018), a line of research uses the Neural Tangent Kernel (NTK) to study the optimization and generalization behavior of neural networks. For CNNs, Arora et al. (2019b) develops the convolutional NTK, while Arora et al. (2019a) derives explicit kernel expressions and shows how the kernel spectrum determines the convergence rate. Mei et al. (2021) further relates the kernel regime to invariance properties of neural networks; see Golikov et al. (2022) for a survey of NTK analyses. Beyond the kernel regime, Lu et al. (2020) and Xiao et al. (2018) study mean-field limits of CNNs and establish convergence results. Another line of work builds problem-specific feature models to characterize training dynamics. For example, Allen-Zhu & Li (2020) shows the importance of multi-view data structures for ensemble learning, and Cao et al. (2022) derives phase-transition results for benign overfitting. More recently, several works have studied the learning dynamics of transformers. Jelassi et al. (2022) and Huang et al. (2024) analyze how transformers learn spatial relationships among visual features. Our work also falls into this category: we study how the attention module contributes to CFG in AR image generation.

Figure 7. Generated images from $\text{logit}^{\text{cond}}$, logit^{λ}, and $\text{logit}^{\text{diff}} = \text{logit}^{\text{cond}} - \text{logit}^{\text{uncond}}$, shown from left to right in each grid. Prompts: **Left:** A street at night is lit up by streetlights. **Right:** An old man is wearing an odd hat.

B. Experimental Details

B.1. Experimental Details and Additional Results of Image Generation in Section 3.1

In this section, we describe the image generation setup used in our Section 3.1 and provide more results.

We generate images using LlamaGen (0.7B) (Sun et al., 2024), Lumina-mGPT 2.0 (7B) (Xin et al., 2025), and Emu3 (8B) (Wang et al., 2026b). We employ the GenEval benchmark (Ghosh et al., 2023), which consists of 553 prompts of varying lengths. Generations are performed via greedy decoding ($\tau = 0$) using three logit configurations: $\text{logit}^{\text{cond}}$, logit^{λ}, and the difference $\text{logit}^{\text{cond}} - \text{logit}^{\text{uncond}}$. For logit^{λ}, we set $\lambda = 10$ across all models to ensure generation quality under greedy settings. Aside from the decoding strategy and guidance scale, all other hyperparameters follow the standard configurations of the respective models. Image resolutions are set to 512×512 for LlamaGen and Emu3, and 768×768 for Lumina-mGPT 2.0. During the image generation process, we also save the generated token sequences prior to decoding. By saving these token sequences, we can efficiently retrieve the corresponding logits via prefilling in any subsequent experiments. Our experiments are run on an A100 Nvidia GPU with 40G memory. The evaluation time used for DPG-Bench in Section 5 follows the standard setting of this benchmark.

More generated examples are provided in Figure 7, which exhibit similar phenomena to those observed in Figure 1 (Section 3.1).

B.2. Experimental Details of Texture Complexity Metrics in Section 3.1

In this section, we provide formal definitions and implementation details for the Texture Complexity metrics analyzed in Section 3.1. Let an image be denoted by $I \in \{0, \ldots, 255\}^{C \times H \times W}$, where $I_{c,i,j}$ represents the pixel value at location (i, j) in channel $c \in \{1, \ldots, C\}$. In our experiments, we use standard RGB images ($C = 3$). We evaluate these metrics across images generated by $\text{logit}^{\text{cond}}$, logit^{λ} and $\text{logit}^{\text{cond}} - \text{logit}^{\text{uncond}}$. We report metrics as averages across images within each of the generation groups established in Section B.1.

Total Variation (TV)**.** To measure the spatial smoothness of an image, we define the Total Variation as the average of normalized ℓ_1-norm of discrete gradients:

$$\mathrm{TV}(I) = \frac{1}{C}\sum_{c=1}^{C}\left(\frac{1}{(H-1)W}\sum_{i=1}^{H-1}\sum_{j=1}^{W}|\nabla_i I_{c,i,j}| + \frac{1}{H(W-1)}\sum_{i=1}^{H}\sum_{j=1}^{W-1}|\nabla_j I_{c,i,j}|\right),$$

where $\nabla_i I_{c,i,j} = I_{c,i+1,j} - I_{c,i,j}$ and $\nabla_j I_{c,i,j} = I_{c,i,j+1} - I_{c,i,j}$ represent horizontal and vertical forward differences, respectively.

Pixel Entropy (Entropy)**.** This metric captures the global richness of the intensity distribution. For each channel c, let $p_c(v) = (HW)^{-1}\sum_{i,j}\mathbb{I}\{I_{c,i,j} = v\}$ be the empirical intensity distribution for $v \in \{0, \ldots, 255\}$. The pixel entropy is computed as:

$$\mathrm{Entropy}(I) = \frac{1}{C}\sum_{c=1}^{C} H(p_c) = \frac{1}{C}\sum_{c=1}^{C}\left(-\sum_{v=0}^{255} p_c(v)\log_2 p_c(v)\right).$$

Orientation Entropy (OriEntropy)**.** To characterize structural regularity, we measure the uncertainty of gradient orientations within the image. For each channel c, we compute the Sobel gradients $(G^{(c)}_{x,i,j}, G^{(c)}_{y,i,j})$ at position (i,j), which capture the local intensity variations along the horizontal and vertical axes, respectively, and the corresponding magnitude $M^{(c)}_{i,j} = \left((G^{(c)}_{x,i,j})^2 + (G^{(c)}_{y,i,j})^2\right)^{1/2}$. The orientation is defined as $\theta^{(c)}_{i,j} = \mathrm{atan2}(G^{(c)}_{y,i,j}, G^{(c)}_{x,i,j})$, mapped to the undirected range $[0, \pi)$. To focus on salient structures, we only consider the set of pixels $\mathcal{M}_c = \{(i,j) \mid M^{(c)}_{i,j} > \bar{M}^{(c)}\}$, where $\bar{M}^{(c)}$ is the mean magnitude of the channel. We construct a magnitude-weighted orientation distribution q_c over $K = 18$ discrete bins $\{\Theta_k\}_{k=1}^{K}$ partitioning $[0, \pi)$:

$$q_c(k) = \frac{1}{Z_c}\sum_{(i,j)\in\mathcal{M}_c} M^{(c)}_{i,j}\cdot\mathbb{I}\{\theta^{(c)}_{i,j} \in \Theta_k\}, \quad Z_c = \sum_{(i,j)\in\mathcal{M}_c} M^{(c)}_{i,j}.$$

The OriEntropy of an image is defined as the average Shannon entropy across channels, which means $\mathrm{OriEntropy}(I) = C^{-1}\sum_{c=1}^{C} H(q_c)$.

B.3. Experimental Details of the Similarity Metrics in Section 3.2

In this section, we provide the experimental details in Section 3.2 when we measure the similarity between $\mathrm{logit}^{\mathrm{cond}}_t$ and $\mathrm{logit}^{\mathrm{uncond}}_t$ in logit and pixel spaces.

For all models, we collect these logits and their corresponding predicted tokens by prefilling the models with token sequences generated via logit^{λ} (as described in Section B.1). We then compute the difference $\mathrm{logit}^{\mathrm{cond}} - \mathrm{logit}^{\mathrm{uncond}}$. The logit-space similarities $(\alpha^{\mathrm{diff}}, \alpha^{\mathrm{cond}}, \alpha^{\mathrm{uncond}})$ are calculated with $k = 5$ for all three models. In the pixel space, we decode the predicted tokens to obtain $\beta^{\mathrm{diff}}, \beta^{\mathrm{cond}}$, and β^{uncond}. Finally, all metrics are averaged across images generated from the prompts defined in Section B.1.

To provide a qualitative comparison, we present sampled images decoded from $\mathrm{logit}^{\mathrm{cond}}$ and $\mathrm{logit}^{\mathrm{uncond}}$ in Figures 8 and 9. We observe that the images generated by $\mathrm{logit}^{\mathrm{cond}}$ and $\mathrm{logit}^{\mathrm{uncond}}$ are highly similar across all cases, a finding that aligns with the results in Figure 3 and the theoretical analysis in Section 3.2.

B.4. Experimental Details and Additional Results of Section 3.3

In this section, we provide experimental details of quantifying repetition behavior in Section 3.3, and additional results of Emu3 and LlamaGen models.

In our experiments, we first identify high-frequency visual tokens and use them to construct prefixes. For each evaluated model, we pool the tokens generated across all images to compute their occurrence frequency. We define the set of common visual tokens as the top 1% most frequent tokens for each model, which comprises more than 100 tokens. Then, for each prompt in the GenEval benchmark (Ghosh et al., 2023), we randomly sample one such token to construct a prefix covering 20% of the visual token positions, as described in Section 3.3. Specifically, the prefix occupies the first 20% of rows in the 2D latent space: 7 rows for LlamaGen (32×32), 13 for Emu3 (64×64), and 20 for Lumina-mGPT (96×96). Following

Figure 8. Decoded images from $\text{logit}^{\text{cond}}$, $\text{logit}^{\text{uncond}}$, and $\text{logit}^{\text{diff}} = \text{logit}^{\text{cond}} - \text{logit}^{\text{uncond}}$ collected in Section B.3, shown from left to right in each grid. Prompts: **Left:** A glass bowl filled with oranges on a table. **Right:** An orange-reddish rose in a vase filled with water on top of a table.

generation conditioned on the prompts and prefixes, we compute r_t^{rep} at every position and average these values across all non-prefix positions to obtain r^{rep}. Finally, all results are averaged across the generated images.

Figures 10 and 11 present the average r_t^{rep} across images for LlamaGen and Emu3. The results are consistent with the findings in Section 3.3, demonstrating that tokens generated from $\text{logit}^{\text{cond}}$ and $\text{logit}^{\text{uncond}}$ tend to repeat the prefix tokens.

B.5. Experimental Details and Additional Results of Section 5

In this section, we provide additional details on AttnReuse. We first present the complete AttnReuse algorithm in Algorithm 1, which describes how attention weights are reused in selected layers while the remaining layers are computed normally. We then summarize the evaluation settings for each model in Table 3. Finally, we present qualitative results in Fig. 12, showing that AttnReuse preserves the semantic content and overall visual structure of the generated images with no noticeable degradation compared to standard CFG.

In Table 3, we list the evaluation settings for each models. For both AR and diffusion models, AttnReuse is applied only to a selected subset of layers during generation to balance efficiency and fidelity. For diffusion models, where computation is additionally distributed across denoising timesteps, we further apply AttnReuse to selected timesteps.

For the three evaluated AR models, each backbone contains 32 Transformer layers. We apply AttnReuse to the middle 16 layers and keep the early and late layers unchanged. This design is motivated by the observation that early layers are more sensitive to low-level token representations, while late layers more directly affect the final prediction logits. Therefore, applying AttnReuse to the middle layers provides a better trade-off between computational efficiency and output fidelity.

For SD3-Medium, which contains 24 Transformer layers, we similarly apply AttnReuse to middle layers. In addition, since the early denoising timesteps largely determine the coarse semantics and global layout of the final image, we keep these timesteps unchanged and apply AttnReuse only to the last 90% of denoising timesteps. This timestep selection reduces computation while preserving the semantic structure established in the early generation stage.

For the stochastic setting $\tau = 1$, we follow the officially released inference configuration of each model. For the greedy setting $\tau = 0$, directly using the official CFG scale often produces semantically incomplete outputs. We therefore search the CFG scale under $\tau = 0$ for each model and report the selected setting used in our evaluation. Across all settings, AttnReuse

Figure 9. Decoded images from $\text{logit}^{\text{cond}}$, $\text{logit}^{\text{uncond}}$, and $\text{logit}^{\text{diff}} = \text{logit}^{\text{cond}} - \text{logit}^{\text{uncond}}$ collected in Section B.3, shown from left to right in each grid. Prompts: **Left:** A street at night is lit up by streetlights. **Right:** An old man is wearing an odd hat.

Model	Type	τ	CFG	AttnReuse Layers	Timesteps
Emu3	AR	$\{0, 1\}$	$\{14.0, 7.0\}$	8–23	–
Janus-Pro 7B	AR	$\{0, 1\}$	5.0	8–23	–
Lumina-GPT 2.0	AR	$\{0, 1\}$	$\{20.0, 4.0\}$	8–23	–
Stable Diffusion 3	Diffusion	–	7.0	9–17	Last 90%

Table 3. Evaluation settings across models. For AR models, AttnReuse is applied to a subset of attention layers (Layers 8–23). For diffusion models, it is applied to selected layers (Layers 9–17) and across the last 90% of timesteps.

introduces no additional learnable parameters and is fully compatible with pretrained checkpoints.

C. Detailed Training Algorithm in Section 4

We train the model by (two-step) population gradient descent on the data distribution defined above, as specified in Algorithm 2. The population cross-entropy loss is

$$\mathcal{L}_{\text{vis}}(\theta) = \mathbb{E}_{(z_1, x_{1:T_{\text{vis}}}) \sim P_{\text{data}}} \left[T_{\text{vis}}^{-1} \sum_{t=0}^{T_{\text{vis}}-1} -\log P_\theta(x_{t+1} \mid z_1, x_{1:t}) \right].$$

To train the parameters $\theta = (W_V, W_{QK})$, we initialize them as $\theta^{(0)} = (0, 0)$. We first update W_V by one gradient step, and then optimize W_{QK} for one step while keeping W_V fixed. This two-stage training protocol follows prior analyses of transformer training dynamics and is used for theoretical tractability (Wen et al., 2024; Nichani et al.).

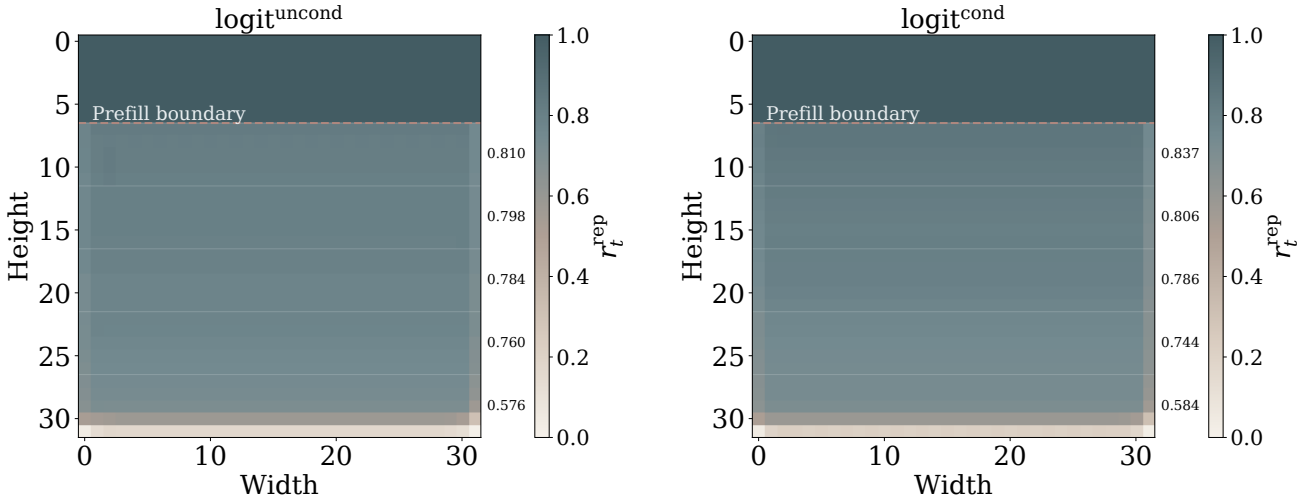

Figure 10. Position-wise repetition ratio r_t^{rep} of LlamaGen, averaged over samples. Both $\text{logit}^{\text{cond}}$ and $\text{logit}^{\text{uncond}}$ show strong repetition of the prescribed prefix token.

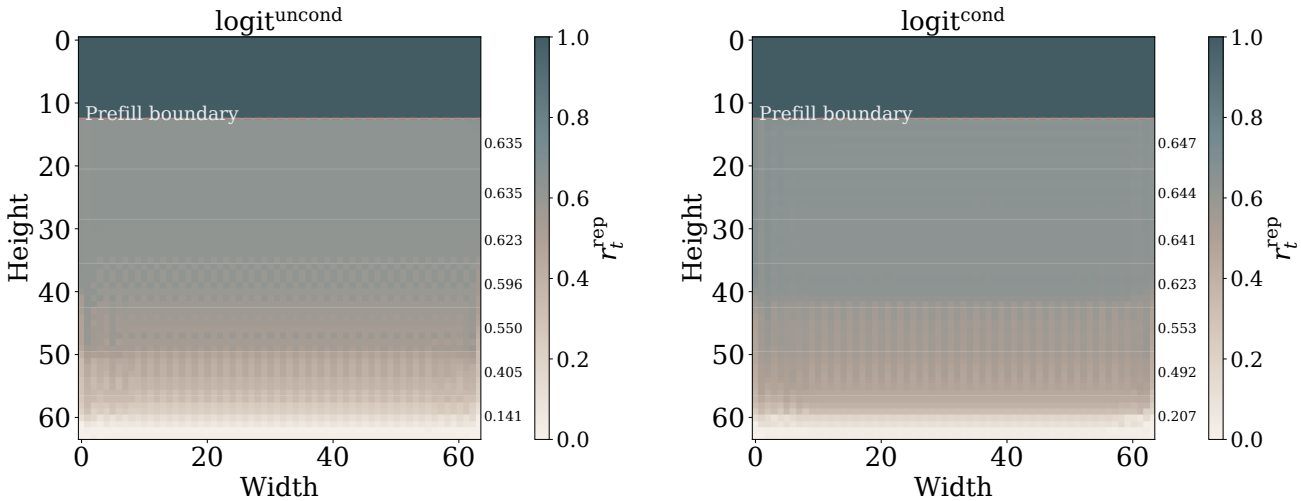

Figure 11. Position-wise repetition ratio r_t^{rep} of Emu3, averaged over samples. Both $\text{logit}^{\text{cond}}$ and $\text{logit}^{\text{uncond}}$ show strong repetition of the prescribed prefix token.

Algorithm 2 Two-step population gradient descent

Require: Transformer P_θ; learning rates $\eta_V, \eta_{QK} > 0$; population loss $\mathcal{L}_{\text{vis}}(\cdot)$.

$\theta^{(0)} \leftarrow (0,\, 0)$

$W_V^{(1)} \leftarrow -\eta_V \, \nabla_{W_V} \mathcal{L}_{\text{vis}}(\theta^{(0)}); \quad \theta^{(1)} \leftarrow \big(W_V^{(1)},\, 0\big)$

$W_{QK}^{(1)} \leftarrow -\eta_{QK} \, \nabla_{W_{QK}} \mathcal{L}_{\text{vis}}(\theta^{(1)}); \quad \theta^{(2)} \leftarrow \big(W_V^{(1)},\, W_{QK}^{(1)}\big)$

return $\theta^{(2)}$

D. Technical Preliminaries for the Proof

In this section, we introduce the notation and technical results that will be used in the proof of the main theorem. We clarify the notation and make the intermediate outputs of the model explicit to facilitate the subsequent analysis. We then derive explicit expressions for the gradients with respect to the model parameters, which will play a central role in the analysis.

Transformer Computation Decomposition To analyze the computational flow of the Transformer defined in Section 4, we first decompose the attention mechanism under the given parameter constraints. For notational simplicity, we denote the initial text token z_1 as x_0 and fix its position at $t = 0$. Let s_{tl} and α_{tl} denote the pre-softmax attention score and the corresponding weight from query position t to key position l. The attention output $\tilde{o}_t$ and the resulting logit logit_t at position t are computed as follows:

$$s_{tl} = h_t^\top \widetilde{W}_{QK} h_l, \quad \alpha_{tl} = \frac{\exp(s_{tl})}{\sum_{s=0}^{t} \exp(s_{ts})}, \quad \tilde{o}_t = \sum_{l=0}^{t} \alpha_{tl} \widetilde{W}_V h_l, \quad \mathrm{logit}_t = \widetilde{W}_{\mathrm{lm}}^\top \tilde{o}_t,$$

where for simplicity, we denote logit_t for $\mathrm{logit}_t^{\mathrm{cond}}$. Because the parameterization is restricted to

$$\widetilde{W}_{QK} = \begin{bmatrix} 0 & 0 \\ W_{QK} & 0 \end{bmatrix}, \quad \widetilde{W}_V = \begin{bmatrix} W_V & 0 \\ 0 & 0 \end{bmatrix}, \quad \widetilde{W}_{\mathrm{lm}} = \begin{bmatrix} W_{\mathrm{lm}} \\ 0 \end{bmatrix},$$

and input satisfies $h_t = [f_{x_t}^\top, e_t^\top]^\top$, we can simplify as

$$s_{tl} = e_t^\top W_{QK} f_{x_l}, \quad \tilde{o}_t = \begin{bmatrix} o_t \\ 0 \end{bmatrix}, \quad o_t = \sum_{l=0}^{t} \alpha_{tl} W_V f_{x_l}, \quad \mathrm{logit}_t = W_{\mathrm{lm}}^\top o_t.$$

For notational convenience, we use the superscript uc to denote variables associated with the unconditional case (uncond). Correspondingly, the attention weights $\alpha_{tl}^{\mathrm{uc}}$, attention outputs $\tilde{o}_t^{\mathrm{uc}}$, and logits $\mathrm{logit}_t^{\mathrm{uc}}$ for the unconditional model (simplified version of $\mathrm{logit}_t^{\mathrm{uncond}}$) are defined as:

$$\alpha_{tl}^{\mathrm{uc}} = \frac{\exp(s_{tl})}{\sum_{s=1}^{t} \exp(s_{ts})}, \quad \tilde{o}_t^{\mathrm{uc}} = \sum_{l=1}^{t} \alpha_{tl}^{\mathrm{uc}} \widetilde{W}_V h_l, \quad \mathrm{logit}_t^{\mathrm{uc}} = \widetilde{W}_{\mathrm{lm}}^\top \tilde{o}_t^{\mathrm{uc}}.$$

Similarly, by exploiting the previously defined block-matrix parameterization, these expressions can be simplified in terms of the token-level representations:

$$\tilde{o}_t^{\mathrm{uc}} = \begin{bmatrix} o_t^{\mathrm{uc}} \\ 0 \end{bmatrix}, \quad o_t^{\mathrm{uc}} = \sum_{l=1}^{t} \alpha_{tl}^{\mathrm{uc}} W_V f_{x_l}, \quad \mathrm{logit}_t^{\mathrm{uc}} = W_{\mathrm{lm}}^\top o_t^{\mathrm{uc}}$$

For clarity, we denote the i-th element of $\mathrm{logit}_t \in \mathbb{R}^{K+J}$ as $\mathrm{logit}_t[i]$, which also applies for $\mathrm{logit}_t^{\mathrm{uc}}$ and $\mathrm{logit}_t^{\lambda}$.

