

Seg4Diff: Unveiling Open-Vocabulary Segmentation in Text-to-Image Diffusion Transformers

Chaehyun Kim¹ Heeseong Shin¹ Eunbeen Hong¹ Heeji Yoon¹
 Anurag Arnab Paul Hongsuck Seo² Sunghwan Hong^{3,†} Seungryong Kim^{1,†}
¹KAIST AI ²Korea University ³ETH Zürich AI Center, CVG, PRS
<https://cvlab-kaist.github.io/Seg4Diff>

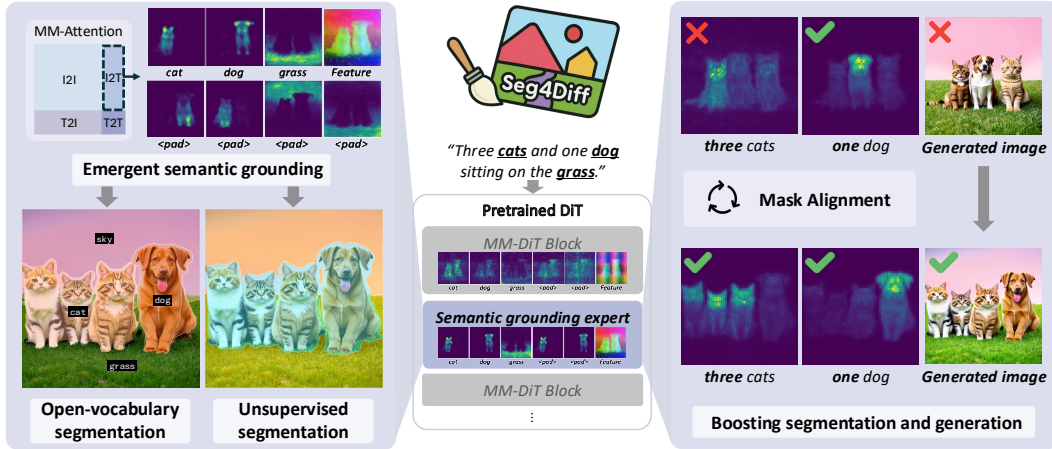


Figure 1: We introduce **Seg4Diff**, a systematic framework designed to *analyze and enhance* the emergent semantic grounding capabilities of multi-modal diffusion transformer (MM-DiT) blocks in text-to-image diffusion transformers (DiTs). Here, *semantic grounding expert* refers to a specific MM-DiT block responsible for establishing semantic alignment between text and image features.

Abstract

Text-to-image diffusion models excel at translating language prompts into photorealistic images by implicitly grounding textual concepts through their cross-modal attention mechanisms. Recent multi-modal diffusion transformers extend this by introducing joint self-attention over concatenated image and text tokens, enabling richer and more scalable cross-modal alignment. However, a detailed understanding of how and where these attention maps contribute to image generation remains limited. In this paper, we introduce **Seg4Diff (Segmentation for Diffusion)**, a systematic framework for analyzing the attention structures of MM-DiT, with a focus on how specific layers propagate semantic information from text to image. Through comprehensive analysis, we identify a *semantic grounding expert* layer, a specific MM-DiT block that consistently aligns text tokens with spatially coherent image regions, naturally producing high-quality semantic segmentation masks. We further demonstrate that applying a lightweight fine-tuning scheme with mask-annotated image data enhances the semantic grouping capabilities of these layers and thereby improves both segmentation performance and generated image fidelity. Our findings demonstrate that semantic grouping is an emergent property of diffusion transformers and can be selectively amplified to advance both segmentation and generation performance, paving the way for unified models that bridge visual perception and generation.

1 Introduction

The emergence of text-to-image (T2I) diffusion models have revolutionized visual content creation by allowing users to generate photorealistic images from natural-language prompts [46, 22, 52]. The success of these models imply that text-to-image diffusion models are highly capable of associating visual and textual representations, as the generated images accurately reflect the given textual prompt.

In this regard, a growing body of work shows that this semantic grounding is encoded in the cross-modal attention maps of the diffusion models [20, 41, 49, 43], which have already been actively exploited for generative downstream tasks such as image editing, inpainting and personalization [48, 20, 5, 39]. Furthermore, several studies [60, 54] show that the cross-attention maps can be used to perform visual perception tasks such as open-vocabulary semantic segmentation, revealing that diffusion models can double as training-free segmentation models. However, the attention map produced by the standard U-Net-based [47, 46] models is often noisy and spatially fragmented [6], limiting the fidelity of the resulting masks when leveraged without heuristic refinement.