Equivalent Form of CFG on Output Given that the output matrix W_{lm} is fixed, we can establish an operational equivalence between applying CFG at the logit level and the attention output level. At position t, the CFG-modified logits with guidance scale λ are formulated as:

$$\mathrm{logit}_t^{\lambda} = \mathrm{logit}_t + \lambda \left(\mathrm{logit}_t - \mathrm{logit}_t^{\mathrm{uc}}\right) = W_{\mathrm{lm}}^\top \big(o_t + \lambda(o_t - o_t^{\mathrm{uc}})\big), \tag{4}$$

where the first equation follows from the definition in Eqn. (2), and the second relies on the simplified representation introduced previously.

Compact Notation for Sequences To simplify the exposition, we adopt the superscript $\times L$ to denote a token repeated L times. For instance, a representative sample from our data distribution is denoted as $\big(K+J+k_1, (K+j_1)^{\times\tau}, (k_1)^{\times L}, (K+j_1)^{\times(T-\tau-L)}\big)$. This sequence consists of an initial text token $K + J + k_1$, followed by a block of τ background tokens $K + j_1$, a segment of L feature tokens k_1, and a final padding of $T - \tau - L$ background tokens.

Per-sequence loss and Population loss To analyze the training dynamics, we derive a closed-form expression for the population loss. We begin by defining the loss for an individual sequence $(x_0, x_1, \ldots, x_T) \in \mathcal{I}_{k,j}$, where $k \in [K]$ and $j \in [J]$ denote the visual feature and background classes, respectively. The cross-entropy loss for this sequence is given by:

$$\mathcal{L}(k, j; \theta) = -\sum_{t=1}^{T-1} \log P_\theta(x_{t+1} | x_0, x_1, \cdots x_t), \quad \text{where} \quad (x_0, x_1, \cdots, x_T) = I_{k,j}. \tag{5}$$

The population loss under the data distribution P_{data} (defined in Section 4) is then formulated as the expected loss over all sequences:

$$\mathcal{L}_{\text{vis}}(\theta) = \mathbb{E}_{(z_1, x_{1:T}) \sim P_{\text{data}}} \left[\sum_{t=1}^{T-1} -\log P_\theta(x_{t+1} \mid x_0, x_{1:t}) \right] = \sum_{k=1}^{K} p_k \sum_{j=1}^{J} q_j \, \mathcal{L}(k, j; \theta), \tag{6}$$

where the first equation is the definition of population loss in Section 4, and the second equation holds because the probability of sampling visual feature class k is p_k, and sampling visual background class j is q_j.

D.1. Gradient Formulas

To facilitate the subsequent analysis of gradient dynamics, we define several auxiliary quantities that characterize the prediction error and its projection back into the hidden representation space.

Let $P_t = P_\theta(x_{t+1}|x_{0:t}) = \texttt{sm}(\text{logit}_t) \in \mathbb{R}^{K+J}$ be the predicted probability distribution over the vocabulary at position t. We define the prediction residual δ_t and its pullback $\tilde{g}_t$ through the language modeling head $\widetilde{W}_{\text{LM}}$ as follows:

$$y_t = \mathbf{1}_{x_{t+1}} \in \mathbb{R}^{K+J}, \quad \delta_t = P_t - y_t \in \mathbb{R}^{K+J}, \quad \tilde{g}_t = \widetilde{W}_{\text{lm}} \delta_t = \begin{bmatrix} g_t \\ 0 \end{bmatrix} \in \mathbb{R}^D, \quad g_t = W_{\text{lm}} \delta_t \in \mathbb{R}^d,$$

where $\mathbf{1}_i$ is the i-th standard basis vector in $\mathbb{R}^{K+J}$, and here $\mathbf{1}_{x_{t+1}}$ acts as indicator vector for class x_{t+1}. The structure of $\tilde{g}_t$ comes from the simplified structure of parameters.

Lemma 2 (Single-sequence gradients)**.** *Fix any token sequence $(x_0, \ldots, x_T)$ and parameters $\theta = (W_V, W_{QK})$. The gradients of the single-sequence loss $\mathcal{L}$ in Eqn.* (5) *with respect to W_V and W_{QK} are*

$$\nabla_{W_V} \mathcal{L} = \sum_{t=1}^{T-1} \sum_{l=0}^{t} \alpha_{tl} \, g_t \, f_{x_l}^\top, \tag{7}$$

$$\nabla_{W_{QK}} \mathcal{L} = \sum_{t=1}^{T-1} \sum_{l=0}^{t} \alpha_{tl} \left[g_t^\top (W_V f_{x_l} - o_t) \right] e_t f_{x_l}^\top. \tag{8}$$

Proof. **Gradient w.r.t. W_V.** Write the per-step loss as $\mathcal{L}_t = -\log P_t(x_{t+1}|x_{0:t})$. Because $P_t = \texttt{sm}(\text{logit}_t) = \texttt{sm}(W_{\text{lm}}^\top o_t)$. Applying the standard gradient property of the softmax cross-entropy loss and the chain rule via $\text{logit}_t = W_{\text{lm}}^\top o_t$, we obtain

$$\frac{\partial \mathcal{L}_t}{\partial \text{logit}_t} = P_t - y_t = \delta_t, \quad \frac{\partial \mathcal{L}_t}{\partial o_t} = W_{\text{lm}} \delta_t = g_t.$$

Since $o_t = \sum_{l=0}^{t} \alpha_{tl} W_V f_{x_l}$ and α_{tl} does not depend on W_V, differentiating entry-wise gives

$$\frac{\partial o_t}{\partial (W_V)_{ab}} = \sum_{l=0}^{t} \alpha_{tl} \, (f_{x_l})_b \, \mathbf{1}_a,$$

where $\mathbf{1}_a$ is the a-th standard basis vector in $\mathbb{R}^{K+J}$. Applying the chain rule from $\frac{\partial \mathcal{L}_t}{\partial o_t}$, $\frac{\partial o_t}{\partial (W_V)_{ab}}$ and collecting into matrix form, we have

$$\frac{\partial \mathcal{L}_t}{\partial W_V} = g_t \Big(\sum_{l=0}^{t} \alpha_{tl} \, f_{x_l} \Big)^\top = \sum_{l=0}^{t} \alpha_{tl} \, g_t \, f_{x_l}^\top.$$

Summing over $t = 1, \ldots, T-1$, we have

$$\nabla_{W_V} \mathcal{L} = \sum_{t=1}^{T-1} \frac{\partial \mathcal{L}_t}{\partial W_V} = \sum_{t=1}^{T-1} \sum_{l=0}^{t} \alpha_{tl} \, g_t \, f_{x_l}^\top.$$

Gradient w.r.t. W_{QK}**.** We decompose the chain via the intermediate variables α_{tj} and s_{tl}.

Since $o_t = \sum_{l=0}^{t} \alpha_{tl} W_V f_{x_l}$, varying α_{tj} while keeping all other attention weights fixed gives $\partial o_t / \partial \alpha_{tj} = W_V f_{x_j}$. Combined with $\partial \mathcal{L}_t / \partial o_t = g_t$, by chain rule, we have

$$\frac{\partial \mathcal{L}_t}{\partial \alpha_{tj}} = g_t^\top W_V f_{x_j}.$$

From the softmax formula $\alpha_{tj} = \exp(s_{tj}) / \sum_{s'=0}^{t} \exp(s_{ts'})$, standard differentiation yields the Jacobian

$$\frac{\partial \alpha_{tj}}{\partial s_{tl}} = \alpha_{tj} \big(\mathbb{I}\{j = l\} - \alpha_{tl} \big).$$

By the chain rule over j, we have

$$\begin{aligned}
\frac{\partial \mathcal{L}_t}{\partial s_{tl}} &= \sum_{j=0}^{t} \frac{\partial \mathcal{L}_t}{\partial \alpha_{tj}} \cdot \frac{\partial \alpha_{tj}}{\partial s_{tl}} \\
&= \sum_{j=0}^{t} \left(g_t^\top W_V f_{x_j} \right) \alpha_{tj} (\mathbb{I}\{j = l\} - \alpha_{tl}) \\
&= \alpha_{tl}\, g_t^\top W_V f_{x_l} - \alpha_{tl} \sum_{j=0}^{t} \alpha_{tj}\, g_t^\top W_V f_{x_j} \\
&= \alpha_{tl}\, g_t^\top (W_V f_{x_l} - o_t),
\end{aligned}$$

where the second and third equalities follow from previous results of $\frac{\partial \mathcal{L}_t}{\partial \alpha_{tj}}, \frac{\partial \alpha_{tj}}{\partial s_{tl}}$ and direct expansion, the last step invokes the structural form of $o_t = \sum_{j=0}^{t} \alpha_{tj} g_t^\top W_V f_{x_j}$.

From the simplification, we have $s_{tl} = e_t^\top W_{QK} f_{x_l}$, the differentiation gives $\partial s_{tl} / \partial W_{QK} = e_t f_{x_l}^\top$. Combining the previous results and applying the chain rule, we obtain

$$\frac{\partial \mathcal{L}_t}{\partial W_{QK}} = \sum_{l=0}^{t} \frac{\partial \mathcal{L}_t}{\partial s_{tl}} \cdot e_t f_{x_l}^\top = \sum_{l=0}^{t} \alpha_{tl}\, g_t^\top (W_V f_{x_l} - o_t)\, e_t f_{x_l}^\top,$$

where the first equation comes from chain rule expansion, and the second equation follow from previous results of $\frac{\partial \mathcal{L}_t}{\partial s_{tl}}, \frac{\partial \mathcal{L}_t}{\partial W_{QK}}$.

Summing over $t = 1, \ldots, T-1$, we have

$$\nabla_{W_{QK}} \mathcal{L} = \sum_{t=1}^{T-1} \frac{\partial \mathcal{L}_t}{\partial W_{QK}} = \sum_{t=1}^{T-1} \sum_{l=0}^{t} \alpha_{tl} \left[g_t^\top (W_V f_{x_l} - o_t) \right] e_t f_{x_l}^\top.$$

Thus, we conclude the proof of Lemma 2. □

E. Proof of Theorem 1

In the proof of Theorem 1, we follow the training protocol in Appendix C and analyze the approximation errors of $\theta^{(1)}$ and $\theta^{(2)}$ sequentially. For ease of notation, we will write T_{vis} as T in the proof. Before the proof, we will first state the formal version of Theorem 1 as follows.

Theorem 3. *Under Assumptions 1 and 2, suppose the following parameter regime holds as $T_{vis} \to \infty$: the start positions of features satisfy $\tau_k = \tau = \Theta(T_{vis}^{\beta})$ for all $k \in [K]$, and some fixed $\beta \in (1/2, 1)$; the number of visual background tokens satisfies $J = \lfloor \tau \log T_{vis} / cL \rfloor$ for a constant c with $1 < c < 1/(2 - 2\beta)$; the number of visual feature tokens satisfies $K = \Theta(1)$; the distribution of images satisfies $p_k = \Theta(1/K)$, $q_j = 1/J$; the learning rates satisfy $\eta_V = \Theta(1)$, $\eta_{QK} = \Theta(T_{vis}^{\beta})$. Let $\theta^{(2)} = (W_V^{(1)}, W_{QK}^{(1)})$ be the output of Algorithm 2 applied to the Transformer model P_θ Then the following three statements hold.*

1. (Training loss.) *The population loss is asymptotically small, with*

$$\mathcal{L}_{vis}(\theta^{(2)}) = \Theta\Big(\frac{(\log T_{vis})^2}{T_{vis}}\Big).$$

2. (Conditional generation fails and repeats.) *With guidance scale $\lambda = 0$ (pure conditional generation), the generated sequence is not a valid image and repeats backgrounds for any query:*

$$\hat{I}^0_{k,j}(\theta^{(2)}) = (K+j, \cdots, K+j) \neq I_{k,j}, \qquad \text{for all } k \in [K],\ j \in [J].$$

3. (CFG recovers valid images.) *There exists $\lambda^* = \Theta(1)$ with $\lambda^* > 0$ such that CFG at guidance scale λ^* generates a valid image for every query:*

$$\hat{I}^{\lambda^*}_{k,j}(\theta^{(2)}) = I_{k,j}, \qquad \text{for all } k \in [K],\ j \in [J].$$

4. (Alignment of conditional and unconditional prediction.) *The conditional model and unconditional model has the same prediction when generating image with guidance scale λ^*, which means*

$$\arg\max_{x \in \mathcal{X}_{vis}} \text{logit}^{\text{cond}}_t = \arg\max_{x \in \mathcal{X}_{vis}} \text{logit}^{\text{uncond}}_t$$

Proof Roadmap. We begin by outlining the roadmap for the proof of Theorem 3. Following the two-stage training procedure in Algorithm 2, we first analyze the update of $W_V^{(1)}$ and then characterize the subsequent update of $W_{QK}^{(1)}$ based on the first-stage results. We then demonstrate that the training loss is asymptotically minimized. Finally, we establish the model's generation behavior: while the model successfully generates target images using CFG with a suitable guidance scale, it fails when restricted to conditional decoding. We also prove that the predictions of the conditional and unconditional branches remain aligned.

Steps 1–4 establish the first-stage parameter estimates: Step 1 computes the population gradient of W_V at initialization; Step 2 derives the asymptotic order of its coefficients; Step 3 verifies a pivotal asymptotic property induced by the parameter regime; and Step 4 converts these results into an explicit one-step approximation for $W_V^{(1)}$, analyzing how embeddings are projected into the value space. Step 5 examines the logits and probabilities induced by $W_V^{(1)}$, laying the groundwork for the second stage. Steps 6–7 analyze the population gradient and one-step update of W_{QK}, yielding an explicit approximation for $W_{QK}^{(1)}$ and characterizing the computation of position-feature correlations. Step 8 leverages these estimates to analyze CFG behavior across different time regimes and defines the valid range for the guidance scale. Step 9 proves that the training loss is asymptotically small. Steps 10–11 establish our two generation claims: CFG-based decoding correctly generates the target image, whereas pure conditional decoding collapses into repetitive background patterns. Step 12 proves the alignment between conditional and unconditional predictions. Finally, Step 13 synthesizes these intermediate results to complete the proof of Theorem 3.

E.1. Step 1: Population Gradient of W_V at Initialization

To compute the first update in Algorithm 2, we first derive $\nabla_{W_V}\mathcal{L}_{\text{vis}}(\theta^{(0)})$ at initialization, where $\theta^{(0)} = (W_V^{(0)}, W_{QK}^{(0)}) = (0, 0)$.

Lemma 4 (Population gradient of W_V at initialization)**.** *At initialization of Algorithm 2, where $\theta^{(0)} = (W_V^{(0)}, W_{QK}^{(0)}) = (0, 0)$, the population gradient is*

$$\nabla_{W_V}\mathcal{L}_{vis}(\theta^{(0)}) = \sum_{k=1}^{K} p_k \sum_{j=1}^{J} q_j \nabla_{W_V}\mathcal{L}(k, j; \theta^{(0)}),$$

where for each (k, j),

$$\begin{aligned}\nabla_{W_V}\mathcal{L}(k, j; \theta^{(0)}) = \bar{f}\Big(&A_{K+J+k} f_{K+J+k}^\top + A_k f_k^\top + A_{K+j} f_{K+j}^\top\Big)\\ &- f_{K+j}\Big(B_{K+J+k} f_{K+J+k}^\top + B_k f_k^\top + B_{K+j} f_{K+j}^\top\Big)\\ &- f_k\Big(C_{K+J+k} f_{K+J+k}^\top + C_k f_k^\top + C_{K+j} f_{K+j}^\top\Big),\end{aligned} \tag{9}$$

for some positive constant $A_{K+J+k}, A_k, A_{K+j}, B_{K+J+k}, B_k, B_{K+j}, C_{K+J+k}, C_k, C_{K+j}$.

Proof. **Uniform attention and uniform prediction at initialization.** When $W_{QK} = 0$, because $s_{tl} = e_t^\top W_{QK} f_{x_l}$, all pre-softmax attention scores s_{tl} are zero, so the attention score $\alpha_{tl} = 1/(t+1)$ for all $l = 0, \ldots, t$. When $W_V = 0$, $o_t = \sum_{l=0}^{t} \alpha_{tl} W_V f_{x_l} = 0$ for all t, hence

$$P_t = \mathtt{sm}(\big[\mathbf{0}\big]) = \frac{1}{K+J}\mathbf{1}, \quad g_t = W_{\text{lm}}\delta_t = \bar{f} - f_{x_{t+1}}, \quad \bar{f} = \frac{1}{K+J}\sum_{a=1}^{K+J} f_a,$$

where $\mathbf{1} \in \mathbb{R}^{K+J}$ denotes all one vector, $\mathbf{0} \in \mathbb{R}^{K+J}$ denotes all zero vector, and the second equations holds because of the fixed languages model head $W_{\text{lm}} = [f_1, \cdots .f_{K+J}]$. The piecewise structure of $f_{x_{t+1}}$ determined by the data model given visual feature class $k \in [K]$ and visual background class $j \in [J]$ is

$$f_{x_{t+1}} = \begin{cases} f_{K+j}, & 1 \le t \le \tau - 2, \\ f_k, & \tau - 1 \le t \le \tau + L - 2, \\ f_{K+j}, & \tau + L - 1 \le t \le T - 1. \end{cases} \tag{10}$$

Applying Eqn. (7) in Lemma 2 with $\alpha_{tl} = 1/(t+1)$, the gradient of the single-sequence loss $\mathcal{L}(k, j; \theta)$ in Eqn. (5) at initialization is

$$\nabla_{W_V}\mathcal{L}(k, j; \theta^{(0)}) = \sum_{t=1}^{T}\sum_{l=0}^{t} \alpha_{tl} g_t f_{x_l}^\top = \sum_{t=1}^{T-1} \frac{1}{t+1}\, g_t \big(\sum_{l=0}^{t} f_{x_l}\big)^\top \tag{11}$$

Token prefix sum. Define the number of feature in prefix at position t as

$$n_F(t) = \big|\{\tau, \ldots, \tau + L - 1\} \cap \{1, \ldots, t\}\big| = \begin{cases} 0, & t < \tau, \\ t - \tau + 1, & \tau \le t \le \tau + L - 1, \\ L, & t \ge \tau + L - 1. \end{cases} \tag{12}$$

Since position $t = 0$ is the text token $x_0 = K + J + k$ for visual feature class k and all non-feature positions carry the same background token $x_t = K + j$ for visual background class j, the prefix sum is

$$\sum_{l=0}^{t} f_{x_l} = f_{K+J+k} + n_F(t)\, f_k + \big(t - n_F(t)\big)\, f_{K+j}. \tag{13}$$

Expansion. Substituting Eqn. (13) into Eqn. (11) and expand $g_t = \bar{f} - f_{x_{t+1}}$ gives

$$\begin{aligned} \nabla_{W_V}\mathcal{L}(k, j; \theta^{(0)}) = & \sum_{t=1}^{T-1} \frac{1}{t+1} \bar{f} f_{K+J+k}^\top + \sum_{t=1}^{T-1} \frac{n_F(t)}{t+1} \bar{f} f_k^\top + \sum_{t=1}^{T-1} \frac{t - n_F(t)}{t+1} \bar{f} f_{K+j}^\top \\ & - \sum_{t=1}^{T-1} \frac{1}{t+1} f_{x_{t+1}} f_{K+J+k}^\top - \sum_{t=1}^{T-1} \frac{n_F(t)}{t+1} f_{x_{t+1}} f_k^\top - \sum_{t=1}^{T-1} \frac{t - n_F(t)}{t+1} f_{x_{t+1}} f_{K+j}^\top, \end{aligned}$$

where we decompose and group the terms into several sums of outer products. To investigate $f_{x_{t+1}}$ within the summation, we then split each resulting time sum according to the next token:

$$\mathcal{B} = \{1, \ldots, T-1\} \setminus \{\tau - 1, \ldots, \tau + L - 2\}, \quad \mathcal{F} = \{\tau - 1, \ldots, \tau + L - 2\}, \tag{14}$$

where on $\mathcal{B}$ we have $f_{x_{t+1}} = f_{K+j}$, while on $\mathcal{F}$ we have $f_{x_{t+1}} = f_k$. Grouping coefficients of the same outer products yields

$$\begin{aligned} \sum_{t=1}^{T-1} \frac{1}{t+1} f_{x_{t+1}} f_{K+J+k}^\top &= \Big(\sum_{t\in\mathcal{B}} \frac{1}{t+1}\Big) f_{K+j} f_{K+J+k}^\top + \Big(\sum_{t\in\mathcal{F}} \frac{1}{t+1}\Big) f_k f_{K+J+k}^\top, \\ \sum_{t=1}^{T-1} \frac{n_F(t)}{t+1} f_{x_{t+1}} f_k^\top &= \Big(\sum_{t\in\mathcal{B}} \frac{n_F(t)}{t+1}\Big) f_{K+j} f_k^\top + \Big(\sum_{t\in\mathcal{F}} \frac{n_F(t)}{t+1}\Big) f_k f_k^\top, \\ \sum_{t=1}^{T-1} \frac{t - n_F(t)}{t+1} f_{x_{t+1}} f_{K+j}^\top &= \Big(\sum_{t\in\mathcal{B}} \frac{t - n_F(t)}{t+1}\Big) f_{K+j} f_{K+j}^\top + \Big(\sum_{t\in\mathcal{F}} \frac{t - n_F(t)}{t+1}\Big) f_k f_{K+j}^\top. \end{aligned}$$

So we plug group the terms in the summation according to their left factors in the outer product and get

$$\begin{aligned}\nabla_{W_V}\mathcal{L}(k,j;\theta^{(0)}) = \bar{f}\Big(&A_{K+J+k}f_{K+J+k}^\top + A_k f_k^\top + A_{K+j}f_{K+j}^\top\Big)\\ - f_{K+j}\Big(&B_{K+J+k}f_{K+J+k}^\top + B_k f_k^\top + B_{K+j}f_{K+j}^\top\Big)\\ - f_k\Big(&C_{K+J+k}f_{K+J+k}^\top + C_k f_k^\top + C_{K+j}f_{K+j}^\top\Big),\end{aligned}$$

where the coefficients can be interpreted as the learned association strengths from specific input features to target features, mediated by the linear transformation W_V prior to their aggregation via the attention mechanism. A_{K+J+k}, A_k, A_{K+j} means the contribution from f_{K+J+k}, f_k, f_{K+j} to $\bar{f}$. B_{K+J+k}, B_k, B_{K+j} means the contribution from f_{K+J+k}, f_k, f_{K+j} to f_{K+j}. C_{K+J+k}, C_k, C_{K+j} means the contribution from f_{K+J+k}, f_k, f_{K+j} to f_k. We can further derive the closed form of these coefficient as follows:

$$\begin{aligned}A_{K+J+k} &= \sum_{t=1}^{T-1}\frac{1}{t+1}, \quad A_k = \sum_{t=1}^{T-1}\frac{n_F(t)}{t+1}, \quad A_{K+j} = \sum_{t=1}^{T-1}\frac{t-n_F(t)}{t+1},\\ B_{K+J+k} &= \sum_{t\in\mathcal{B}}\frac{1}{t+1}, \quad B_k = \sum_{t\in\mathcal{B}}\frac{n_F(t)}{t+1}, \quad B_{K+j} = \sum_{t\in\mathcal{B}}\frac{t-n_F(t)}{t+1},\\ C_{k+J+k} &= \sum_{t\in\mathcal{F}}\frac{1}{t+1}, \quad C_k = \sum_{t\in\mathcal{F}}\frac{n_F(t)}{t+1}, \quad C_{K+j} = \sum_{t\in\mathcal{F}}\frac{t-n_F(t)}{t+1},\end{aligned} \tag{15}$$

which is a results of direct computation from the previous expansion.

From the closed form in Eqn. (6), we have

$$\nabla_{W_V}\mathcal{L}_{\text{vis}}(\theta^{(0)}) = \sum_{k=1}^{K}p_k\sum_{j=1}^{J}q_j\,\nabla_{W_V}\mathcal{L}(k,j;\theta^{(0)}).$$

Thus, we conclude the proof of Lemma 4. □

E.2. Step 2: Asymptotic Oder of Coefficients

In this step, we derive the asymptotic order of the embedding coefficients in Lemma 4, which is crucial for our subsequent analysis.

Proposition 5 (Asymptotic Order of B, C coefficients)**.** *Under assumptions in Theorem 3, the asymptotic order of coefficients $B_{K+J+k}, B_k, B_{K+j}, C_{K+J+k}, C_k, C_{K+j}$ in Lemma 4 are*

$$\begin{aligned}&B_{K+J+k} = \Theta(\log T), \quad B_k = \Theta((\log T)^2), \quad B_{K+j} = \Theta(T),\\ &C_{K+J+k} = \Theta\Big(\tfrac{\log T}{T^\beta}\Big), \quad C_k = \Theta\Big(\tfrac{(\log T)^2}{T^\beta}\Big), \quad C_{K+j} = \Theta(\log T).\end{aligned} \tag{16}$$

Proof. Let $H_n = \sum_{i=1}^{n} 1/i$ denote the n-th harmonic number, and $H_n = \log n + \gamma + O(1/n)$.

Order of C_{K+J+k}. From the closed form of coefficient in Eqn. (15), we have

$$C_{K+J+k} = \sum_{t\in\mathcal{F}}\frac{1}{t+1} = \sum_{r=0}^{L-1}\frac{1}{\tau+r} = \frac{L}{\tau}(1+o(1)) = \Theta\Big(\frac{\log T}{T^\beta}\Big),$$

where the second equation comes from expansion using Eqn. (14), the third and fourth equation come from assumption $L = \Theta(\log T), \tau = \Theta(T^\beta)$.

Order of B_{K+J+k}. From the closed form in Eqn. (15), we have

$$B_{K+J+k} = \sum_{t\in\mathcal{B}}\frac{1}{t+1} = \sum_{t=1}^{T-1}\frac{1}{t+1} - \sum_{t=\tau-1}^{\tau+L-2}\frac{1}{t+1} = (H_{\tau-1}-H_1) + (H_T - H_{\tau+L-1}),$$

where the second equation comes from Eqn. (14), and the third equation rewrites the terms as a harmonic sum. Expand it via the harmonic approximation,

$$B_{K+J+k} = \log\tau + O(1) + \log T - \log(\tau + L - 1) + O(1) = \Theta(\log T),$$

where the last equation comes from assumption $L = \Theta(\log T), \tau = \Theta(T^\beta)$.

Order of C_k. From the closed form in Eqn. (15), we have

$$C_k = \sum_{t\in\mathcal{F}} \frac{n_F(t)}{t+1} = \sum_{t=\tau}^{\tau+L-2} \frac{t-\tau+1}{t+1} = (L-1) - \tau\sum_{r=1}^{L-1}\frac{1}{\tau+r},$$

where the second equation uses the results in Eqn. (12) and Eqn. (14), the third equation is a direct expansion. Using Taylor expansion $1/(\tau+r) = 1/\tau - r/\tau^2 + O(r^2/\tau^3)$, we have

$$\sum_{r=1}^{L-1}\frac{1}{\tau+r} = \frac{L-1}{\tau} - \frac{L(L-1)}{2\tau^2} + O\left(\frac{L^3}{\tau^3}\right),$$

and plug it in the previous results, we have

$$C_k = (L-1) - \tau\sum_{r=1}^{L-1}\frac{1}{\tau+r} = \frac{L(L-1)}{2\tau} + O\left(\frac{L^3}{\tau^2}\right) = \Theta\left(\frac{(\log T)^2}{T^\beta}\right),$$

where the last equation comes from assumption $L = \Theta(\log T), \tau = \Theta(T^\beta)$.

Order of B_k. From the closed form in Eqn. (15), we have

$$B_k = \sum_{t\in\mathcal{B}}\frac{n_F(t)}{t+1} = \sum_{t=\tau+L-1}^{T-1}\frac{L}{t+1} = L(H_T - H_{\tau+L-1}) = L\big(\log T - \log(\tau+L-1) + O(1)\big),$$

where the second equation uses the results in Eqn. (12) and Eqn. (14), the third equation rewrites the terms as harmonic sum, and the fourth equation uses the approximation results of harmonic sum. Using the assumption that $L = \Theta(\log T), \tau = \Theta(T^\beta)$, we have $\log(\tau+L-1) = \beta\log T + O(1)$, so plug it in the previous results, we have

$$B_k = L\big((1-\beta)\log T + O(1)\big) = \Theta\big((\log T)^2\big).$$

Order of C_{K+j}. From the closed form in Eqn. (15), we have

$$C_{K+j} = \sum_{t\in\mathcal{F}}\frac{t-n_F(t)}{t+1} = \sum_{t=\tau-1}^{\tau+L-2}\frac{t-n_F(t)}{t+1} = \frac{\tau-1}{\tau+1} + \sum_{t=\tau}^{\tau+L-2}\frac{t-t+\tau-1}{t+1} = (\tau-1)\sum_{t=\tau-1}^{\tau+L-2}\frac{1}{t+1},$$

where the second equation uses the results in Eqn. (14), the third equation uses the results in Eqn. (12) the fourth equation is a direct computation. Reorganize the summation, we have

$$C_{K+j} = (\tau-1)\sum_{r=0}^{L-1}\frac{1}{\tau+r} = (\tau-1)\cdot\frac{L}{\tau}(1+o(1)) = \Theta(L) = \Theta(\log T),$$

where the last two equations use the assumption that $L = \Theta(\log T), \tau = \Theta(T^\beta)$.

Order of B_{K+j}. From the closed form in Eqn. (15), we have

$$\begin{aligned}
B_{K+j} &= \sum_{t\in\mathcal{B}}\frac{t-n_F(t)}{t+1} \\
&= \sum_{t=1}^{\tau-2}\frac{t}{t+1} + \sum_{t=\tau+L-1}^{T-1}\frac{t-L}{t+1} \\
&= (T-L-1) - (H_{\tau-1} - H_1) - (L+1)(H_T - H_{\tau+L-1}) \\
&= \Theta(T)
\end{aligned}$$

where the second equation uses results in Eqn. (12) and Eqn. (14), and the third equation rewrites the terms as harmonic sum. The fourth equation uses assumption $L = \Theta(\log T), \tau = \Theta(T^\beta)$, to get $T - L - 1 = \Theta(T)$, $H_{\tau-1} - H_1 = \Theta(\log T)$, and $(L+1)(H_T - H_{\tau+L-1}) = \Theta\big((\log T)^2\big)$, thus we can conclude that $B_{K+j} = \Theta(T)$.

Thus, we conclude the proof of Proposition 5. □

E.3. Step 3: Asymptotic Sandwich Property

We now verify that, under the parameter regime of Theorem 3, q_j is asymptotically sandwiched between the two critical ratios C_{K+J+k}/B_{K+J+k} and C_k/B_k. Specifically, we show that $q_j B_{K+J+k} > C_{K+J+k}$ and $q_j B_k < C_k$ for sufficiently large T. These conditions are pivotal for the subsequent analysis, as they ensure that the projection via $W_V^{(1)}$ maps f_{K+J+k} predominantly to background tokens, while preserving f_k as itself.

Proposition 6 (Asymptotic sandwich on q_j)**.** *Under the assumptions of Theorem 3, we know that $q_j = 1/J$, with $J = \lfloor \tau \log T / cL \rfloor$ for some constant $c \in \big(1, \frac{L-1}{2(1-\beta)L}\big)$. Then, for sufficiently large T,*

$$\frac{C_{K+J+k}}{B_{K+J+k}} < q_j < \frac{C_k}{B_k}.$$

Proof. Using the exact expressions from Proposition 5, we compute the two ratio bounds:

$$\frac{C_{K+J+k}}{B_{K+J+k}} = \frac{\frac{L}{\tau}(1+o(1))}{\log T + O(1)} = \frac{L}{\tau \log T} + o(1),$$
$$\frac{C_k}{B_k} = \frac{L(L-1)/(2\tau) + O(L^3/\tau^2)}{L(1-\beta)\log T + O(1)} = \frac{L-1}{2(1-\beta)\tau \log T} + o(1),$$

where for both row, the first equations come from the closed form in proof of Proposition 5, and the second equations use the asymptotic results in Eqn. (16). Since $q_j = 1/J$, with $J = \lfloor \tau \log T / cL \rfloor$, when $c \in \big(1, \frac{L-1}{2(1-\beta)L}\big)$, we know that

$$\frac{L}{\tau \log T} < \frac{cL}{\tau \log T} < \frac{L-1}{2(1-\beta)L} \frac{L}{\tau \log T} = \frac{L-1}{2(1-\beta)\tau \log T}, \tag{17}$$

so $C_{K+J+k}/B_{K+J+k} < q_j < C_k/B_k$ holds for sufficiently large T. Equivalently, we know $q_j B_{K+J+k} > C_{K+J+k}$ and $q_j B_k < C_k$.

Because when $\beta \in (1/2, 1)$, the upper bound satisfies $\frac{L-1}{2(1-\beta)L} \to \frac{1}{2(1-\beta)} > 1$ as $L \to \infty$, such a constant c exists. Because $L = \Theta(\log T)$, the interval is non-empty for sufficiently large T.

Thus, we conclude the proof of Proposition 5. □

E.4. Step 4: One-Step Update of W_V

Now we compute $W_V^{(1)}$, and analyze the projection of embeddings into the value space via it.

Lemma 7 (Projection of embeddings via $W_V^{(1)}$)**.** *After the first step update in Algorithm 2, the value matrix is*

$$W_V^{(1)} = -\eta_V \sum_{k=1}^{K} p_k \sum_{j=1}^{J} q_j \nabla_{W_V} \mathcal{L}(k, j; \theta^{(0)}).$$

Projecting f_{K+j_1}, f_{k_1}, and f_{K+J+k_1}, with feature class $k_1 \in [K]$, background class $j_1 \in [J]$, into the value space via $W_V^{(1)}$ yields:

$$W_V^{(1)} f_{K+j_1} = \eta_V \big[-q_{j_1} A_{K+j_1} \bar{f} + q_{j_1} B_{K+j_1} f_{K+j_1} + q_{j_1} C_{K+j_1} \sum_k p_k f_k \big],$$
$$W_V^{(1)} f_{k_1} = \eta_V \big[-p_{k_1} A_{k_1} \bar{f} + p_{k_1} \sum_j q_j B_{k_1} f_{K+j} + p_{k_1} C_{k_1} f_{k_1} \big],$$
$$W_V^{(1)} f_{K+J+k_1} = \eta_V \big[-p_{k_1} A_{K+J+k_1} \bar{f} + p_{k_1} \sum_j q_j B_{K+J+k_1} f_{K+j} + p_{k_1} C_{K+J+k_1} f_{k_1} \big].$$

Proof. The first step gradient descent in Algorithm 2 using population loss will help us get

$$W_V^{(1)} = 0 - \eta_V \nabla_{W_V} \mathcal{L}_{\text{vis}}(\theta^{(0)}) = -\eta_V \sum_{k=1}^{K} p_k \sum_{j=1}^{J} q_j \nabla_{W_V} \mathcal{L}(k, j; \theta^{(0)}),$$

which is a direct expansion using the results in Lemma 4. For $k_1 \in [K], j_1 \in [J]$, We can analyze the projection of an embedding into the value space via the weight matrix $W_V^{(1)}$.

For $W_V^{(1)} f_{K+j_1}$, we have

$$\begin{aligned} W_V^{(1)} f_{K+j_1} &= -\eta_V \sum_k p_k \sum_j q_j \nabla_{W_V} \mathcal{L}(k, j; \theta^{(0)}) f_{K+j_1} \\ &= \eta_V \left[-q_{j_1} A_{K+j_1} \bar{f} + q_{j_1} B_{K+j_1} f_{K+j_1} + q_{j_1} C_{K+j_1} \textstyle\sum_k p_k f_k \right], \end{aligned}$$

where the second equality comes from expansion of $\nabla_{W_V} \mathcal{L}(k, j; \theta^{(0)})$ using Lemma 4, and direct computation using the orthonormality of embeddings and property $\sum_{k=1}^{K} p_k = 1, \sum_{j=1}^{J} q_j = 1$.

Similarly, we apply the same technique to expand $W_V^{(1)} f_{k_1}, W_V^{(1)} f_{K+J+k_1}$ and can get

$$\begin{aligned} W_V^{(1)} f_{k_1} &= \eta_V [-p_{k_1} A_{k_1} \bar{f} + p_{k_1} \sum_j q_j B_{k_1} f_{K+j} + p_{k_1} C_{k_1} f_{k_1}], \\ W_V^{(1)} f_{K+J+k_1} &= \eta_V [-p_{k_1} A_{K+J+k_1} \bar{f} + p_{k_1} \sum_j q_j B_{K+J+k_1} f_{K+j} + p_{k_1} C_{K+J+k_1} f_{k_1}]. \end{aligned}$$

Thus, we conclude the proof of Lemma 7. □

E.5. Step 5: Logits and Probabilities under $W_V^{(1)}$

After knowing how $W_V^{(1)}$ works, we can analyze the logits and probabilities of the model after the frist update, which is crucial for the analysis of the second update.

Lemma 8 (Logits of tokens after First step update)**.** *With the parameter after the first step optimization $\theta^{(1)} = (W_V^{(1)}, 0)$, when the input is an image with visual feature class $k_1 \in [K]$ and visual background class $j_1 \in [J]$, i.e. $x_0 = K + J + k_1$ and $x_{1:T} \in \mathcal{I}_{k_1, j_1}$, $\text{logit}_t \in \mathbb{R}^{K+J}$ satisfies*

$$\begin{aligned} \text{logit}_t[k] &= \tilde{A}(t) + \frac{\eta_V}{t+1} p_k q_{j_1} \big(t - n_F(t)\big) C_{K+j_1}, \quad k \in [K], k \neq k_1, \\ \text{logit}_t[k_1] &= \tilde{A}(t) + \frac{\eta_V}{t+1} p_{k_1} \Big(q_{j_1} \big(t - n_F(t)\big) C_{K+j_1} + C_{K+J+k_1} + n_F(t) C_{k_1} \Big), \\ \text{logit}_t[K+j] &= \tilde{A}(t) + \frac{\eta_V}{t+1} q_j \Big[p_{k_1} \big(B_{K+J+k_1} + n_F(t) B_{k_1} \big) \Big], \quad j \in [J], j \neq j_1, \\ \text{logit}_t[K+j_1] &= \tilde{A}(t) + \frac{\eta_V}{t+1} q_{j_1} \Big[p_{k_1} \big(B_{K+J+k_1} + n_F(t) B_{k_1} \big) \Big] + \frac{\eta_V}{t+1} (t - n_F(t)) q_{j_1} B_{K+j_1}, \end{aligned}$$

where the shared terms above also has a closed form

$$\tilde{A}(t) = \frac{\eta_V}{(t+1)(K+J)} \Big[p_{k_1} \big(A_{K+J+k_1} + n_F(t) A_{k_1} \big) + (t - n_F(t)) q_{j_1} A_{K+j_1} \Big].$$

The coefficients A_i, B_i, C_i, for $i \in [K+J]$, comes from Lemma 4.

Proof. With the input sequence $x_{0:T}$, because $W_{QK} = 0$, $s_{tl} = e_t^\top W_{QK} f_{x_l = 0}$ for $0 \leq l \leq t$, so $\alpha_{tl} = 1/(t+1)$, using the

definition of o_t with $W_V = W_V^{(1)}$, we have

$$
\begin{aligned}
o_t &= \sum_{l=0}^{t} \frac{1}{t+1} W_V^{(1)} f_{x_l} \\
&= \frac{1}{t+1} W_V^{(1)} \Big(f_{K+J+k_1} + n_F(t) f_{k_1} + \big(t - n_F(t)\big) \Big) \\
&= \frac{\eta_V}{t+1} \Big(- p_{k_1} A_{K+J+k_1} \bar{f} + p_{k_1} \sum_j q_j B_{K+J+k_1} f_{K+j} + p_{k_1} C_{K+J+k_1} f_{k_1} \\
&\quad + n_F(t)(-p_{k_1} A_{k_1} \bar{f} + p_{k_1} \sum_j q_j B_{k_1} f_{K+j} + p_{k_1} C_{k_1} f_{k_1}) \\
&\quad + \big(t - n_F(t)\big) \big(- q_{j_1} A_{K+j_1} \bar{f} + (\sum_k p_k q_{j_1} B_{K+j_1}) f_{K+j_1} + \sum_k p_k q_{j_1} C_{K+j_1} f_k \big) \Big),
\end{aligned}
$$

where the second equation use the results in Eqn. (13), and the third equation use the results of $W_V^{(1)} f_{k_1}, W_V^{(1)} f_{K+J+k_1}, W_V^{(1)} f_{K+j_1}$ in Lemma 7. Reorganize it with Eqn. (12), we have

$$
\begin{aligned}
o_t = \frac{\eta_V}{t+1} \Bigg\{ & \bar{f} \Bigg(- \Big[p_{k_1} \big(A_{K+J+k_1} + n_F(t) A_{k_1} \big) + \big(t - n_F(t)\big) q_{j_1} A_{K+j_1} \Big] \Bigg) \\
& + \sum_k f_k \Bigg(p_k \Big[\mathbb{I}\{k = k_1\} \big(C_{K+J+k_1} + n_F(t) C_{k_1} \big) + (t - n_F(t)) q_{j_1} C_{K+j_1} \Big] \Bigg) \\
& + \sum_j f_{K+j} \Bigg(q_j \Big[p_{k_1} \big(B_{K+J+k_1} + n_F(t) B_{k_1} \big) \Big] + \mathbb{I}\{j = j_1\} \big(t - n_F(t)\big) q_{j_1} B_{K+j_1} \Bigg) \Bigg\},
\end{aligned}
$$

where we group the terms according to the embeddings. Because $\bar{f} = \sum_{i=1}^{K+J} f_i/(K+J)$, we can further group the coefficient of embeddings and rewrite o_t as linear combination of embeddings. Due to the weight tying W_{lm}, we have $o_t = \sum_{i=1}^{K+J} \text{logit}_t[i] f_i$, where

$$
\begin{aligned}
\text{logit}_t[k] &= \tilde{A}(t) + \frac{\eta_V}{t+1} p_k q_{j_1} \big(t - n_F(t)\big) C_{K+j_1}, \quad k \in [K], k \neq k_1, \\
\text{logit}_t[k_1] &= \tilde{A}(t) + \frac{\eta_V}{t+1} p_{k_1} \Big(q_{j_1} \big(t - n_F(t)\big) C_{K+j_1} + C_{K+J+k_1} + n_F(t) C_{k_1} \Big), \\
\text{logit}_t[K+j] &= \tilde{A}(t) + \frac{\eta_V}{t+1} q_j \Big[p_{k_1} \big(B_{K+J+k_1} + n_F(t) B_{k_1} \big) \Big], \quad j \in [J], j \neq j_1, \\
\text{logit}_t[K+j_1] &= \tilde{A}(t) + \frac{\eta_V}{t+1} q_{j_1} \Big[p_{k_1} \big(B_{K+J+k_1} + n_F(t) B_{k_1} \big) \Big] + \frac{\eta_V}{t+1} (t - n_F(t)) q_{j_1} B_{K+j_1},
\end{aligned}
$$

where we directly decompose the factors and $\mathbb{I}\{k = k_1\}, \mathbb{I}\{j = j_1\}$. The common factor of every embedding is provided by coefficient of $\bar{f}$, which is

$$
\tilde{A}(t) = \frac{\eta_V}{(t+1)(K+J)} \Big[p_{k_1} \big(A_{K+J+k_1} + n_F(t) A_{k_1} \big) + (t - n_F(t)) q_{j_1} \sum_k p_k A_{K+j_1} \Big],
$$

which will be eliminated in the computation of softmax function. Thus, we conclude the proof of Lemma 8. □

Lemma 9 (Concentration of probability on the background token after first step)**.** *Under assumptions in Theorem 3, after the first step optimization, for all $t \in \{1, \ldots, T-1\}$ and all three phases of $n_F(t)$ in Eqn.* (12)*, with an input sequence with visual feature class $k_1 \in [K]$ and visual background class $j_1 \in [J]$, i.e. $x_0 = K + J + k_1, x_{1:T} \in \mathcal{I}_{k_1, j_1}$, the probability satisfies*

$$
P_t[i] = \begin{cases} e^{-\Theta(T^{1-\beta})}, & i \neq K + j_1, \\ 1 - \Theta\big(T^{\beta} e^{-\Theta(T^{1-\beta})}\big), & i = K + j_1. \end{cases}
$$

The dominant logit gap is driven by $\mathcal{C}_{K+j_1} = \Theta(T^{1-\beta})$, while all other logits are $\Theta((\log T)^3/T^{\beta})$ or smaller.

Proof. From Lemma 8, we know the closed form of logits after the first step with an input sequence with feature k_1 and background j_1 in Lemma 8. Note that $\tilde{A}(t)$ is the same for every logits, so when we feed $\mathrm{logit}_t^{(1)}$ into softmax, the term $\tilde{A}(t)$ will be eliminated, which means

$$\mathtt{sm}(\mathrm{logit}_t)[a] = \mathtt{sm}\Big(\big[\mathrm{logit}_t[1] - \tilde{A}(t), \cdots, \mathrm{logit}_t[K+J] - \tilde{A}(t)\big]^\top\Big)[a],$$

which holds because we divide $\tilde{A}(t)$ from the numerator and denominator, and then use the definition of softmax function. Therefore, the probability $P_t[i]$ only determined by $\mathrm{logit}_t[i] - \tilde{A}(t)$. Now we analysis their order at different region using the results in Eqn. (16).

When $t < \tau$, from Eqn. (12), $n_F(t) = 0$, we plug them in results in Lemma 8 and get

$$\begin{aligned}
\mathrm{logit}_t[k] - \tilde{A}(t) &= \frac{\eta_V}{t+1} p_k q_{j_1} t C_{K+j_1}, \quad k \in [K], k \neq k_1, \\
\mathrm{logit}_t[k_1] - \tilde{A}(t) &= \frac{\eta_V}{t+1} p_{k_1} (q_{j_1} t C_{K+j_1} + C_{K+J+k_1}), \\
\mathrm{logit}_t[K+j] - \tilde{A}(t) &= \frac{\eta_V}{t+1} q_j p_{k_1} B_{K+J+k_1}, \quad j \in [J], j \neq j_1, \\
\mathrm{logit}_t[K+j_1] - \tilde{A}(t) &= \frac{\eta_V}{t+1} q_{j_1} p_{k_1} B_{K+J+k_1} + \frac{\eta_V}{t+1} t q_{j_1} B_{K+j_1}.
\end{aligned}$$

Then we plug in the order of the coefficients computed in Eqn. (16), and use the assumption in Theorem 3 with direct computation, finally we get

$$\mathrm{logit}_t[i] - \tilde{A}(t) = \begin{cases} \Theta\big(\frac{\log T}{T^\beta}\big), & i = k \in [K], k \neq k_1, \\ \Theta\big(\frac{\log T}{T^\beta}\big), & i = k_1, \\ \Theta\big(\frac{\log T}{(t+1)T^\beta}\big), & i = K+j, j \in [J], j \neq j_1, \\ \Theta(T^{1-\beta}), & i = K+j_1. \end{cases}$$

Similarly, from Eqn. (12), when $\tau \leq t \leq \tau + L - 1$, $t - n_F(t) = \tau - 1 = \Theta(1)$, we plug them in Lemma 8 and use the coefficient order in Eqn. (16) with assumption in Theorem 3, we can get

$$\mathrm{logit}_t[i] - \tilde{A}(t) = \begin{cases} \Theta\big(\frac{\log T}{T^\beta}\big), & i = k \in [K], k \neq k_1, \\ \Theta\big(\frac{\log T}{T^\beta}\big), & i = k_1, \\ \Big[\Theta\big(\frac{(\log T)^2}{T^{2\beta}}\big), \Theta\big(\frac{(\log T)^3}{T^{2\beta}}\big)\Big], & i = K+j, j \in [J], j \neq j_1, \\ \Theta(T^{1-\beta}), & i = K+j_1, \end{cases}$$

where those for $i = K + j, j \neq j_1$ lies in an interval because from Eqn. 12, when $\tau \leq t \leq \tau + L - 1$, $n_F(t) \in [1, L]$, and $L = \Theta(\log T)$.

Similarly, from Eqn. (12), when $t \geq \tau + L - 1$, $n_F(t) = L$, $t - n_F = t - L$, t can be at most T, so we also plug them in Lemma 8 and use the coefficient order in Eqn. (16) with assumption in Theorem 3, we can get

$$\mathrm{logit}_t[i] - \tilde{A}(t) = \begin{cases} \Theta\big(\frac{\log T}{T^\beta}\big), & i = k \in [K], k \neq k_1, \\ \Theta\big(\frac{\log T}{T^\beta}\big), & i = k_1, \\ \Theta\big(\frac{(\log T)^3}{(t+1)T^\beta}\big), & i = K+j, j \in [J], j \neq j_1, \\ \Theta(T^{1-\beta}), & i = K+j_1. \end{cases}$$

Now feed them into softmax. Firstly, we can find that in different time region, $K + j_1$ always dominates, but we also need to consider that $J = \lfloor \tau \log T / cL \rfloor$. Therefore, for all t, we apply softmax function and get

$$\mathtt{sm}(\mathrm{logit}_t)[K+j_1] = \frac{1}{K \exp\big(-\Theta(T^{1-\beta})\big) + (J-1)\exp\big(-\Theta(T^{1-\beta})\big) + 1} = 1 - \Theta(T^\beta e^{-\Theta(T^{1-\beta})}),$$

where the second equation is a results of divide $\exp\big(\mathrm{logit}_t[K+j_1]-\tilde{A}(t)\big)$ from the numerator and denominator, and $\mathrm{logit}_t[K+j_1]-\tilde{A}(t)=\Theta(T^{1-\beta})$ for every $t\in[1,T-1]$ and dominate others. The third equation comes from direct computation. So for $i\neq K+j_1$, we have $P_t[i]=e^{-\Theta(T^{1-\beta})}$. Conclude the above results, for $t\in[1,T-1]$ we have

$$P_t[i]=\begin{cases} e^{-\Theta(T^{1-\beta})}, & i\neq K+j_1\\ 1-\Theta\big(T^{\beta}e^{-\Theta(T^{1-\beta})}\big), & i=K+j_1.\end{cases}$$

Thus, we conclude the proof of Lemma 8. □

E.6. Step 6: Population Gradient of W_{QK} after first Update

In this step, we can compute the population gradient of W_{QK}, which will be used in the second update.

Lemma 10 (Population gradient of W_{QK})**.** *At $\theta^{(1)}=(W_V^{(1)},0)$, the negative gradient of W_{QK} on a sample with visual feature class $k_1\in[K]$ and visual background class $j_1\in[J]$ is*

$$\begin{aligned}
&-\nabla_{W_{QK}}\mathcal{L}(k_1,j_1;\theta^{(1)})\\
&=\sum_{t=\tau-1}^{\tau+L-2}\gamma(k_1,j_1,t,K+J+k_1)\,e_tf_{K+J+k_1}^\top+\sum_{t=\tau-1}^{\tau+L-2}(\tau-1)\,\gamma(k,j,t,K+j)\,e_tf_{K+j}^\top\\
&\quad+\sum_{t=\tau-1}^{\tau+L-2}(t-\tau+1)\,\gamma(k_1,j_1,t,k_1)\,e_tf_{k_1}^\top+\sum_{t=1}^{T-1}\sum_{l=0}^{t}\xi(k_1,j_1,t,x_l)e_tf_{x_l}^\top,
\end{aligned}$$

where the scalar $\gamma(k_1,j_1,t,a)$ parameterd by class k_1,j_1, position t and token a satisfies

$$\begin{aligned}
&\gamma(k_1,j_1,t,K+j_1)\\
&=\frac{1}{(t+1)^2}\Big(p_{k_1}q_{j_1}B_{K+J+k_1}-p_{k_1}C_{K+J+k_1}\Big)+\frac{n_F(t)}{(t+1)^2}\Big(p_{k_1}q_{j_1}B_{k_1}-p_{k_1}C_{k_1}\Big)\\
&\quad-\frac{1+n_F(t)}{(t+1)^2}\Big(q_{j_1}B_{K+j_1}-p_{k_1}q_{j_1}C_{K+j_1}\Big),\\
&\gamma(k_1,j_1,t,k_1)\\
&=\frac{1}{(t+1)^2}\Big(p_{k_1}q_{j_1}B_{K+J+k_1}-p_{k_1}C_{K+J+k_1}\Big)-\frac{t+1-n_F(t)}{(t+1)^2}\Big(p_{k_1}q_{j_1}B_{k_1}-p_{k_1}C_{k_1}\Big)\\
&\quad+\frac{t-n_F(t)}{(t+1)^2}\Big(q_{j_1}B_{K+j_1}-p_{k_1}q_{j_1}C_{K+j_1}\Big),\\
&\gamma(k_1,j_1,t,K+J+k_1)\\
&=-\frac{t}{(t+1)^2}\Big(p_{k_1}q_{j_1}B_{K+J+k_1}-p_{k_1}C_{K+J+k_1}\Big)+\frac{n_F(t)}{(t+1)^2}\Big(p_{k_1}q_{j_1}B_{k_1}-p_{k_1}C_{k_1}\Big)\\
&\quad+\frac{t-n_F(t)}{(t+1)^2}\Big(q_{j_1}B_{K+j_1}-p_{k_1}q_{j_1}C_{K+j_1}\Big).
\end{aligned}$$

Another group of scaler $\xi(k_1,j_1,t,a)$ parameterd by class k_1,j_1, position t and token a satisfies: $\xi(k_1,j_1,t,K+j_1)=O(\frac{T}{t+1}e^{-\Theta(T^{1-\beta})})$, $\xi(k_1,j_1,t,K+J+k_1)=O(\frac{T}{t+1}e^{-\Theta(T^{1-\beta})})$, $\xi(k_1,j_1,t,k_1)=O(\frac{T}{t+1}e^{-\Theta(T^{1-\beta})})$, which are exponentially small.

Proof. Recall the deinition in Section D.1, for each position t, we have $\delta_t=P_t-y_t$ and $g_t=W_{\mathrm{lm}}\delta_t$. For a sequence with visual feature class $k_1\in[K]$, and visual background class $j_1\in[J]$, by Lemma 9, we know that the probability $P_t^{(1)}$ concentrates on $K+j_1$, which means

$$\delta_t=P_t-y_t=(1-\epsilon_{K+j_1})\mathbf{1}_{K+j_1}+\sum_{i\neq K+j_1}\epsilon_i\mathbf{1}_i-\mathbf{1}_{x_{t+1}},$$

where $\epsilon_{K+j_1} = \Theta(T^\beta e^{-\Theta(T^{1-\beta})})$, and $\epsilon_i = e^{-\Theta(T^{1-\beta})}, i \neq K + j_1$, and $\sum_{i=1}^{K+J} \epsilon_i = 1$. So we have

$$g_t = \begin{cases} f_{K+j_1} - f_{k_1} + \sum_{i \neq K+j_1} \epsilon_i f_i - \epsilon_{K+j_1} f_{K+j_1}, & x_{t+1} = k_1, \\ \sum_{i \neq K+j_1} \epsilon_i f_i - \epsilon_{K+j_1} f_{K+j_1}, & x_{t+1} = K + j_1, \end{cases} \tag{18}$$

which is direct expansion using weight tying language model head W_{lm}. Because all the $\epsilon_i, i \in [K + J]$ are exponentially small, when $x_{t+1} = k_1$, $g_t \approx f_{K+j_1} - f_{k_1}$; when $x_{t+1} = K + j_1$, $g_t \approx 0$, for sufficiently large T. We can divide them into the main part $f_{K+j_1} - f_{k_1}$ and the residual part $\sum_{i \neq K+j_1} \epsilon_i f_i - \epsilon_{K+j_1} f_{K+j_1}$.

Because $s_{tl} = e_t^\top W_{QK} f_{x_l} = 0$, the attention score is uniform, i.e. $\alpha_{tl} = 1/(t + 1)$, for all $0 \leq l \leq t$. So from Eqn. (8), we know that, at $(W_V, W_{QK}) = (W_V^{(1)}, 0)$, the negative of the exact per-sample gradient is

$$-\nabla_{W_{QK}} \mathcal{L} = \sum_{t=1}^{T-1} \sum_{l=0}^{t} \frac{1}{t+1} \left[g_t^\top \left(o_t - W_V^{(1)} f_{x_l} \right) \right] e_t f_{x_l}^\top, \tag{19}$$

which is a simple reverse of sign in the Eqn. (8). To compute the closed form of the gradient of W_{QK}, we should consider to decompose the term $g_t^\top \left(o_t - W_V^{(1)} f_{x_l} \right)$. Frisly, we expand the $o_t - W_V^{(1)} f_{x_l}$ part by considering 3 kinds of embeddings respectively, so we can get

$$\begin{aligned} o_t - W_V^{(1)} f_{K+j_1} &= W_V^{(1)} \Big(\frac{1}{t+1} f_{K+J+k_1} + \frac{n_F(t)}{t+1} f_{k_1} - \frac{1 + n_F(t)}{t+1} f_{K+j_1} \Big), \\ o_t - W_V^{(1)} f_{k_1} &= W_V^{(1)} \Big(\frac{1}{t+1} f_{K+J+k_1} - \frac{t + 1 - n_F(t)}{t+1} f_{k_1} + \frac{t - n_F(t)}{t+1} f_{K+j_1} \Big), \\ o_t - W_V^{(1)} f_{K+J+k_1} &= W_V^{(1)} \Big(- \frac{t}{t+1} f_{K+J+k_1} + \frac{n_F(t)}{t+1} f_{k_1} + \frac{t - n_F(t)}{t+1} f_{K+j_1} \Big), \end{aligned}$$

where the equation all comes from the direct computation using definition $o_t = \sum_{l=0}^{t} \alpha_{tl} W_V^{(1)} f_{x_l}$ with uniform attention $\alpha_{tl} = 1/(t + 1)$ and the results in Eqn. (13).

Then we should consider the inner product $g_t^\top \left(o_t - W_V^{(1)} f_{x_l} \right)$. But due to the structure of g_t in Eqn. (18), we need to consider the main part and the residual part seperately, and then merge them.

Main Part: From the structure of g_t in Eqn. (18), the main part $f_{K+j_1} - f_{k_1}$ exists when $t \in \mathcal{F}$ in Eqn. (14), because $x_{t+1} = k_1$. We firsly consider the quadratic form when $f_{K+j_1} - f_{k_1}$ work on $W_V^{(1)} f_{K+j_1}$ to simplify the computation. Combine the results in Lemma 7, we have

$$\begin{aligned} &(f_{K+j_1} - f_{k_1})^\top W_V^{(1)} f_{K+j_1} \\ &\quad = -\eta_V (f_{K+j_1} - f_{k_1})^\top \sum_k p_k q_{j_1} A_{K+j_1} \bar{f} + \eta_V (\sum_k p_k q_{j_1} B_{K+j_1}) (f_{K+j_1} - f_{k_1})^\top f_{K+j_1} \\ &\qquad + \eta_V \sum_k p_k q_{j_1} C_{K+j_1} (f_{K+j_1} - f_{k_1})^\top f_k \\ &\quad = \eta_V q_{j_1} B_{K+j_1} - \eta_V p_{k_1} q_{j_1} C_{K+j_1}, \end{aligned}$$

where the first equation is a direct expansion of the $W_V^{(1)} f_{K+j_1}$, and the second equation comes from the orthonormality of the embeddings. Similarly, using the same technique, we can get

$$\begin{aligned} (f_{K+j_1} - f_{k_1})^\top W_V^{(1)} f_{k_1} &= \eta_V p_{k_1} q_{j_1} B_{k_1} - \eta_V p_{k_1} C_{k_1}, \\ (f_{K+j_1} - f_{k_1})^\top W_V^{(1)} f_{K+J+k_1} &= \eta_V p_{k_1} q_{j_1} B_{K+J+k_1} - \eta_V p_{k_1} C_{K+J+k_1}. \end{aligned}$$

To simplify the computation, we denote the scaler $\gamma(k_1, j_1, t, a) = (f_{K+j} - f_k)^\top (o_t - W_V^{(1)} f_a)/(t + 1)$ to pack the term befor $e_t f_{x_l}^\top$ in Eqn. (19) for a sequence with feature class k_1 and background class j_1 when computing the loss, and $x_l = a$.

When $x_l = K + j_1$, we have

$$\begin{aligned}\gamma(k_1, j_1, t, K + j_1) &= \frac{1}{t+1}(f_{K+j_1} - f_{k_1})^\top (o_t - W_V^{(1)} f_{K+j_1}) \\ &= \frac{1}{(t+1)^2}(p_{k_1} q_{j_1} B_{K+J+k_1} - p_{k_1} C_{K+J+k_1}) + \frac{n_F(t)}{(t+1)^2}(p_{k_1} q_{j_1} B_{k_1} - p_{k_1} C_{k_1}) \\ &\quad - \frac{1 + n_F(t)}{(t+1)^2}(q_{j_1} B_{K+j_1} - p_{k_1} q_{j_1} C_{K+j_1}),\end{aligned}$$

where the second equation use the results of $o_t - W_V^{(1)} f_{K+j_1}$ and expand them with the closed form of quadratic form $(f_{K+j_1} - f_{k_1})^\top W_V^{(1)} f_{K+j_1}$, $(f_{K+j_1} - f_{k_1})^\top W_V^{(1)} f_{k_1}$ and $(f_{K+j_1} - f_{k_1})^\top W_V^{(1)} f_{K+J+k_1}$. Similarly, we can use the same technique to derive those form when $x_l = k_1$ and $x_l = K + J + k_1$ as follows

$$\begin{aligned}&\gamma(k_1, j_1, t, k_1) \\ &\quad = \frac{1}{(t+1)^2}(p_{k_1} q_{j_1} B_{K+J+k_1} - p_{k_1} C_{K+J+k_1}) - \frac{t + 1 - n_F(t)}{(t+1)^2}(p_{k_1} q_{j_1} B_{k_1} - p_{k_1} C_{k_1}) \\ &\qquad + \frac{t - n_F(t)}{(t+1)^2}(q_{j_1} B_{K+j_1} - p_{k_1} q_{j_1} C_{K+j_1}), \\ &\gamma(k_1, j_1, t, K + J + k_1) \\ &\quad = -\frac{t}{(t+1)^2}(p_{k_1} q_{j_1} B_{K+J+k_1} - p_{k_1} C_{K+J+k_1}) + \frac{n_F(t)}{(t+1)^2}(p_{k_1} q_{j_1} B_{k_1} - p_{k_1} C_{k_1}) \\ &\qquad + \frac{t - n_F(t)}{(t+1)^2}(q_{j_1} B_{K+j_1} - p_{k_1} q_{j_1} C_{K+j_1}).\end{aligned}$$

Residual Part: Because ϵ_i are exponentially small, so for any $i \in [K + J]$, we have $\epsilon_i f_i^\top W_V^{(1)} f_{K+J+k_1} = o(1)$, $\epsilon_i f_i^\top W_V^{(1)} f_{k_1} = o(1)$, and $\epsilon_i f_i^\top W_V^{(1)} f_{K+j_1} = o(1)$. We explicitly describe these in the part. For the residuals $\sum_{i \neq K+j_1} \epsilon_i f_i - \epsilon_{K+j_1} f_{K+j_1}$, we also compute the quadratic form made by it and $W_V^{(1)} f_{K+j_1}, W_V^{(1)} f_{K+J+k_1}, W_V^{(1)} f_{k_1}$. For $W_V^{(1)} f_{K+j_1}$, we have

$$\begin{aligned}&(\sum_{i \neq K+j_1} \epsilon_i f_i - \epsilon_{K+j_1} f_{K+j_1})^\top W_V^{(1)} f_{K+j_1} \\ &\quad = -\eta_V (\sum_{i \neq K+j_1} \epsilon_i f_i - \epsilon_{K+j_1} f_{K+j_1})^\top \sum_k p_k q_{j_1} A_{K+j_1} \bar{f} \\ &\qquad + \eta_V (\sum_k p_k q_{j_1} B_{K+j_1})(\sum_{i \neq K+j_1} \epsilon_i f_i - \epsilon_{K+j_1} f_{K+j_1})^\top f_{K+j_1} \\ &\qquad + \eta_V \sum_k p_k q_{j_1} C_{K+j_1} (\sum_{i \neq K+j_1} \epsilon_i f_i - \epsilon_{K+j_1} f_{K+j_1})^\top f_k \\ &\quad = -\epsilon_{K+j_1} \eta_V (\sum_k p_k q_{j_1} B_{K+j_1}) + \eta_V \sum_k \epsilon_k p_k q_{j_1} C_{K+j_1} \\ &\quad = -\Theta(T e^{-\Theta(T^{1-\beta})}),\end{aligned}$$

where the first equation comes from direct expansion of $W_V^{(1)} f_{K+j_1}$ using results in Lemma 7, the second equation uses the orthonormality of embeddings, and the third equation use the assumption in Theorem 3 and the order of the coefficient in Eqn. (16). Use the same technique, we have

$$\begin{aligned}(\sum_{i \neq K+j_1} \epsilon_i f_i - \epsilon_{K+j_1} f_{K+j_1})^\top W_V^{(1)} f_{k_1} &= \Theta\big((\log T)^2 e^{-\Theta(T^{1-\beta})}\big), \\ (\sum_{i \neq K+j_1} \epsilon_i f_i - \epsilon_{K+j_1} f_{K+j_1})^\top W_V^{(1)} f_{K+J+k_1} &= \Theta\big(\frac{\log T}{T^\beta} e^{-\Theta(T^{1-\beta})}\big).\end{aligned}$$

To simplify the computation, we denote a scaler similar to the previous $\gamma(k_1, j_1, t, a)$, which is $\xi(k_1, j_1, t, a) = (\sum_{i \neq K+j_1} \epsilon_i f_i - \epsilon_{K+j_1} f_{K+j_1})^\top (o_t - W_V^{(1)} f_a)/(t+1)$ to pack the term befor $e_t f_{x_l}^\top$ in Eqn. (19) for a sequence with feature class k_1 and background class j_1 when computing the loss, and $x_l = a$. And for different x_l, we have

$$\begin{aligned}\xi(k_1, j_1, t, K+j_1) &= \frac{1}{t+1}(\sum_{i \neq K+j_1} \epsilon_i f_i - \epsilon_{K+j_1} f_{K+j_1})^\top (o_t - W_V^{(1)} f_{K+j_1}) \\ &= O\big(\frac{T}{t+1} e^{-\Theta(T^{1-\beta})}\big),\end{aligned}$$

where the second equation uses the previous expansion results of $o_t - W_V^{(1)} f_{K+j_1}$ and plug in the results of the order of the previous quadratic form of the residual part $\sum_{i \neq K+j_1} \epsilon_i f_i - \epsilon_{K+j_1} f_{K+j_1}$. Use similar technique for $x_l = k_1$ and $x_l = K + J + k_1$, we have

$$\xi(k_1, j_1, t, k_1) = O\big(\frac{T}{t+1} e^{-\Theta(T^{1-\beta})}\big), \quad \xi(k_1, j_1, t, K+J+k_1) = O\big(\frac{T}{t+1} e^{-\Theta(T^{1-\beta})}\big).$$

For the negative gradient for general per sequence in Eqn. (19), we can fixed the input sequence with a visual feature class k_1 and visual background class j_1, and expand the summation using the structure of g_t in Eqn. (18) and the next-token structure in Eqn. (10), we have

$$\begin{aligned}-\nabla_{W_{QK}} \mathcal{L}(k_1, j_1, \theta^{(1)}) &= \sum_{t=1}^{T-1} \sum_{l=0}^{t} \frac{1}{t+1} \left[g_t^\top (o_t - W_V^{(1)} f_{x_l})\right] e_t f_{x_l}^\top \\ &= \sum_{t=\tau-1}^{\tau+L-2} \sum_{l=0}^{t} \big(\gamma(k_1, j_1, t, x_l) + \xi(k_1, j_1, t, x_l)\big) e_t f_{x_l}^\top \\ &\quad + \sum_{t=1}^{\tau-2} \sum_{l=0}^{t} \xi(k_1, j_1, t, x_l) e_t f_{x_l}^\top + \sum_{t=\tau+L-1}^{T-1} \sum_{l=0}^{t} \xi(k_1, j_1, t, x_l) e_t f_{x_l}^\top \\ &= \sum_{t=\tau-1}^{\tau+L-2} \sum_{l=0}^{t} \gamma(k_1, j_1, t, x_l) e_t f_{x_l}^\top + \sum_{t=1}^{T-1} \sum_{l=0}^{t} \xi(k_1, j_1, t, x_l) e_t f_{x_l}^\top,\end{aligned}$$

where the second equation is a direct expansion using Eqn. (10) and Eqn. (18). We can further expand the formula using the data model structure and get

$$\begin{aligned}&-\nabla_{W_{QK}} \mathcal{L}(k_1, j_1, \theta^{(1)}) \\ &= \sum_{t=\tau-1}^{\tau+L-2} \gamma(k_1, j_1, t, K+J+k_1)\, e_t f_{K+J+k_1}^\top + \sum_{t=\tau-1}^{\tau+L-2} (\tau-1)\, \gamma(k_1, j_1, t, K+j_1)\, e_t f_{K+j_1}^\top \\ &\quad + \sum_{t=\tau-1}^{\tau+L-2} (t-\tau+1)\, \gamma(k_1, j_1, t, k_1)\, e_t f_{k_1}^\top + \sum_{t=1}^{T-1} \sum_{l=0}^{t} \xi(k_1, j_1, t, x_l) e_t f_{x_l}^\top,\end{aligned}$$

where the equation comes from group of x_l, because at $\tau - 1 \leq t \leq \tau + L - 2$, the prefix contain 1 $K + J + k_1$ token, $\tau - 1$ $K + j_1$ token, and $t - \tau + 1$ k_1 token. The x_l in the last term is the l-th token in the sequence with feature and background k_1, j_1. We don't explicitly expand those because we will show that their effect is exponentially small.

Thus, we conclude the proof of Lemma 10. □

E.7. Step 7: One-Step Update of W_{QK}

With the population gradient of W_{QK}, we can compute the updated version $W_{QK}^{(1)}$ and analyze its properties.

Proposition 11 (Position-Feature correlation computed by $W_{QK}^{(1)}$)**.** *After the second update in Algorithm 2, the query-key matrix $W_{QK}^{(1)}$ will have the following properties when computes the position-feature correlation on a sampled sequence with*

visual feature class $k_1 \in [K]$ and visual background class $j_1 \in [J]$:

$$e_t^\top W_{QK}^{(1)} f_{K+j_1} = \eta_{QK} \sum_{k=1}^{K} p_k q_{j_1} \, (\tau - 1) \, \gamma(k, j_1, t, K + j_1) \mathbf{1}\{t \in \mathcal{F}\} + o(1),$$

$$e_t^\top W_{QK}^{(1)} f_{k_1} = \eta_{QK} \, p_{k_1} \sum_{j=1}^{J} q_j \, (t - \tau + 1) \, \gamma(k_1, j, t, k_1) \mathbf{1}\{t \in \mathcal{F}\} + o(1),$$

$$e_t^\top W_{QK}^{(1)} f_{K+J+k_1} = \eta_{QK} \, p_{k_1} \sum_{j=1}^{J} q_j \, \gamma(k_1, j, t, K + J + k_1) \mathbf{1}\{t \in \mathcal{F}\} + o(1),$$

where $\mathcal{F}$ is the set of positions whose subsequent token is a feature, as defined in Eqn. 14, and the $o(1)$ term is specifically of order $O\big(T^{2+\beta} e^{-\Theta(T^{1-\beta})}\big)$.

Proof. From the closed form of population loss in Eqn. (6), the population gradient is

$$\nabla_{W_{QK}} \mathcal{L}_{\text{vis}}(\theta^{(1)}) = \sum_{k=1}^{K} p_k \sum_{j=1}^{J} q_j \, \nabla_{W_{QK}} \mathcal{L}(k, j; \theta^{(1)}). \tag{20}$$

After the second update in Algorithm 2, and plug in the results in Lemma 10, we can have the following form

$$\begin{aligned} W_{QK}^{(1)} = -\eta_{QK} \sum_{k=1}^{K} p_k \sum_{j=1}^{J} q_j \nabla_{W_{QK}} \mathcal{L}(k, j; \theta^{(1)}) \gamma(k, j, t, K + J + k) \, e_t f_{K+J+k}^\top \\ + \sum_{t=\tau-1}^{\tau+L-2} (\tau - 1) \, \gamma(k, j, t, K + j) \, e_t f_{K+j}^\top + \sum_{t=\tau-1}^{\tau+L-2} (t - \tau + 1) \, \gamma(k, j, t, k) \, e_t f_k^\top \\ + \sum_{t=1}^{T-1} \sum_{l=0}^{t} \xi(k, j, t, x_l) e_t f_{x_l}^\top \Big], \end{aligned}$$

where the second step use the results in Lemma 10. Suppose the input sequence has feature k_1 and background j_1, from e_t, we find the attention between feature $f_{k_1}, f_{K+j_1}, f_{K+J+k_1}$.

For $e_t^\top W_{QK}^{(1)} f_{K+j_1}$, we directly plug in the expansion in Lemma 10 and get

$$\begin{aligned} & e_t^\top W_{QK}^{(1)} f_{K+j_1} \\ & = \eta_{QK} \sum_{k=1}^{K} p_k q_{j_1} \, (\tau - 1) \, \gamma(k, j_1, t, K + j_1) \mathbf{1}\{t \in \mathcal{F}\} + \eta_{QK} \sum_{k=1}^{K} p_k q_{j_1} O(T^2 e^{-\Theta(T^{1-\beta})}) \\ & = \eta_{QK} \sum_{k=1}^{K} p_k q_{j_1} \, (\tau - 1) \, \gamma(k, j_1, t, K + j_1) \mathbf{1}\{t \in \mathcal{F}\} + O(T^2 e^{-\Theta(T^{1-\beta})}), \end{aligned}$$

where the first equation also use the orthornormality of embeddings and the last step use the assumptions in Theorem 3, which are $\eta_{QK} = \Theta(T^\beta), q_j = 1/J = \Theta(T^{-\beta})$.

For $e_t^\top W_{QK}^{(1)} f_{k_1}$ and $e_t^\top W_{QK}^{(1)} f_{K+J+k_1}$, with the same technique and results in Lemma 10 and assumption in Theorem 3, we can derive

$$e_t^\top W_{QK}^{(1)} f_{k_1} = \eta_{QK} \sum_{k=1}^{K} p_k q_{j_1} \, (\tau - 1) \, \gamma(k, j_1, t, K + j_1) \mathbf{1}\{t \in \mathcal{F}\} + O(T^{2+\beta} e^{-\Theta(T^{1-\beta})}),$$

$$e_t^\top W_{QK}^{(1)} f_{K+J+k_1} = \eta_{QK} \, p_{k_1} \sum_{j=1}^{J} q_j \, \gamma(k_1, j, t, K + J + k_1) \mathbf{1}\{t \in \mathcal{F}\} + O(T^{2+\beta} e^{-\Theta(T^{1-\beta})}).$$

We can notice that the small error provided by the residual in g_t in Eqn. 18 is exponentially small, and at most $O\big(T^{2+\beta} e^{-\Theta(T^{1-\beta})}\big)$. Thus, we conclude the proof of proposition 11. □

E.8. Step 8: CFG Analysis by Time Regime

From Eqn. (4), we know applying CFG is equivalent to applying CFG to the attention output. We now analyze the CFG on the attention output $o_t^\lambda = o_t + \lambda(o_t - o_t^{\text{uc}})$ at the trained parameters $(W_V^{(1)}, W_{QK}^{(1)})$ for each time regime. Recall the settings above, for simplicity, we denote logit_t for $\text{logit}_t^{\text{cond}}$, and $\text{logit}_t^{\text{uc}}$ for $\text{logit}_t^{\text{uncond}}$. For clarity, but we denote the i-th element of $\text{logit}_t \in \mathbb{R}^{K+J}$ as $\text{logit}_t[i]$, which also apply for $\text{logit}_t^{\text{uc}}$ and logit_t^λ.

Lemma 12 (Background positions: $1 \le t \le \tau - 2$)**.** *When $1 \le t < \tau - 1$, suppose the input tokens are $\big(K + J + k_1, (K + j_1)^{\times t}\big)$, when $\lambda < 1 + o(1)$, where $o(1)$ denotes a positive vanishing term, the CFG logits satisfies $\arg\max_i \text{logit}_t^\lambda[i] = K + j_1$.*

Proof. From Proposition 11, when $1 \le t < \tau - 1$, suppose the input tokens are $[K + J + k_1, (K + j_1)^{\times t}]$, we have

$$o_t = \alpha_{t0} W_V^{(1)} f_{K+J+k_1} + \sum_{l=1}^{t} \alpha_{tl} W_V^{(1)} f_{K+j_1}, \quad o_t^{\text{uc}} = \sum_{l=1}^{t} \alpha_{tl}^{\text{uc}} W_V^{(1)} f_{K+j_1} = W_V^{(1)} f_{K+j_1},$$

The output of CFG with λ is

$$\begin{aligned} o_t^\lambda &= o_t + \lambda(o_t - o_t^{\text{uc}}) \\ &= (1+\lambda)\big(\alpha_{t0} W_V^{(1)} f_{K+J+k_1} + \sum_{l=1}^{t} \alpha_{tl} W_V^{(1)} f_{K+j_1}\big) - \lambda W_V^{(1)} f_{K+j_1} \\ &= \big(1 - (1+\lambda)\alpha_{t0}\big) W_V^{(1)} f_{K+j_1} + (1+\lambda)\alpha_{t0} W_V^{(1)} f_{K+J+k_1}, \end{aligned}$$

where the second equation comes from expansion of o_t, o_t^{uc}, and the last equation comes from direct computation. For attention score α_{t0}, $1 \le t \le \tau - 1$, we have

$$\alpha_{t0} = \frac{\exp(e_t^\top W_{QK}^{(1)} f_{K+J+k_1})}{\exp(e_t^\top W_{QK}^{(1)} f_{K+J+k_1}) + \sum_{s=1}^{t} \exp(e_s^\top W_{QK}^{(1)} f_{K+J+k_1})} = \frac{1}{t+1}(1 + o(1)),$$

where the last equation using the results in Proposition 11, that when not inside the feature block, i.e. $t \notin \mathcal{F}$, $e_t^\top W_{QK}^{(1)} f_{K+J+k_1} = o(1)$, where the error is exponentially small. Similarly, we know that $\alpha_{tl} = \big(1 + o(1)\big)/(t+1)$.

Therefore, we can know that the attention score is asymptotically uniform. With the attention score, we can further expand $W_V^{(1)} f_{K+j_1}, W_V^{(1)} f_{K+J+k_1}$ using results in Lemma 7, which can change o_t^λ into a linear combination of embeddings. Because we use weight tying language model head W_{lm}, the coefficient before each embedding in o_t^λ is the corresponing value in logit_t^λ. Observe that in Lemma 7: f_{K+j_1} has an additional positive term than other f_{K+j} in $W_V^{(1)} f_{K+j_1}$; f_{k_1} has an additional positive term than other f_k in $W_V^{(1)} f_{K+J+k_1}$. We can know that $\text{logit}_t^\lambda[K + j_1] > \text{logit}_t^\lambda[K + j], j \ne j_1$, $\text{logit}_t^\lambda[k_1] > \text{logit}_t^\lambda[k], k \ne k_1$, so we only need to compare those of f_{K+j_1} and f_{k_1}. By expansion using Lemma 7 and direct substraction of the coefficient, we have

$$\begin{aligned} \eta_V^{-1}\big(\text{logit}_t^\lambda[K + j_1] - \text{logit}_t^\lambda[k_1]\big) = &\big(1 - (1+\lambda)\alpha_{t0}\big)(q_{j_1} B_{K+j_1} - p_{k_1} q_{j_1} C_{K+j_1}) \\ &+ (1+\lambda)\alpha_{t0}(p_{k_1} q_{j_1} B_{K+J+k_1} - p_{k_1} C_{K+J+k_1}), \end{aligned}$$

where the shared factor in coefficient of $\bar{f}$ is eliminated. We know that $\text{logit}_t^\lambda[K + j_1] - \text{logit}_t^\lambda[k_1] > 0$ holds when

$$\begin{aligned} \lambda &< \frac{q_{j_1} B_{K+j_1} - p_{k_1} q_{j_1} C_{K+j_1}}{\alpha_{t0}(q_{j_1} B_{K+j_1} - p_{k_1} q_{j_1} C_{K+j_1} p_{k_1} q_{j_1} - p_{k_1} q_{j_1} B_{K+J+k_1} + p_{k_1} C_{K+J+k_1})} - 1 \\ &= t\big(1 + o(1)\big) \end{aligned}$$

where the first equality comes from direct computation to make the gap positive, and the second equality follows from the asymptotic orders of the coefficients in Eqn. (16), noting that B_{K+j_1} is the dominant term, as well as the previously established order of α_{t0}. The residual $o(1)$ here is a positive vanishing term. Therefore, when $\lambda < 1 + o(1)$, $K + j_1 = \arg\max_i \text{logit}_t^\lambda[i]$, so in greedy decoding at position $1 \le t < \tau - 1$, the model will predict $K + j_1$ in the next position.

Thus, we conclude the proof of Lemma 12. □

Lemma 13 (Transition position: $t = \tau - 1$)**.** *When $t = \tau - 1$, suppose the input tokens are $\big(K + J + k_1, (K + j_1)^{\times(\tau-1)}\big)$, when $\lambda > o(1)$, where $o(1)$ denotes a positive vanishing term, the CFG logits satisfies* $\arg\max_i \operatorname{logit}_t^{\lambda}[i] = k_1$.

Proof. Similar to Lemme 12, from Proposition 11, when $t = \tau - 1$, suppose the input tokens are $[K+J+k_1, (K+j_1)^{\times(\tau-1)}]$, we have

$$o_t = \alpha_{t0} W_V^{(1)} f_{K+J+k_1} + \sum_{l=1}^{\tau-1} \alpha_{tl} W_V^{(1)} f_{K+j_1}, \qquad o_t^{\text{uc}} = \sum_{l=1}^{\tau-1} \alpha_{tl}^{\text{uc}} W_V^{(1)} f_{K+j_1} = W_V^{(1)} f_{K+j_1},$$

The output of CFG with λ is

$$o_t^{\lambda} = o_t + \lambda(o_t - o_t^{\text{uc}}) = \big(1 - (1+\lambda)\alpha_{t0}\big) W_V^{(1)} f_{K+j_1} + \lambda \alpha_{t0} W_V^{(1)} f_{K+J+k_1},$$

where we use the same computation as Lemma 12. For attention score α_{t0}, $t = \tau - 1$, we have

$$\begin{aligned}
\alpha_{t0} &= \frac{\exp(e_t^\top W_{QK}^{(1)} f_{K+J+k_1})}{\exp(e_t^\top W_{QK}^{(1)} f_{K+J+k_1}) + \sum_{s=1}^{\tau-1} \exp(e_t^\top W_{QK}^{(1)} f_{K+j_1})} \\
&= \frac{1}{1 + (\tau-1)\exp\big(e_t^\top W_{QK}^{(1)} (f_{K+j_1} - f_{K+J+k_1})\big)}.
\end{aligned}$$

where we divide $\exp(e_t^\top W_{QK}^{(1)} f_{K+J+k_1})$ from numerator and denominator. So we need to focus on the order of $e_t^\top W_{QK}^{(1)} (f_{K+j_1} - f_{K+J+k_1})$. Using the results in Lemma 10, because $t = \tau - 1$, $n_F(t) = 0$, we can know that

$$\begin{aligned}
\gamma(k, j, t, K + j_1) &= \frac{1}{(t+1)^2}(p_{k_1} q_{j_1} B_{K+J+k_1} - p_{k_1} C_{K+J+k_1}) \\
&\quad - \frac{1}{(t+1)^2}(q_{j_1} B_{K+j_1} - p_{k_1} q_{j_1} C_{K+j_1}), \\
\gamma(k, j, t, K + J + k_1) &= -\frac{t}{(t+1)^2}(p_{k_1} q_{j_1} B_{K+J+k_1} - p_{k_1} C_{K+J+k_1}) \\
&\quad + \frac{t - n_F(t)}{(t+1)^2}(q_{j_1} B_{K+j_1} - p_{k_1} q_{j_1} C_{K+j_1}).
\end{aligned}$$

Using the results in Proposition 11 to compute the position-feature correlation, and substract those with f_{K+j_1} and f_{K+J+k_1}, we have

$$\begin{aligned}
&e_t^\top W_{QK}^{(1)} (f_{K+j_1} - f_{K+J+k_1}) \\
&\quad = \eta_{QK} \sum_{k=1}^{K} p_k q_{j_1} (\tau - 1) \gamma(k, j_1, t, K + j_1) - \eta_{QK} p_{k_1} \sum_{j=1}^{J} q_j \gamma(k_1, j, t, K + J + k_1) + o(1) \\
&\quad = -\Theta(T^{1-\beta}),
\end{aligned}$$

where the second equation uses the order in Eqn. (16) and the assumptions in Theorem 3 that $\tau = \Theta(T^{\beta}), q_j = \Theta(T^{-\beta})$. Therefore, we can get the attention score

$$\alpha_{t0} = \frac{1}{1 + \Theta(T^{\beta}) \exp\big(-\Theta(T^{1-\beta})\big)} = 1 + o(1),$$

where the first equation use the assumption that $\tau = \Theta(T^{\beta})$. So we can observe that the attention score for conditional model is concentrates on the text token at $t = 0$. With the attention score, we can further expand $W_V^{(1)} f_{K+j_1}, W_V^{(1)} f_{K+J+k_1}$ using results in Lemma 7, which can change o_t^{λ} into a linear combination of embeddings. Similar to Lemma 12, we compare the logits by comparing the coefficient before embeddings. Observe that in Lemma 7: f_{K+j_1} has an additional positive term than other f_{K+j} in $W_V^{(1)} f_{K+j_1}$; f_{k_1} has an additional positive term than other f_k in $W_V^{(1)} f_{K+J+k_1}$. At $t = \tau - 1$ we can

observe that $\mathrm{logit}_t^\lambda[K+j_1] > \mathrm{logit}_t^\lambda[K+j], j \neq j_1, \mathrm{logit}_t^\lambda[k_1] > \mathrm{logit}_t^\lambda[k], k \neq k_1$, so we only need to compare those of f_{K+j_1} and f_{k_1}. By expansion using Lemma 7 and direct substraction of the coefficient, we have

$$\eta_V^{-1}\big(\mathrm{logit}_t^\lambda[K+j_1] - \mathrm{logit}_t^\lambda[k_1]\big) = \big(1-(1+\lambda)\alpha_{t0}\big)(q_{j_1}B_{K+j_1} - p_{k_1}q_{j_1}C_{K+j_1}) + (1+\lambda)\alpha_{t0}(p_{k_1}q_{j_1}B_{K+J+k_1} - p_{k_1}C_{K+J+k_1}),$$

where the shared factor in coefficient of $\bar{f}$ is eliminated. We know that $\mathrm{logit}_t^\lambda[K+j_1] - \mathrm{logit}_t^\lambda[k_1] < 0$ holds when

$$\lambda > \frac{q_{j_1}B_{K+j_1} - p_{k_1}q_{j_1}C_{K+j_1}}{\alpha_{t0}(q_{j_1}B_{K+j_1} - p_{k_1}q_{j_1}C_{K+j_1}p_{k_1}q_{j_1} - B_{K+J+k_1} + p_{k_1}C_{K+J+k_1})} - 1 = o(1),$$

where the first equality comes from direct computation to make the gap positive, and the second equality follows from the asymptotic orders of the coefficients in Eqn. (16), noting that B_{K+j_1} is the dominant term, as well as the previously established order of α_{t0}. The $o(1)$ here is a positive vanishing term. Therefore, when $\lambda > o(1)$, $k_1 = \arg\max_i \mathrm{logit}_{\tau-1}^\lambda[i]$, so in greedy decoding, the model will predict k_1 in the next position, thus successully predict the feature tokens.

Thus, we conclude the proof of Lemma 13. □

Lemma 14 (Feature block: $\tau \leq t \leq \tau + L - 2$)**.** *When $\tau \leq t \leq \tau + L - 2$, suppose the input tokens are $\big(K+J+k_1, (K+j_1)^{\times(\tau-1)}, (k_1)^{\times(t-\tau+1)}\big)$, when $\lambda < 1 + o(1)$, where $o(1)$ denotes a positive vanishing term, the CFG logits satisfies* $\arg\max_i \mathrm{logit}_t^\lambda[i] = k_1$.

Proof. When $\tau \leq t \leq \tau + L - 2$, there are some feature token after transition. Suppose the input tokens are $\big(K+J+k_1, (K+j_1)^{\times(\tau-1)}, k_1^{\times(t-\tau+1)}\big)$, we have

$$o_t = \alpha_{t0}W_V^{(1)}f_{K+J+k_1} + \sum_{l=1}^{\tau-1}\alpha_{tl}W_V^{(1)}f_{K+j_1} + \sum_{l=\tau}^{t}\alpha_{tl}W_V^{(1)}f_{k_1},$$
$$o_t^{\mathrm{uc}} = \sum_{l=1}^{t}\alpha_{tl}^{\mathrm{uc}}W_V^{(1)}f_{K+j_1} + \sum_{l=\tau}^{t}\alpha_{tl}^{\mathrm{uc}}W_V^{(1)}f_{k_1},$$

which are direct expansion using the definition. Therefore, we know that

$$\begin{aligned} o_t^\lambda &= o_t + \lambda(o_t - o_t^{\mathrm{uc}}) \\ &= \big(-\lambda\sum_{l=1}^{\tau-1}\alpha_{tl}^{\mathrm{uc}} + (1+\lambda)\sum_{t=1}^{\tau-1}\alpha_{tl}\big)W_V^{(1)}f_{K+j_1} + \big(-\lambda\sum_{l=\tau}^{t}\alpha_{tl}^{\mathrm{uc}} + (1+\lambda)\sum_{l=\tau}^{t}\alpha_{tl}\big)W_V^{(1)}f_{k_1} \\ &\quad + (1+\lambda)\alpha_{t0}W_V^{(1)}f_{K+J+k_1}, \end{aligned}$$

where the second equation comes from direct expansion of o_t, o_t^{uc} and group according to embeddings. Now we need to compute the attention scores. Because of the structure, we group the attention score by the embeddings. Firstly, for conditional model, we have

$$\begin{aligned} \alpha_{t0} &= \frac{\exp(e_t^\top W_{QK}^{(1)}f_{K+J+k_1})}{\exp(e_t^\top W_{QK}^{(1)}f_{K+J+k_1}) + (\tau-1)\exp(e_t^\top W_{QK}^{(1)}f_{K+j_1}) + (t-\tau+1)\exp(e_t^\top W_{QK}^{(1)}f_{k_1})} \\ &= \frac{1}{1 + (\tau-1)\exp\big(e_t^\top W_{QK}^{(1)}(f_{K+j_1} - f_{K+J+k_1})\big) + (t-\tau+1)\exp\big(e_t^\top W_{QK}^{(1)}(f_{k_1} - f_{K+J+k_1})\big)}, \end{aligned}$$

where we divide $\exp(e_t^\top W_{QK}^{(1)}f_{K+J+k_1})$ from the numerator and denominator. Similarly, using the same technique, we have

$$\sum_{l=\tau}^{t}\alpha_{tl} = \frac{t-\tau+1}{\exp\big(e_t^\top W_{QK}^{(1)}(f_{K+J+k_1} - f_{k_1})\big) + (\tau-1)\exp\big(e_t^\top W_{QK}^{(1)}(f_{K+j_1} - f_{k_1})\big) + (t-\tau+1)},$$
$$\sum_{l=1}^{\tau-1}\alpha_{tl} = \frac{\tau-1}{\exp\big(e_t^\top W_{QK}^{(1)}(f_{K+J+k_1} - f_{K+j_1})\big) + (\tau-1) + (t-\tau+1)\exp\big(e_t^\top W_{QK}^{(1)}(f_{k_1} - f_{K+j_1})\big)},$$

where we also write the softmax function explicitly, but divide $\exp(e_t^\top W_{QK}^{(1)} f_{K+j_1})$ and $\exp(e_t^\top W_{QK}^{(1)} f_{k_1})$ respectively.

Similar to conditional model, but expand with a different softmax structure in unconditional model, we have

$$\begin{aligned}\sum_{l=\tau}^{t} \alpha_{tl}^{\text{uc}} &= \frac{\sum_{l=\tau}^{t} \exp(e_t^\top W_{QK}^{(1)} f_{k_1})}{\sum_{l=1}^{\tau-1} \exp(e_t^\top W_{QK}^{(1)} f_{K+j_1}) + \sum_{l=\tau}^{t} \exp(e_t^\top W_{QK}^{(1)} f_{k_1})} \\ &= \frac{t-\tau+1}{(\tau-1)\exp\left(e_t^\top W_{QK}^{(1)}(f_{K+j_1} - f_{k_1})\right) + (t-\tau+1)},\end{aligned}$$

where we divide $\exp(e_t^\top W_{QK}^{(1)} f_{k_1})$ from the numerator and denominator. Similarly, using the same technique, we have

$$\sum_{l=1}^{\tau-1} \alpha_{tl}^{\text{uc}} = \frac{\tau-1}{(\tau-1) + (t-\tau+1)\exp\left(e_t^\top W_{QK}^{(1)}(f_{k_1} - f_{K+j_1})\right)},$$

where we also write the softmax function explicitly, but divide $\exp(e_t^\top W_{QK}^{(1)} f_{K+j_1})$.

From Eqn. (12), we know that within $\tau \le t \le \tau + L - 1$, we have $n_F(t) = t - \tau + 1$, so using results in Lemma 10, we know that

$$\begin{aligned}&\gamma(k, j, t, K + j_1) \\ &\quad= \frac{1}{(t+1)^2}(p_{k_1} q_{j_1} B_{K+J+k_1} - p_{k_1} C_{K+J+k_1}) + \frac{t-\tau+1}{(t+1)^2}(p_{k_1} q_{j_1} B_{k_1} - p_{k_1} C_{k_1}) \\ &\qquad - \frac{t-\tau+2}{(t+1)^2}(q_{j_1} B_{K+j_1} - p_{k_1} q_{j_1} C_{K+j_1}), \\ &\gamma(k, j, t, k_1) \\ &\quad= \frac{1}{(t+1)^2}(p_{k_1} q_{j_1} B_{K+J+k_1} - p_{k_1} C_{K+J+k_1}) - \frac{\tau}{(t+1)^2}(p_{k_1} q_{j_1} B_{k_1} - p_{k_1} C_{k_1}) \\ &\qquad + \frac{\tau-1}{(t+1)^2}(q_{j_1} B_{K+j_1} - p_{k_1} q_{j_1} C_{K+j_1}), \\ &\gamma(k, j, t, K + J + k_1) \\ &\quad= -\frac{t}{(t+1)^2}(p_{k_1} q_{j_1} B_{K+J+k_1} - p_{k_1} C_{K+J+k_1}) + \frac{t-\tau+1}{(t+1)^2}(p_{k_1} q_{j_1} B_{k_1} - p_{k_1} C_{k_1}) \\ &\qquad + \frac{\tau-1}{(t+1)^2}(q_{j_1} B_{K+j_1} - p_{k_1} q_{j_1} C_{K+j_1}),\end{aligned}$$

where we substitute the value of $n_F(t)$ explicitly into γ. So from the structure of attention value, we need to compute the value of the different of position-feature correlation. So using the results in Proposition 11 to compute the position-feature correlation, and substract those with f_{K+j_1} and f_{K+J+k_1}, we have

$$\begin{aligned}&e_t^\top W_{QK}^{(1)}(f_{K+j_1} - f_{K+J+k_1}) \\ &\quad= \eta_{QK} \sum_{k=1}^{K} p_k q_{j_1} (\tau - 1)\gamma(k, j_1, t, K + j_1) - \eta_{QK} p_{k_1} \sum_{j=1}^{J} q_j \gamma(k_1, j, t, K + J + k_1) + o(1) \\ &\quad= -\Theta(T^{1-\beta}),\end{aligned}$$

where the second equation uses the asymptotic order in Eqn. (16) and the assumptions in Theorem 3 that $\tau = \Theta(T^\beta), q_j = \Theta(T^{-\beta})$. With the same technique, we can similarly get

$$e_t^\top W_{QK}^{(1)}(f_{K+j_1} - f_{k_1}) = -\Theta((t - \tau + 1)T^{1-\beta}).$$

However, the computation of $e_t^\top W_{QK}^{(1)}(f_{K+j_1} - f_{k_1})$ is more difficult, and we should firstly compute a intermediate results

$$
\begin{aligned}
&(t-\tau+1)\gamma(k_1,j,t,k_1) - \gamma(k_1,j,t,K+J+k_1)\\
=&\frac{(t-\tau)(\tau-1)}{(t+1)^2}\big(p_k q_{j_1} B_{K+j_1} - p_{k_1} q_{j_1} C_{K+j_1}\big) + \frac{(2t-\tau+1)}{(t+1)^2}\big(p_{k_1} q_{j_1} B_{K+J+k_1} - p_{k_1} C_{K+J+k_1}\big)\\
&- \frac{(t-\tau+1)(\tau+1)}{(t+1)^2}\Big(p_{k_1} q_{j_1} B_{k_1} - p_{k_1} C_{k_1}\Big)\\
=&\begin{cases}\Theta((t-\tau)T^{1-2\beta}), & \tau < t \le \tau+L-2,\\ O\big(\frac{(\log T)^2}{T^{2\beta}}\big), & t=\tau,\end{cases}
\end{aligned}
$$

where the first equation comes from direct expansion using previous explicit form of γ, and the second equation uses the asymptotic order in Eqn. (16) with the assumptions in Theorem 3 that $\tau = \Theta(T^\beta), q_j = \Theta(T^{-\beta})$. The case when $t = \tau$ is special because $t = \tau$ will make the dominant term B_{K+j_1} disappear. Therefore, expand $e_t^\top W_{QK}^{(1)}(f_{K+j_1} - f_{k_1})$ using Proposition 11, we have

$$
\begin{aligned}
&e_t^\top W_{QK}^{(1)}(f_{k_1} - f_{K+J+k_1})\\
&\quad= \eta_{QK} p_{k_1} \sum_{j=1}^{J} q_j\big((t-\tau+1)\gamma(k_1,j,t,k_1) - \gamma(k_1,j,t,K+J+k_1)\big) + o(1)\\
&\quad= \begin{cases}\Theta((t-\tau)T^{1-\beta}), & \tau < t \le \tau+L-2,\\ O\big(\frac{(\log T)^2}{T^{\beta}}\big), & t=\tau,\end{cases}
\end{aligned}
$$

where the last equation use the assumption in Theorem 3 that $\eta_{QK} = \Theta(T^\beta), q_j = \Theta(T^{-\beta})$. Next, we discuss different cases.

When $\tau < t \le \tau + L - 2$: From the assumption in Theorem 3 that $L = \Theta(\log T)$, we know that $0 \le t - \tau = O(\log T)$. So plug the results of order of the terms back to the attention score computed above, we have

$$
\sum_{l=\tau}^{t} \alpha_{tl}^{\text{uc}} = 1 + o(1), \quad \sum_{l=\tau}^{t} \alpha_{tl} = 1 + o(1), \quad \alpha_{t0} = o(1),
$$

where we also use the assumption in Theorem 3 that $\tau = \Theta(T^\beta)$, and the error $o(1)$ is exponentially small. Therefore, we can find that the attention score of conditional and unconditional model concentrate on the feature block $\tau \le l \le t$, so from the construction of o_t^λ, we have $o_t^\lambda \approx W_V^{(1)} f_{k_1}$ with exponentially small error, for $\lambda < 1 + o(1)$, for sufficiently large T. From Lemma 7 and order in Eqn. (16), we only need to compare feature k_1 and any background $j \in [J]$, so substract the coefficient of theem and consider the error, we can get the logit

$$
\eta_V^{-1}\big(\text{logit}_t^\lambda[k_1] - \text{logit}_t^\lambda[K+j]\big) = p_{k_1} q_j B_{k_1} - p_{k_1} C_{k_1} f_{k_1} + o(1) > 0,
$$

where we use the results of order in Eqn. (16) and the assumption in Theorem 3 that $q_j = \Theta(T^{-\beta})$. The error $o(1)$ here is exponentially small.

When $t = \tau$: Because $\sum_{l=\tau}^{t} \alpha_{tl}^{\text{uc}}$ doew not contain $e_t^\top W_{QK}^{(1)}(f_{k_1} - f_{K+J+k_1})$, we have the same results $\sum_{l=\tau}^{t} \alpha_{tl}^{\text{uc}} = 1 + o(1)$. From the assumption in Theorem 3 that $L = \Theta(\log T)$, we know that $t = \tau = O(T^\beta)$. So plug the results of order of the terms back to the attention score computed above, we have

$$
\begin{aligned}
\sum_{l=\tau}^{t} \alpha_{tl} &= \frac{t-\tau+1}{\exp\Big(O\big(\frac{(\log T)^2}{T^{2\beta}}\big)\Big) + (\tau-1)\exp\Big(-\Theta\big((t-\tau+1)T^{1-\beta}\big)\Big) + t-\tau+1} = \frac{1}{2} + o(1),\\
\alpha_{t0} &= \frac{1}{1 + (\tau-1)\exp\big(-\Theta(T^{1-\beta})\big) + (t-\tau+1)\exp\Big(O\big(\frac{(\log T)^2}{T^{2\beta}}\big)\Big)} = \frac{1}{2} + o(1),
\end{aligned}
$$

where the error $o(1)$ is exponentially small. Therefore, at $t = \tau$, the conditional model seperate half of the attention to the text token, and the other half still lies on the feature block. From the construction of o_t^λ, we have $o_t^\lambda \approx \big(1 -$

$(\lambda+1)/2\big)W_V^{(1)}f_{k_1}+\big((\lambda+1)/2\big)W_V^{(1)}f_{K+J+k_1}$ with exponentially small error under finite λ. For sufficiently large T, considering results in Lemma 7 and order in Eqn. (16), we only need to compare feature k_1 and any background $j\in[J]$, so substract the coefficient of theem and consider the error, we can get the logit

$$\eta_V^{-1}\big(\mathrm{logit}_t^\lambda[k_1]-\mathrm{logit}_t^\lambda[K+j]\big)=\frac{\lambda+1}{2}(p_{k_1}C_{K+J+k_1}-p_{k_1}C_{k_1}+p_{k_1}q_jB_{k_1}-p_{k_1}q_jB_{K+J+k_1})$$
$$+p_{k_1}C_{k_1}-p_{k_1}q_jB_{k_1}+o(1),$$

which comes from directly expansion using Lemma 7, and the error term is exponentially small. From Proposition 6, we know $q_jB_{k_1}<C_{k_1}, q_jB_{K+J+k_2}>C_{K+J+k_1}$. Because $\eta_V>0$, to get $\mathrm{logit}_t^\lambda[k_1]-\mathrm{logit}_t^\lambda[K+j]>0$, it should satisfies

$$\lambda<\frac{2(p_{k_1}C_{k_1}-p_{k_1}q_{j_1}B_{k_1})}{p_{k_1}C_{k_1}-p_{k_1}q_jB_{k_1}+p_{k_1}q_jB_{K+J+k_1}-p_{k_1}C_{K+J+k_1}}-1=1+o(1),$$

where the first equality comes from direct computation to make the gap position, and the second equality follows from the asymptotic orders of the coesfficients in Eqn. (16), noting that B_{K+j_1} is the dominant term. The $o(1)$ here is a positive vanishing term.

In conclusion, at position $\tau\le t<\tau+L-1$, when $\lambda<1+o(1)$, with positive vanishing term $o(1)$, $k_1=\arg\max_i \mathrm{logit}_t^\lambda[i]$, so in greedy decoding, the model will predict k_1 in the next position, thus successully predict the feature tokens.

Thus, we conclude the proof of Lemma 14. □

Lemma 15 (Post-feature positions: $t\ge\tau+L-1$)**.** *When $t\ge\tau+L-1$, suppose the input tokens are $\big(K+J+k_1,(K+j_1)^{\times(\tau-1)},k_1^{\times L},(K+j_1)^{\times(t-\tau-L+1)}\big)$, when $\lambda<1+o(1)$, the CFG logits satisfies* $\arg\max_i \mathrm{logit}_t^\lambda[i]=K+j_1$.

Proof. When $t\ge\tau+L-1$, the feature block completes, and the target tokens are the background tokens. Suppose the input tokens are $\big(K+J+k_1,(K+j_1)^{\times(\tau-1)},k_1^{\times L},(K+j_1)^{\times(t-\tau-L+1)}\big)$, from the definition of output, we have

$$o_t=\alpha_{t0}W_V^{(1)}f_{K+J+k_1}+(\sum_{l=1}^{\tau-1}\alpha_{tl}+\sum_{l=\tau+L}^{t}\alpha_{tl})W_V^{(1)}f_{K+j_1}+\sum_{l=\tau}^{\tau+L-1}\alpha_{tl}W_V^{(1)}f_{k_1},$$
$$o_t^{\mathrm{uc}}=(\sum_{l=1}^{\tau-1}\alpha_{tl}^{\mathrm{uc}}+\sum_{l=\tau+L}^{t}\alpha_{tl}^{\mathrm{uc}})W_V^{(1)}f_{K+j_1}+\sum_{l=\tau}^{\tau+L-1}\alpha_{tl}^{\mathrm{uc}}W_V^{(1)}f_{k_1},$$

where group the terms according to the embeddings for simplification. Therefore, we know that

$$\begin{aligned}o_t^\lambda&=o_t+\lambda(o_t-o_t^{\mathrm{uc}})\\&=\big(-\lambda(\sum_{l=1}^{\tau-1}\alpha_{tl}^{\mathrm{uc}}+\sum_{l=\tau+L}^{t}\alpha_{tl}^{\mathrm{uc}})+(1+\lambda)(\sum_{l=1}^{\tau-1}\alpha_{tl}+\sum_{l=\tau+L}^{t}\alpha_{tl})\big)W_V^{(1)}f_{K+j_1}\\&\quad+\big(-\lambda\sum_{l=\tau}^{\tau+L-1}\alpha_{tl}^{\mathrm{uc}}+(1+\lambda)\sum_{l=\tau}^{\tau+L-1}\alpha_{tl}\big)W_V^{(1)}f_{k_1}+(1+\lambda)\alpha_{t0}W_V^{(1)}f_{K+J+k_1}.\end{aligned}$$

Apply softmax function on the position-feature correlation on the sequence to compute the attention score, we have

$$\begin{aligned}\alpha_{t0}&=\frac{\exp(e_t^\top W_{QK}^{(1)}f_{K+J+k_1})}{\exp(e_t^\top W_{QK}^{(1)}f_{K+J+k_1})+(\tau-1)\exp(e_t^\top W_{QK}^{(1)}f_{K+j_1})+(t-\tau+1)\exp(e_t^\top W_{QK}^{(1)}f_{k_1})}\\&=\frac{1}{1+(\tau-1)\exp\big(e_t^\top W_{QK}^{(1)}(f_{K+j_1}-f_{K+J+k_1})\big)+(t-\tau+1)\exp\big(e_t^\top W_{QK}^{(1)}(f_{k_1}-f_{K+J+k_1})\big)}\\&=\frac{1}{1+t}(1+o(1)),\end{aligned}$$

where the second equation divides $\exp(e_t^\top W_{QK}^{(1)} f_{K+J+k_1})$ from the numerator and denomenator, and the thrid equation comes from the properties in Proposition 11 that when $t \notin \mathcal{F}$, we have $\exp(e_t^\top W_{QK}^{(1)} f_{K+J+k_1}) = o(1)$, $\exp(e_t^\top W_{QK}^{(1)} f_{K+j_1}) = o(1)$, $\exp(e_t^\top W_{QK}^{(1)} f_{k_1}) = o(1)$, and the $o(1)$ is exponentially small. Similarly, applying the same technique to expand softmax function, we have

$$\sum_{l=\tau}^{\tau+L-1} \alpha_{tl} = \frac{L}{1+t}(1+o(1)), \quad \sum_{l=1}^{\tau-1} \alpha_{tl} + \sum_{l=\tau+L}^{t} \alpha_{tl} = \frac{t-L}{1+t}\big(1+o(1)\big).$$

For unconditional model, the only difference is the structure in softmax function, so we can also directly expand them and use the properties in Proposition 11, and get

$$\sum_{l=\tau}^{\tau+L-1} \alpha_{tl}^{\text{uc}} = \frac{L}{(t-L)\exp\big(e_t^\top W_{QK}^{(1)}(f_{K+j_1} - f_{k_1})\big) + L} = \frac{L}{t}(1+o(1)),$$

where the first equation divides $\exp(e_t^\top W_{QK}^{(1)} f_{k_1})$ from the numerator and denomenator, and the second equation apply the fact that $\exp(e_t^\top W_{QK}^{(1)} f_{K+j_1}) = o(1)$ and $\exp(e_t^\top W_{QK}^{(1)} f_{k_1}) = o(1)$. Similarly, we apply the same technique to the background token block, we have

$$\sum_{l=1}^{\tau-1} \alpha_{tl}^{\text{uc}} + \sum_{l=\tau+L}^{t} \alpha_{tl}^{\text{uc}} = \frac{t-L}{t}(1+o(1)),$$

where the error is exponentially small. Therefore, we can know that the attention pattern is almost uniform. So we can plug in the attention score to the output applied CFG

$$\begin{aligned} o_t^\lambda = &\big(-\lambda\frac{t-L}{t} + (1+\lambda)\frac{t-L}{t+1} + o(1)\big)W_V^{(1)} f_{K+j_1} \\ &+ \big(-\lambda\frac{L}{t} + (1+\lambda)\frac{L}{t+1} + o(1)\big)W_V^{(1)} f_{k_1} + (1+\lambda)\frac{1}{t+1}(1+o(1))W_V^{(1)} f_{K+J+k_1}. \end{aligned}$$

We can further expand $W_V^{(1)} f_{K+j_1}, W_V^{(1)} f_{K+J+k_1}, W_V^{(1)} f_{k_1}$ using results in Lemma 7, which can change o_t^λ into a linear combination of embeddings. Because we use weight tying language model head W_{lm}, the coefficient before each embedding in o_t^λ is the corresponing value in logit_t^λ. Observe that in Lemma 7: f_{K+j_1} has an additional positive term than other f_{K+j} in $W_V^{(1)} f_{K+j_1}$; f_{k_1} has an additional positive term than other f_k in $W_V^{(1)} f_{K+J+k_1}$. We can know that $\text{logit}_t^\lambda[K+j_1] > \text{logit}_t^\lambda[K+j], j \neq j_1$, $\text{logit}_t^\lambda[k_1] > \text{logit}_t^\lambda[k], k \neq k_1$, so we only need to compare those of f_{K+j_1} and f_{k_1}. By expansion using Lemma 7 and direct substraction of the coefficient, we have

$$\begin{aligned} &\eta_V^{-1}\big(\text{logit}_t^\lambda[K+j_1] - \text{logit}_t^\lambda[k_1]\big) \\ &= \big((1-\lambda)\frac{t-L}{t} + \lambda\frac{t-L}{t+1}\big)q_{j_1}B_{K+j_1} + \lambda\frac{1}{t+1}p_{k_1}q_{j_1}B_{K+J+k_1} \\ &\quad + \big((1-\lambda)\frac{L}{t} + \lambda\frac{L}{t+1}\big)p_{k_1}q_{j_1}B_{k_1} - \big((1-\lambda)\frac{t-L}{t} + \lambda\frac{t-L}{t+1}\big)p_{k_1}q_{j_1}C_{K+j_1} \\ &\quad - \lambda\frac{1}{t+1}p_{k_1}C_{K+J+k_1} - \big((1-\lambda)\frac{L}{t} + \lambda\frac{L}{t+1}\big)p_{k_1}C_{k_1} + o(1) \\ &= \big((1-\lambda)\frac{t-L}{t} + \lambda\frac{t-L}{t+1}\big)q_{j_1}B_{K+j_1} + o(T^{1-\beta}), \end{aligned}$$

where the first equation comes from direct expansion using Lemma 7, and the shared factor in coefficient of $\bar{f}$ is eliminated. The the second equality holds because of Eqn. (16). Because $\eta_V > 0$, we know that $\text{logit}_t^\lambda[K+j_1] - \text{logit}_t^\lambda[k_1] < 0$ holds for sufficiently large T when

$$-\lambda\frac{t-L}{t} + (1+\lambda)\frac{t-L}{t+1} < 0,$$

which is a direct computation by setting the gap negative. This result is equivalent to $\lambda < t+1$. Therefore, when $\lambda < 1 + o(1) < t$, at $t > \tau + L - 1$, $k_1 = \arg\max_i \text{logit}_t^{\lambda}[i]$, so in greedy decoding, the model will predict k_1 in the next position, thus successully predict the feature tokens.

Thus, we conclude the proof of Lemma 15. □

E.9. Step 9: Loss Bound and Proof of Theorem 3

Notice that the input prefix discuss in the lemmas of Section E.8 are the prefix of sequences from data distribution. So we can reuse the results in these lemmas to analyze the training loss.

Proposition 16 (Minimal Training Loss)**.** *Under assumption in Theorem 3, running the conditional model with parameter $\theta^{(2)} = (W_V^{(1)}, W_{QK}^{(1)})$, the per-position losses of the conditional model are:*

- *For $t \notin [\tau - 1, \tau + L - 2]$ (background): $-\log P_t[I_{t+1}] = \Theta\big(T^{\beta} e^{-\Theta(T^{1-\beta})}\big)$.*
- *For $t = \tau - 1$ (transition): $-\log P_t[k_1] = \Theta(\log T)$.*
- *For $t \in [\tau, \tau + L - 2]$ (feature block): $-\log P_t[k_1] = \Theta(\log T)$.*

The population loss of the model is $\mathcal{L}_{vis}(\theta^{(2)}) = \Theta\big((\log T)^2/T\big)$, which is asymptotically small.

Proof. We firstly consider an input sequence with feature class k_1 and background class j_1. Note that in Lemma 12, Lemma 13, Lemma 14, and Lemma 15, the input are prefix of all standard sequence with feature class k_1 and background class j_1. So we can directly use these lemma to analyze the loss with sequence from the data distribution. Here we denote $P_t[k]$ as the probability of token k in probability $P_t \in \mathbb{R}^{K+J}$.

For $t \notin \mathcal{F}$: by Lemma 12 and Lemma 15, we know that when $t \notin \mathcal{F}$, the attention score is asymptotically uniform, i.e. $\alpha_{tl} = 1/(t+1) + o(1)$, with an exponential error. So the probability the same with those in Lemma 9 (before the second update) for sufficiently large T. So we know that

$$-\log P_t[x_{t+1}] = -\log P_t[K + j_1] = -\log\Big(1 - \Theta\big(T^{\beta} e^{-\Theta(T^{1-\beta})}\big)\Big) = \Theta\big(T^{\beta} e^{-\Theta(T^{1-\beta})}\big),$$

where the first equation uses $x_{t+1} = K + j_1$ when $t \in \mathcal{F}$, the second equation use results in Lemma 9 because the attention score is asymptotic uniform with exponentially small, and the third equation directly computation use $-\log(1-x) = x + O(x^2)$ when x is small.

For $t \in \mathcal{F}$: by Lemma 14, we know that when $\tau < t \le \tau + L - 2$, the attention concentrates on the feature block, i.e. $\sum_{l=\tau}^{t} \alpha_{tl} = 1 + o(1)$, with exponentially small error, so we have

$$\begin{aligned} o_t &= W_V^{(1)} f_{k_1} + o(1)(W_V^{(1)} f_{K+J+k_1} + W_V^{(1)} f_{K+j_1}) \\ &= \eta_V[-p_{k_1} A_{k_1} \bar{f} + p_{k_1} \sum_j q_j B_{k_1} f_{K+j} + p_{k_1} C_{k_1} f_{k_1}] + o(1)(W_V^{(1)} f_{K+J+k_1} + W_V^{(1)} f_{K+j_1}), \end{aligned}$$

where the second equation is a direct expansion. Use the results in Eqn. (16), because $f_k, k \neq k_1$ only show up in $\bar{f}$, so the logits is $\text{logit}_t[k] = -\eta_V p_{k_1} A_{k_1}/(K+J) + o(1), k \neq k_1$, and this comes from the exponentially small $o(1)$. Since the coefficient of $\bar{f}$ contribute the same to all the embeddings, we have the results of other logits

$$\begin{aligned} \text{logit}_t[k_1] &= \frac{\eta_V p_{k_1} A_{k_1}}{K+J} + \Theta\big(\frac{L(L-1)}{2\tau}\big), \\ \text{logit}_t[K+j] &= \frac{\eta_V p_{k_1} A_{k_1}}{K+J} + \Theta\big(\frac{cL^2(1-\beta)}{\tau}\big), j \in [J] \end{aligned}$$

which are results plug in the asymptotic order in Eqn. (16) and the assumption in Theorem 3. So we can apply softmax

function to compute the probability

$$P_t[k_1] = \frac{\exp\big(\text{logit}_t[k_1]\big)}{\sum_{j=1}^{J}\exp\big(\text{logit}_t[K+j]\big) + \sum_{k=1}^{K}\exp\big(\text{logit}_t[k]\big)\big)}$$
$$= \frac{1}{\Theta(T^{\beta})\exp\Big(\Theta\big(\frac{cL^2(1-\beta)}{\tau}\big) - \Theta\big(\frac{L(L-1)}{2\tau}\big)\Big) + \Theta\Big(\exp\big(-T^{-\beta}(\log T)^2\big)\Big) + \Theta(1)},$$

where the second equation divide $\exp\big(\text{logit}_t[k_1]\big)$ from the numerator and denominator, and then plug in the asymptotic results. Therefore, we can directly compute the log probability, which is $-\log P_t[k_1] = \Theta(\log T)$.

For $t = \tau$, also from Lemma 13, we know that the conditional model distribute half of attention to the text token, i.e. $\alpha_{t0} = \frac{1}{2}(1+o(1))$ and $\alpha_{t\tau} = \frac{1}{2}(1+o(1))$, with exponentially small error $o(1)$, so we have

$$o_t = \frac{1}{2}\eta_V[-p_{k_1}A_{k_1}\bar{f} + p_{k_1}\sum_j q_j B_{k_1} f_{K+j} + p_{k_1}C_{k_1}f_{k_1}]$$
$$+ \frac{1}{2}\eta_V\big[-p_{k_1}A_{K+J+k_1}\bar{f} + p_{k_1}\sum_j q_j B_{K+J+k_1} f_{K+j} + p_{k_1}C_{K+J+k_1}f_{k_1}\big] + o(1)W_V^{(1)}f_{K+j_1},$$

which is a direct results from Lemma 7. Similarly, use the results in Eqn. (16) and notice that the coefficient of $\bar{f}$ contribute the same to all the embeddings, we have

$$\text{logit}_t[k_1] = -\frac{\eta_V(p_{k_1}A_{k_1} + p_{k_1}A_{K+J+k_1})}{2(K+J)} + \Theta\big(\frac{L(L-1)}{4\tau}\big),$$
$$\text{logit}_t[K+j] = -\frac{\eta_V(p_{k_1}A_{k_1} + p_{k_1}A_{K+J+k_1})}{2(K+J)} + \Theta\big(\frac{cL^2(1-\beta)}{2\tau}\big), j \in [J],$$
$$\text{logit}_t[k] = -\frac{\eta_V(p_{k_1}A_{k_1} + p_{k_1}A_{K+J+k_1})}{2(K+J)} + o(1), k \in [K]$$

which are results plug in the asymptotic order in Eqn. (16) and the assumption in Theorem 3. So we can apply softmax function to compute the probability

$$P_t[k_1] = \frac{1}{\Theta(T^{\beta})\exp\Big(\Theta\big(\frac{cL^2(1-\beta)}{2\tau}\big) - \Theta\big(\frac{L(L-1)}{4\tau}\big)\Big) + \Theta\Big(\exp\big(-T^{-\beta}(\log T)^2\big)\Big) + \Theta(1)},$$

where we also apply the same technique to divide $\exp\big(\text{logit}_t[k_1]\big)$ from the numerator and denominator, and then plug in the asymptotic results. Therefore, we can directly compute the log probability, which is $-\log P_t[k_1] = \Theta(\log T)$.

So in conclusion, we know that for $t \in \mathcal{F}$, i.e. $\tau \le t \le \tau + L - 2$, every position contributes to loss with $-\log P_t[I_{t+1}] = -\log P_t[k_1] = \Theta(\log T)$.

For $t = \tau - 1$: From Lemma 13, we know that when $t = \tau - 1$, the attention concentrates on the text token, i.e. $\alpha_{t0} = 1 + o(1)$, with exponentially small error, so we have

$$o_t = W_V^{(1)}f_{K+J+k_1} + o(1)W_V^{(1)}(f_{K+J+k_1} + f_{K+j_1} - f_{K+J+k_0})$$

where the equation comes from direct expansion and the error $o(1)$ is exponentially samll. Similarly, applying results from Lemma 7 and using the results in Eqn. (16) and notice that the coefficient of $\bar{f}$ contribute the same to all the embeddings, we have so the logits is

$$\text{logit}_t[k_1] = -\frac{\eta_V p_{k_1}A_{K+J+k_1}}{K+J} + \Theta\big(\frac{L}{\tau}\big),$$
$$\text{logit}_t[K+j] = -\frac{\eta_V p_{k_1}A_{K+J+k_1}}{K+J} + \Theta\big(\frac{cL}{\tau}\big), j \in [J],$$
$$\text{logit}_t[k] = -\frac{\eta_V p_{k_1}A_{K+J+k_1}}{K+J} + o(1), k \in [K]$$

which are results plug in the asymptotic order in Eqn. (16) and the assumption in Theorem 3. So we can apply softmax function to compute the probability

$$P_t[k_1] = \frac{1}{\Theta(T^\beta)\exp\Big(\Theta\big((c-1)T^{-\beta}\log T\big)\Big) + \Theta\big(\exp(-T^{-\beta}\log T)\big) + \Theta(1)}$$

where we also apply the same technique to divide $\exp\big(\mathrm{logit}_t[k_1]\big)$ from the numerator and denominator, and then plug in the asymptotic results. Therefore, we can directly compute the log probability, which is $-\log P_t[k_1] = \Theta(\log T)$.

In conclusion, we can compute the loss of models on a sequence with feature k_1 and background j_1

$$\begin{aligned}\mathcal{L}(k_1, j_1; \theta^{(2)}) &= -\frac{1}{T-1}\sum_{t=1}^{T-1}\log P_t[x_{t+1}] \\ &= \frac{1}{T-1}\Big((T-L-1)\Theta\big(T^\beta e^{-\Theta(T^{1-\beta})}\big) + L\Theta(\log T)\Big) \\ &= \Theta\Big(\frac{(\log T)^2}{T}\Big),\end{aligned}$$

where these first equations use the definition in Eqn. (5), the second equation plug in the previous asymptotic results of the error at every position, and the last one use the assumption in Theorem 3 to expand. Therefore, compute the population loss use definition in Eqn. (6), we have

$$\mathcal{L}_{\text{vis}}(\theta^{(2)}) = \sum_{k=1}^{K} p_k \sum_{j=1}^{J} q_j \mathcal{L}(k_1, j_1; \theta^{(2)}) = \Theta\Big(\frac{(\log T)^2}{T}\Big).$$

Notably, the loss is primarily driven by the errors on the feature block, which scale as $\Theta(\log T)$ at each position. In contrast, the error on the background block is exponentially small, as the model consistently predicts the background tokens with high accuracy.

Thus, we conclude the proof of Proposition 16. □

E.10. Step 10: Correctly Generation with CFG

With the analyze of CFG in Section E.8, we can know analyze the correctness of generation using suitable guidance scale. Here, we use $\hat{I}^{\lambda}_{k,j}[t_1 : t_2]$ to denote the token between position t_1, t_2 in the generated image $\hat{I}^{\lambda}_{k,j}$ with feature class k and background j.

Lemma 17 (Correctness of Generation with CFG)**.** *With guidance scale $o(1) < \lambda < 1 + o(1)$, where $o(1)$ denotes a positive vanishing term, the trained Transformer generates sequence $\hat{I}^{\lambda}_{k,j}(\theta^{(2)}) = I_{k,j}$, for all $k \in [K], j \in [J]$.*

Proof. Consider feature class k_1 and background class j_1, follow the inference procedure defined in Sec. 4, starting with prefix $(K + J + k_1, K + j_1)$.

Before feature: From Lemma 12, when $1 \le t < \tau - 1$, we know that with $o(1) < \lambda < 1 + o(1)$, given input $\big(K + J + k_1, (K + j_1)^{\times t}\big)$, the greedy decoding satisfies $\arg\max_i \mathrm{logit}_t[i] = K + j_1$. So the background feature at $t = 2, \cdots, \tau - 1$ can be correctly generated, i.e. the generated image satisfies $\hat{I}^{\lambda}_{k_1,j_1}[1 : \tau - 1] = \big((K + j_1)^{\times(\tau-1)}\big)$.

Transition: From Lemma 13, when $t = \tau - 1$, we know that with $o(1) < \lambda < 1 + o(1)$, given input $\big(K + J + k_1, (K + j_1)^{\times(\tau-1)}\big)$, the greedy decoding satisfies $\arg\max_i \mathrm{logit}_t[i] = k_1$. So the required feature at $t = \tau$ can be correctly generated, i.e. we have $\hat{I}^{\lambda}_{k_1,j_1}[1 : \tau] = \big((K + j_1)^{\times(\tau-1)}, k_1\big)$.

Feature: From Lemma 14, when $\tau \le t \le \tau + L - 2$, we know that with $o(1) < \lambda < 1 + o(1)$, given input $\big(K + J + k_1, (K + j_1)^{\times(\tau-1)}, (k_1)^{\times(t-\tau+1)}\big)$, the greedy decoding will output k_1, i.e. $\arg\max_i \mathrm{logit}_t[i] = k_1$. So the required feature at $t = \tau + 1, \cdots, \tau + L - 1$ can be correctly generated, i.e. the generated image satisfies $\hat{I}^{\lambda}_{k_1,j_1}[1 : \tau + L - 1] = \big((K + j_1)^{\times(\tau-1)}, (k_1)^{\times L}\big)$.

After feature: From Lemma 15, when $\tau \le t \le \tau + L - 2$, we know that with $o(1) < \lambda < 1 + o(1)$, given input $\big(K + J + k_1, (K + j_1)^{\times(\tau-1)}, (k_1)^{\times(t-\tau+1)}, (K + j_1)^{\times(t-\tau-L+1)}\big)$, the greedy decoding satisfies $\arg\max_i \text{logit}_t[i] = K + j_1$. So the required feature at $t = \tau + L, \cdots, T$ can be correctly generated, i.e. we have $\hat{I}^{\lambda}_{k_1,j_1}[1:T] = \big((K + j_1)^{\times(\tau-1)}, (k_1)^{\times L}, (K + j_1)^{\times(T-\tau-L+1)}\big)$.

This pattern holds for all $k \in [K], j \in [J]$ because k_1, j_1 in the proof are arbitrary. In conclusion, the final image $\hat{I}^{\lambda}_{k_1,j_1} = I_{k_1}$, when $o(1) < \lambda < 1 + o(1)$. Thus, we conclude the proof of Lemma 17. □

E.11. Step 11: Incorrectly Generation with Conditional Model

Similarly, we now analyze the incorrectness of generation with conditional model. The intuition is that the conditional model fails to predict correctly at the transition position, triggering a cascade of errors in subsequent feature blocks and causing the generation to collapse into repetitive sequences.

Lemma 18 (Incorrectness of Generation with Conditional Model)**.** *Only with conditional model, i.e. with guidance scale $\lambda = 0$, the trained Transformer generates invalid sequence that repeats background $\hat{I}^{0}_{k,j}(\theta^{(2)}) = \big((K + j)^{\times T}\big) \neq I_{k,j}$, for all $k \in [K], j \in [J]$.*

Proof. Consider feature class k_1 and background class j_1, follow the inference procedure defined in Section 4, starting with prefix $[K + J + k_1, K + j_1]$.

Before feature: From Lemma 12, when $1 \le t < \tau - 1$,given input $\big(K + J + k_1, (K + j_1)^{\times t}\big)$, we know that the attention pattern is asymptotic uniform, i.e. $\alpha_{tl} = 1/(t + 1)$, for $l \le t$, so the output logits is the same with those in Lemma 9 for sufficiently large T. Therefore, we know that $\arg\max_i \text{logit}_t[i] = K + j_1$. So the background feature at $t = 2, \cdots, \tau - 1$ can be correctly generated, i.e. the generated image satisfies $\hat{I}^{0}_{k_1,j_1}[1 : \tau - 1] = \big((K + j_1)^{\times(\tau-1)}\big)$.

Transition: From Lemma 13, when $t = \tau - 1$, given input $\big(K + J + k_1, (K + j_1)^{\times(\tau-1)}\big)$, the attention concentrate to f_{K+J+k_1}, i.e. $\alpha_{t0} = 1 + o(1), \sum_{l=1}^{\tau-1} \alpha_{tl} = o(1)$, where the error is exponentially small. With direct expansion, we have $o_t = W_V^{(1)} f_{K+J+k_1} + o(1) W_V^{(1)} (f_{K+j_1} - f_{K+J+k_1})$. Use results in Lemma 7, and use similar technique in Lemma 12–Lemma 15 to compute the logits using the coefficient before each embeddings, we have

$$
\begin{aligned}
\text{logit}_t[k_1] &= -\frac{\eta_V p_{k_1} A_{K+J+k_1}}{K + J} + \eta_V p_{k_1} C_{K+J+k_1} + o(1), \\
\text{logit}_t[K + j] &= -\frac{\eta_V p_{k_1} A_{K+J+k_1}}{K + J} + p_{k_1} q_j B_{K+J+k_1} + o(1), j \in [J], \\
\text{logit}_t[k] &= -\frac{\eta_V p_{k_1} A_{K+J+k_1}}{K + J} + o(1), k \in [K],
\end{aligned}
$$

where the shared term comes from the coefficient of $\bar{f}$. So using Eqn. (16) and Proposition 6, we know that $C_{K+J+k_1} < q_j B_{K+J+k_1}$ with sufficiently large T, so with greedy decoding, we have $\arg\max_i \text{logit}_t[i] = K + j_1$. Therefore, the required feature token k_1 at $t = \tau$ **cannot** be correctly generated. So the generated images is already not valid, i.e. $\hat{I}^{0}_{k_1,j_1}(\theta^{(2)}) \neq I_{k_1,j_1}$.

Feature: Because at $t = \tau - 1$, the model predict $K + j_1$, the given input at $t = \tau$ still only contain text token and background token, i.e. $\big(K + J + k_1, (K + j_1)^{\times\tau}\big)$. Know we use induction to show that the model keep predict $K + j_1$ using greedy decoding in $t = \tau, \cdots, \tau + L - 2$.

Formally, we should prove: for $\tau \le t \le \tau + L - 2$, $\arg\max_i \text{logit}_t[i] = K + j_1$ and the next input is the input is $\big(K + J + k_1, (K + j_1)^{\times(t+1)}\big)$.

Base Case: At $t = \tau$, directly expand the softmax function with feature-position correlation, we can compute the attention

pattern

$$\begin{aligned}\alpha_{t0} &= \frac{\exp(e_t^\top W_{QK}^{(1)} f_{K+J+k_1})}{\exp(e_t^\top W_{QK}^{(1)} f_{K+J+k_1}) + \tau \exp(e_t^\top W_{QK}^{(1)} f_{K+j_1})} \\ &= \frac{1}{1 + \tau \exp\left(e_t^\top W_{QK}^{(1)}(f_{K+j_1} - f_{K+J+k_1})\right)} \\ &= 1 + o(1),\end{aligned}$$

where the second divide $\exp(e_t^\top W_{QK}^{(1)} f_{K+J+k_1})$ from the numerator and denominator, and the last equality using the assumption in Theorem 3 that $\tau = \Theta(T^\beta)$ and results from Lemma 14 that at $t = \tau$, $e_t^\top W_{QK}^{(1)}(f_{K+j_1} - f_{K+J+k_1}) = -\Theta(T^{1-\beta})$. So we know that $\sum_{l=1}^{\tau} \alpha_{tl} = o(1)$, where $o(1)$ is exponentially small. Now this case degenerate to the same case as the transition case above. Similarly, with direct expansion, we have $o_t = W_V^{(1)} f_{K+J+k_1} + o(1) W_V^{(1)}(f_{K+j_1} - f_{K+J+k_1})$. Use results in Lemma 7, and use similar technique in Lemma 12–Lemma 15 to compute the logits using the coefficient before each embeddings, we have

$$\begin{aligned}\text{logit}_t[k_1] &= -\frac{\eta_V p_{k_1} A_{K+J+k_1}}{K+J} + \eta_V p_{k_1} C_{K+J+k_1} + o(1), \\ \text{logit}_t[K+j] &= -\frac{\eta_V p_{k_1} A_{K+J+k_1}}{K+J} + p_{k_1} q_j B_{K+J+k_1} + o(1), j \in [J], \\ \text{logit}_t[k] &= -\frac{\eta_V p_{k_1} A_{K+J+k_1}}{K+J} + o(1), k \in [K],\end{aligned}$$

where the shared term comes from the coefficient of $\bar{f}$. So using Eqn. (16) and Proposition 6, we know that $C_{K+J+k_1} < q_j B_{K+J+k_1}$ with sufficiently large T, so with greedy decoding, we have $\arg\max_i \text{logit}_t[i] = K + j_1$. Therefore, the required feature token k_1 at $t = \tau + 1$ **cannot** be correctly generated. So the generated images is already not valid, i.e. $\hat{I}^0_{k_1,j_1}(\theta^{(2)}) \neq I_{k_1,j_1}$. And the input for next time step is $\big(K + J + k_1, (K + j_1)^{\times(\tau+1)}\big)$.

Induction: Suppose the induction assumptions holds for $\tau < s \leq \tau + L - 3$, we know that at $t = s + 1$, the input is $\big(K + J + k_1, (K + j_1)^{\times(s+1)}\big)$, so by directly expand the softmax function with position-feature correlation, we can compute attention score

$$\alpha_{t0} = \frac{\exp(e_t^\top W_{QK}^{(1)} f_{K+J+k_1})}{\exp(e_t^\top W_{QK}^{(1)} f_{K+J+k_1}) + (s+1)\exp(e_t^\top W_{QK}^{(1)} f_{K+j_1})} = 1 + o(1),$$

where we use the same technique as the base case, because they are almost the same except there are more terms related to background feature f_{K+j_1}, but $\tau < s \leq \tau + L - 3$ is still small compared to the exponential term $\exp\big(-\Theta(T^{1-\beta})\big)$ comes from $\exp\big(e_t^\top W_{QK}^{(1)}(f_{K+j_1} - f_{K+J+k_1})\big)$. So the attention pattern is the same as the base case, which concentrates on the text token at $t = 0$. With the same procedure, using Eqn. (16) and Proposition 6, when T is sufficiently large, we have $\arg\max_i \text{logit}_t[i] = K + j_1$. So the required feature at $t = s + 1$ **cannot** be correctly generated. And the input for next time step is $\big(K + J + k_1, (K + j_1)^{\times(s+2)}\big)$.

Therefore, we prove that in the generated image, the origional position for feature k_1 is now all $K + j_1$, i.e. $\hat{I}^0_{k_1,j_1}[1 : \tau + L - 1] = \big((K + j_1)^{\times(\tau+L-1)}\big)$.

After feature: Because from previous results, at $t = \tau + L - 1$, the input are all background token, i.e. $\hat{I}^0_{k_1,j_1}[1 : \tau + L - 1] = \big((K + j_1)^{\times(\tau+L-1)}\big)$. Know we use induction to show that the model keep predict $K + j_1$ using greedy decoding in $\tau + L - 1 \leq t \leq T - 1$.

Formally, we should prove: for $\tau + L - 1 \leq t \leq T - 1$, $\arg\max_i \text{logit}_t[i] = K + j_1$ and the next input is the input is $\big(K + J + k_1, (K + j_1)^{\times(t+1)}\big)$.

Base Case: At $t = \tau + L - 1$, directly expand the softmax function with feature-position correlation, we can compute the

attention pattern

$$\alpha_{t0} = \frac{\exp(e_t^\top W_{QK}^{(1)} f_{K+J+k_1})}{\exp(e_t^\top W_{QK}^{(1)} f_{K+J+k_1}) + (\tau + L - 1)\exp(e_t^\top W_{QK}^{(1)} f_{K+j_1})} = \frac{1}{\tau + L} + o(1),$$

where the second divide $\exp(e_t^\top W_{QK}^{(1)} f_{K+J+k_1})$ from the numerator and denominator, and the last equality using the assumption in Theorem 3 that $\tau = \Theta(T^\beta)$ and results from Proposition 11 that at $t \notin \mathcal{F}$, $e_t^\top W_{QK}^{(1)}(f_{K+j_1} - f_{K+J+k_1}) = o(1)$, which is exponentially small. With similar technique, we know that $\sum_{l=1}^t \alpha_{tl} = (\tau + L - 1)/(\tau + L) + o(1)$. Therefore, the attention is asymptotically uniform, and using results from Lemma 7, and similar technique above to compute the logits by the coefficient before each embeddings, we have

$$\begin{aligned}
\text{logit}_t[k] &= \tilde{A}(t) + \frac{\eta_V(\tau + L - 1)}{\tau + L} p_k q_{j_1} C_{K+j_1} + o(1), \quad k \in [K], k \neq k_1, \\
\text{logit}_t[k_1] &= \tilde{A}(t) + \frac{\eta_V}{\tau + L} p_{k_1} \Big((\tau + L - 1) q_{j_1} C_{K+j_1} + C_{K+J+k_1}\Big) + o(1), \\
\text{logit}_t[K + j] &= \tilde{A}(t) + \frac{\eta_V}{t+1} q_j p_{k_1} B_{K+J+k_1} + o(1), \quad j \in [J], j \neq j_1, \\
\text{logit}_t[K + j_1] &= \tilde{A}(t) + \frac{\eta_V}{t+1} q_{j_1} p_{k_1} B_{K+J+k_1} + \frac{\eta_V}{t+1}(\tau + L - 1) q_{j_1} B_{K+j_1} + o(1),
\end{aligned}$$

where the $\tilde{A}(t)$ is the shared factor in Lemma 8, because this case is asymptotic the same as the uniform case. So using the asymptotic order in Eqn. 16, with sufficiently large T, B_{K+j_1} dominates, we have greedy decoding $\arg\max_i \text{logit}_t[i] = K + j_1$. And the input for next time step is $\big(K + J + k_1, (K + j_1)^{\times(\tau+1)}\big)$.

Induction: Suppose the induction assumptions holds for $\tau + L - 1 < s \leq T - 2$, we know that at $t = s + 1$, the input is $\big(K + J + k_1, (K + j_1)^{\times(s+1)}\big)$, so attention pattern is

$$\alpha_{t0} = \frac{\exp(e_t^\top W_{QK}^{(1)} f_{K+J+k_1})}{\exp(e_t^\top W_{QK}^{(1)} f_{K+J+k_1}) + (s + 1)\exp(e_t^\top W_{QK}^{(1)} f_{K+j_1})} = \frac{1}{s + 2} + o(1),$$

where we use the same technique as the base case, because they are almost the same except there are more terms related to background feature f_{K+j_1}, and in this area, $t \notin \mathcal{F}$, so we still have $e_t^\top W_{QK}^{(1)} f_{K+j_1} = o(1), e_t^\top W_{QK}^{(1)} f_{k_1} = o(1)$. And we have $\sum_{l=1}^t \alpha_{tl} = (s + 1)/(s + 2) + o(1)$. So the attention pattern is the same as the base case, which is asymptotically uniform. With the same procedure, using Eqn. (16) and Proposition 6, when T is sufficiently large, we have $\arg\max_i \text{logit}_t[i] = K + j_1$. And the input for next time step is $\big(K + J + k_1, (K + j_1)^{\times(s+1)}\big)$.

Therefore, we finish the induction. And we know that, with greedy decoding, the predictions will repeat $K + j_1$ at every position. This pattern holds for all $k \in [K], j \in [J]$ because k_1, j_1 in the proof are arbitrary. In conclusion, $\hat{I}_{k,j}^0(\theta^{(2)}) = \big((K + j)^{\times T}\big) \neq I_{k,j}$. Thus, we conclude the proof of Lemma 18. □

E.12. Step 12: Alignment of Conditional Model and Unconditional Model

Now we prove that during sucessful generation, the conditional model and unconditional model have the same prediction with greedy decoding.

Lemma 19 (Alignment of Conditional Model and Unconditional Model)**.** *With guidance scale $o(1) < \lambda < 1 + o(1)$, where $o(1)$ denotes a positive vanishing term, during the generation, the conditional model and unconditional model has the same prediction, which means*

$$\arg\max_i \text{logit}_t[i] = \arg\max_i \text{logit}_t^{uc}[i].$$

Proof. Because we are now considering the unconditional model and conditional model seperately, without CFG operation. From Lemma 12–Lemma 15, we know that the attention pattern are concentrate on some tokens, with exponentially small error. So we firstly analyze the dominant embedding in $W_V^{(1)} f_{K+J+k_1}$, $W_V^{(1)} f_{K+j_1}$ or $W_V^{(1)} f_{k_1}$, and then discuss for different time regime.

Dominant embedding of $W_V^{(1)} f_{K+J+k_1}, W_V^{(1)} f_{K+j_1}, W_V^{(1)} f_{k_1}$: From Lemma 7, we know that after projection through $W_V^{(1)}$, the value vector is a linear combination of embeddings

$$W_V^{(1)} f_{K+j_1} = \eta_V\big[- q_{j_1} A_{K+j_1} \bar{f} + q_{j_1} B_{K+j_1} f_{K+j_1} + q_{j_1} C_{K+j_1} \sum_k p_k f_k\big],$$
$$W_V^{(1)} f_{k_1} = \eta_V\big[- p_{k_1} A_{k_1} \bar{f} + p_{k_1} \sum_j q_j B_{k_1} f_{K+j} + p_{k_1} C_{k_1} f_{k_1}\big],$$
$$W_V^{(1)} f_{K+J+k_1} = \eta_V\big[- p_{k_1} A_{K+J+k_1} \bar{f} + p_{k_1} \sum_j q_j B_{K+J+k_1} f_{K+j} + p_{k_1} C_{K+J+k_1} f_{k_1}\big].$$

Further applying results in Eqn. (16), we know that: For $W_V^{(1)} f_{K+j_1}$, because $q_{j_1} B_{K+j_1}$ dominates, the dominant embedding is f_{K+j_1}; for $W_V^{(1)} f_{k_1}$, from Proposition 6, we know that $q_j B_{k_1} < C_{k_1}$, so the dominant embedding is f_{k_1}; for $W_V^{(1)} f_{K+J+k_1}$, also from Proposition 6, we know that $q_j B_{B+J+k_1} < C_{K+J+k_1}$, so the dominant embedding is f_{K+J+k_1}. In the above analysis, we ignore the coefficient for $\bar{f}$, because they are the same for every embeddings.

We follow the same structure of proof of Lemma 17, and we already knows that the generation is correct. Consider feature class k_1 and background class j_1, follow the inference procedure stated in Section 4, starting with prefix $\big(K+J+k_1, K+j_1\big)$.

Before feature: From Lemma 12, when $1 \le t < \tau - 1$, we know that, we know that with suitable guidance scale, $o(1) < \lambda < 1 + o(1)$, given input $\big(K + J + k_1, (K + j_1)^{\times t}\big)$, the attention of conditional model is asymptotically uniform, i.e. $\alpha_{tl} = \big(1 + o(1)\big)/(t + 1)$ for $0 \le l \le t$, with exponentially small $o(1)$. Because the dominant embedding of $W_V^{(1)} f_{K+J+k_1}, W_V^{(1)} f_{K+j_1}$ are all f_{K+j_1}, the dominant embedding of o_t is also f_{K+j_1}. Because of weight typing, we can know that $\arg\max_i \text{logit}_t[i] = K + j_1$.

For unconditional model, all the prefix are f_{K+j_1}, so the output is $o_t^{\text{uc}} = W_V^{(1)} f_{K+j_1}$. Because the dominant embedding in $W_V^{(1)} f_{K+j_1}$ is f_{K+j_1} and weight tying, we know that $\arg\max_i \text{logit}_t^{\text{uc}}[i] = K + j_1$.

So when $1 \le t < \tau - 1$, we know that $\arg\max_i \text{logit}_t[i] = \arg\max_i \text{logit}_t^{\text{uc}}[i] = K + j_1$, the conditional model aligns with the unconditional model.

Transition: From Lemma 13, when $t = \tau - 1$, we know that with $o(1) < \lambda < 1 + o(1)$, given input $\big(K + J + k_1, (K + j_1)^{\times(\tau-1)}\big)$, the attention of conditional model concentrate on text token at $t = 0$, i.e. $\alpha_{t0} = 1 + o(1)$, with exponentially small error. Because the dominant embedding of $W_V^{(1)} f_{K+J+k_1}, W_V^{(1)} f_{K+j_1}$ are all f_{K+j_1}, the dominant embedding of o_t is also f_{K+j_1}. Because of weight typing, we can know that $\arg\max_i \text{logit}_t[i] = K + j_1$.

For unconditional model, all the prefix are f_{K+j_1}, so the output is $o_t^{\text{uc}} = W_V^{(1)} f_{K+j_1}$. Because the dominant embedding in $W_V^{(1)} f_{K+j_1}$ is f_{K+j_1} and weight tying, we know that the maximum is still $K + j_1$, i.e. $\arg\max_i \text{logit}_t^{\text{uc}}[i] = K + j_1$.

So when $t = \tau - 1$, we know that $\arg\max_i \text{logit}_t[i] = \arg\max_i \text{logit}_t^{\text{uc}}[i] = K + j_1$, the conditional model aligns with the unconditional model.

Feature: From Lemma 14, we know that with suitable guidance scale $o(1) < \lambda < 1 + o(1)$, given input $\big(K + J + k_1, (K + j_1)^{\times(\tau-1)}, (k_1)^{\times(t-\tau+1)}\big)$, we need to discuss two cases: $t = \tau$ and $\tau < t \le \tau + L - 2$.

When $t = \tau$, from the proof of Lemma 14, the conditional model half the attention score on text token at $t = 0$ and the seen feature block, i.e. $\alpha_{t0} = \big(1 + o(1)\big)/2$, $\alpha_{t\tau} = \big(1 + o(1)\big)/2$, with exponential small error. For sufficiently large T, we only need to find the dominant embedding for $(W_V^{(1)} f_{K+J+k_1} + W_V^{(1)} f_{k_1})/2$. Use Lemma 7 to expand, we know that we need to compare the coefficient for f_{K+j}, f_{k_1}, which are $q_j(B_{K+J+k_1} + B_{k_1})/2$ and $(C_{K+J+k_1} + C_{k_1})/2$. From the results in Proposition 6, we have

$$\frac{C_{K+J+k_1} + C_{k_1}}{B_{K+J+k_1} + B_{k_1}} = \frac{L\big(1 + o(1)\big)/\tau + L(L-1)/(2\tau) + O(L^3/\tau^2)}{\log T + O(1) + L(1-\beta)\log T + O(1)} = \frac{L-1}{2(1-\beta)\tau \log T} + o(1),$$

where the first equation is direct expansion using exact expression in Proposition 5, the second asymptotic equivalence and the last inequality use the assumption in Theorem 3. So use the Eqn. (17) and proof in Proposition 5, we know that for sufficiently large T, $(C_{K+J+k_1} + C_{k_1})/(B_{K+J+k_1} + B_{k_1}) < q_j$. With direct computation, we know that

$q_j(B_{K+J+k_1} + B_{k_1})/2 < (C_{K+J+k_1} + C_{k_1})/2$, so the dominant embedding in o_t is f_{k_1}. Because of weight tying, we know that $\arg\max_i \text{logit}_t^{\text{uc}}[i] = k_1$.

For unconditional model, from Lemma 14, the attention concentrate on the feature block, so the output is $o_t^{\text{uc}} = W_V^{(1)} f_{K+j_1} + o(1) W_V^{(1)}(f_{k_1} - f_{K+j_1})$, with exponentially small $o(1)$. For sufficiently large T, because the dominant embedding in $W_V^{(1)} f_{k_1}$ and weight tying, we know that the maximum logit is at k_1, i.e. $\arg\max_i \text{logit}_t^{\text{uc}}[i] = k_1$, which is the same as unconditional model.

When $\tau < t < \tau + L - 2$, from the proof of Lemma 14, the conditional model and unconditional model concentrate the attention score on the feature block, i.e. $\sum_{l=\tau}^{t} \alpha_{tl} = 1 + o(1)$, $\sum_{l=\tau}^{t} \alpha_{tl}^{\text{uc}} = 1 + o(1)$, with exponentially small error term $o(1)$. For sufficiently large T, because the dominant embedding in $W_V^{(1)} f_{k_1}$ and weight tying, we know that $\arg\max_i \text{logit}_t[i] = \arg\max_i \text{logit}_t^{\text{uc}}[i] = k_1$, which is the same as unconditional model.

In conclusion, for feature block $\tau \le t \le \tau + L - 2$, the conditional model and unconditional model aligns in the prediction, i.e. $\arg\max_i \text{logit}_t[i] = \arg\max_i \text{logit}_t^{\text{uc}}[i]$.

After feature: From Lemma 15, when $\tau \le t \le \tau + L - 2$, we know that with $o(1) < \lambda < 1 + o(1)$, given input $\big(K + J + k_1, (K + j_1)^{\times(\tau-1)}, (k_1)^{\times(t-\tau+1)}, (K + j_1)^{\times(t-\tau-L+1)}\big)$, the attention of conditional model and unconditional model are asymptotically uniform, i.e. $\alpha_{tl} = \big(1 + o(1)\big)/(t + 1)$ for $0 \le l \le t$, with exponentially small $o(1)$. From assumption in Theorem 3, we know that $\tau = \Theta(T^\beta), L = \Theta(\log T)$, so the dominant part is $W_V^{(1)} f_{K+j_1}$, for sufficiently large T, because the number of token $K + j_1$ is the most. Because the dominant embedding of $W_V f_{K+j_1}$ are all f_{K+j_1}, the dominant embedding of o_t is also f_{K+j_1}. Because of weight typing, we can know that $\arg\max_i \text{logit}_t[i] = \arg\max_i \text{logit}_t^{\text{uc}}[i] = K + j_1$.

In conclusion, we know that the conditional model and unconditional model has the same prediction, i.e. $\arg\max_i \text{logit}_t[i] = \arg\max_i \text{logit}_t^{\text{uc}}[i]$. Thus, we conclude the proof of Lemma 19. □

E.13. Step 13: Proof of Theorem 3

Proof of Theorem 3. The three parts of Theorem 3 can be proved with previous results.

Part I (Loss): From Proposition 16, the population loss of the trained model is $\mathcal{L}_{\text{vis}}(\theta^{(2)}) = \Theta\big((\log T)^2/T\big)$.

Part II (Correctness of Generation with CFG): From Lemma 17, we know that with suitable guidance scale $\lambda^* \in \big(o(1), 1 + o(1)\big)$, where $o(1)$ is positive vanishing term, the generated images is valid, i.e. $\hat{I}_{k,j}^{\lambda^*} = I_{k,j}$ for $k \in [K], j \in [J]$.

Part III (Incorrectness of Generation with CFG): From Lemma 18, we know that only with conditional models, i.e. $\lambda = 0$, it generates invalid images that repeats backgrounds, i.e. $\hat{I}_{k,j}^{0}(\theta^{(2)}) = \big((K + j)^{\times T}\big) \ne I_{k,j}$, for all $k \in [K], j \in [J]$.

Part IV (Alignment of Conditional and Unconditional Model): From Lemma 19, we know that with $\lambda^* \in \big(1 + o(1)\big)$, the conditional model and unconditional model has the same prediction (with greedy decoding), which means, $\arg\max_i \text{logit}_t[i] = \arg\max_i \text{logit}_t^{\text{uc}}[i]$.

Thus, we conclude the proof of Theorem 3. □

Algorithm 1 Attention Weight Reuse (AttnReuse) for AR models

Require: Text-token length T_{txt}; current sequence length t; reuse-layer set $\mathcal{L}_{\text{reuse}}$; CFG scale λ.
Ensure: Guided logits logit_t^{λ}.

for each decoding step t **do**
  Run the conditional branch and cache the conditional attention weights $\{A_\ell^{\text{cond}}\}_{\ell \in \mathcal{L}_{\text{reuse}}}$.
  for each layer ℓ **do**
    if $\ell \in \mathcal{L}_{\text{reuse}}$ **then**
      Construct a binary key mask $M \in \{0,1\}^t$:

$$M_j = \begin{cases} 0, & j \le T_{\text{txt}}, \\ 1, & T_{\text{txt}} < j \le t. \end{cases}$$

      Remove the text-token contribution from the conditional attention weights:

$$\widetilde{A}_\ell^{\text{reuse}} = A_\ell^{\text{cond}} \odot M.$$

      Renormalize over the key dimension:

$$A_\ell^{\text{reuse}} = \frac{\widetilde{A}_\ell^{\text{reuse}}}{\sum_{j=1}^{t} \widetilde{A}_{\ell,:,j}^{\text{reuse}} + \epsilon}.$$

      Approximate the unconditional attention output:

$$O_\ell^{\text{reuse}} = A_\ell^{\text{reuse}} V_\ell^{\text{cond}}.$$

      Replace the unconditional attention output at layer ℓ with O_ℓ^{reuse}.
    else
      Compute the unconditional attention output with standard attention:

$$O_\ell^{\text{uncond}} = \text{Attn}\left(Q_\ell^{\text{uncond}}, K_\ell^{\text{uncond}}, V_\ell^{\text{uncond}}\right).$$

    end if
  end for
  Obtain $\text{logit}_t^{\text{cond}}$ from the conditional branch and approximate $\widehat{\text{logit}}_t^{\text{uncond}}$ from the resulting unconditional branch.
  Compute the guided logits:

$$\text{logit}_t^{\lambda} = \widehat{\text{logit}}_t^{\text{uncond}} + \lambda \left(\text{logit}_t^{\text{cond}} - \widehat{\text{logit}}_t^{\text{uncond}} \right).$$

  Decode the next visual token from logit_t^{λ}.
end for

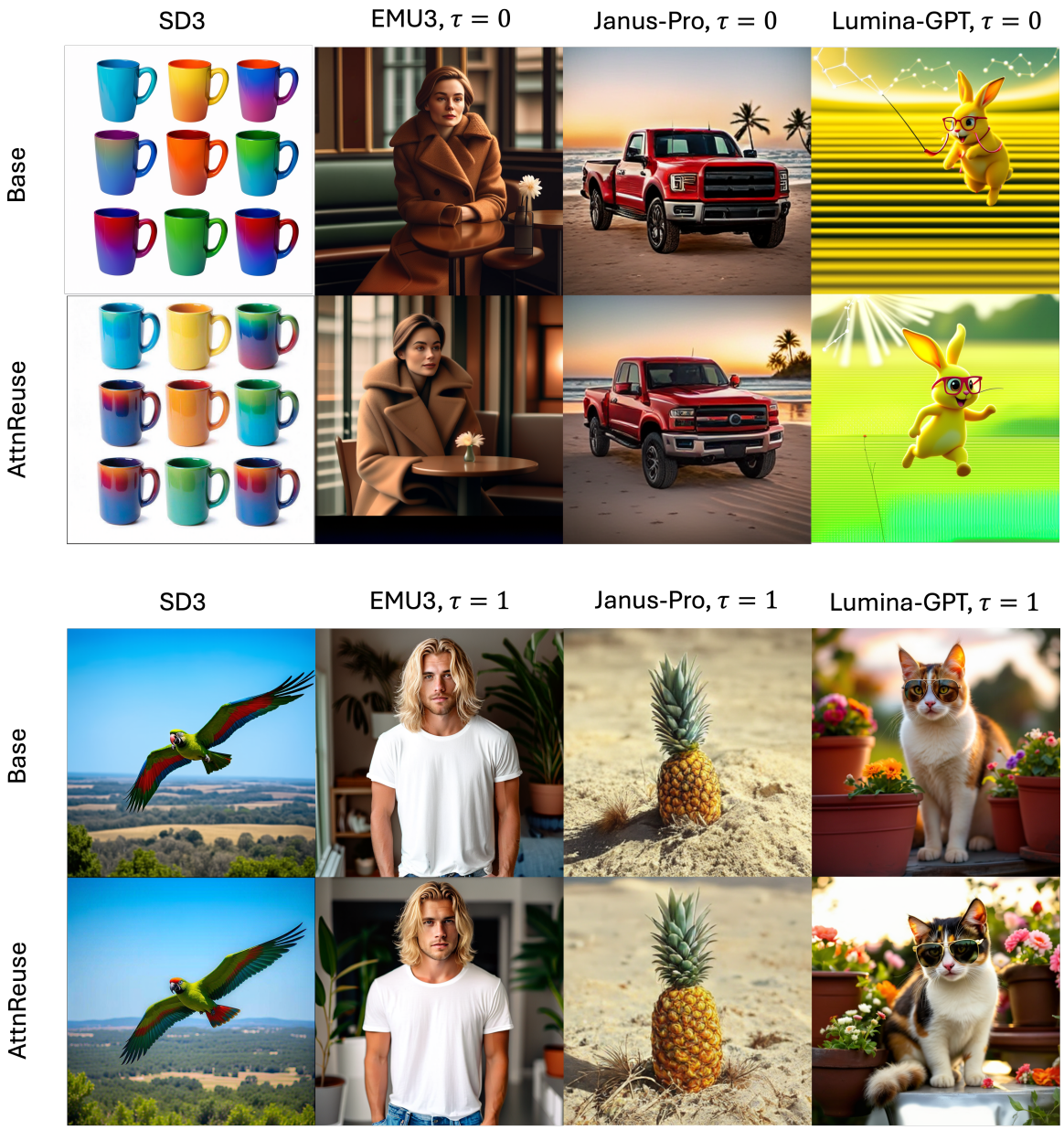

Figure 12. Qualitative comparison between **Base** and **AttnReuse** across different models. AttnReuse preserves structural and semantic fidelity under both greedy ($\tau = 0$) and stochastic ($\tau = 1$) decoding, while reducing attention computation.