Recent advances replace the U-Net architecture with diffusion transformers (DiTs) [42], which benefits from the strong representation power and scalability of transformers for achieving state-of-the-art image generation quality. In particular, the multi-modal diffusion transformer (MM-DiT) [15, 53, 3] introduces multi-modal attention, which concatenates text and image tokens and applies joint self-attention, enabling richer cross-modal interaction. This encourages us to explore the internal representations and the characteristics of the MM-DiT-based models. However, as opposed to U-Net based models, which have been extensively studied [38], DiT-based models and their characteristics remain underexplored, in which a detailed understanding of how and where these attention maps contribute to image generation remains limited.

In this paper, we first conduct an in-depth analysis of the joint attention mechanism in MM-DiT models. We characterize the distribution of attention scores to discover active cross-modal interaction, and complement this with feature similarity measure and norm analysis to assess which modality—textual or visual—exerts greater influence on the output representations. Together, these perspectives isolate a small subset of layers that consistently align textual semantics with contiguous image regions. These observations motivate us to further analyze this emergent semantic grounding capability.

Subsequently, we propose an open-vocabulary segmentation scheme leveraging MM-DiT, which demonstrates that the identified layer, *semantic grounding expert*, yields high-quality segmentation masks, confirming their inherently competitive capability for open-vocabulary semantic segmentation. Extended analysis reveals that these attention map is in fact further decomposed into attention heads, which capture distinct parts of the semantic region and finally sum up to construct a complete semantic mask. Moreover, without explicit class information, we reveal that <pad> tokens can decompose the image into meaningful semantic regions, acting as anchors for the unconditional generation.

Building on these findings, we introduce a lightweight fine-tuning scheme, called *mask alignment for segmentation and generation (MAGNET)*, that explicitly strengthens semantic grouping by leveraging mask-annotated image data in selected MM-DiT layers. Our primary motivation is to reinforce the emergent segmentation capability of these layers, enabling more accurate open-vocabulary semantic masks. Interestingly, this targeted adaptation also yields a modest but consistent improvement in image synthesis quality as a by-product of enhancing the cross-modal semantic alignment. Taken together, our stepwise analysis and selective refinement establish semantic grouping as a salient property of diffusion transformers for text-conditioned image generation—one that can be leveraged with minimal compute to advance both dense recognition and generative fidelity. These results illuminate the internal dynamics of diffusion transformers and point toward unified models that excel at both generation and perception.

In summary, our contributions are as follows:

- We provide in-depth analysis and investigation of learned representations of MM-DiT within its multi-modal attention layers.
- We identify critical layers in MM-DiT that are essential for preserving text-conditioned semantics throughout the generation process, which reveals strong semantic grounding.
- We demonstrate a zero-shot segmentation scheme to extract segmentation masks from the specific layers, and further enhance generation quality by enhancing their localization capability.

2 Related Work

Emergent properties of diffusion models. Diffusion models, which generate images by iteratively denoising Gaussian noise, have transformed generative modeling [22, 52]. U-Net-based text-to-image models [46] surpassed GANs [18], while recent transformer backbones [42] further raised the bar. In particular, multi-modal diffusion transformer (MM-DiT) [15], achieve stronger scalability and semantic alignment by fusing image and text tokens through joint attention. Beyond generation, pretrained diffusion models have been adapted to perception tasks such as correspondence [67, 56], segmentation [59], tracking [37], depth estimation [29, 26] and video understanding [58]. On the other hand, inspired by Prompt-to-Prompt [20], which first proposed to leverage the cross-modal attention maps of text-to-image diffusion models revealing that text information is delicately aggregated into image generation via the attention mechanism, a rich line of works explored attention maps in different areas, such as image editing [7] and conditional generation [1]. Building on this, we focus specifically on analyzing the joint attention mechanism of MM-DiT, uncovering its rich semantic structure and its potential for segmentation. This perspective not only bridges generation and perception but also highlights attention as a central building block for semantic reasoning in diffusion models.

Leveraging diffusion models for perception tasks. Recently, text-to-image diffusion models [46, 43, 15] have been adapted beyond generation to tackle discriminative vision tasks with remarkable success. For segmentation, ODISE [64] integrates a frozen diffusion backbone with CLIP, while adapter-based methods [68] leverage cross-attention for referring segmentation. For geometry, lightweight fine-tuning of diffusion model for depth or normal estimation [63, 17] enhances spatial detail and generalization. In correspondence, methods like DiffMatch [40] achieve robustness to textureless regions, while leveraging frozen features [56] still rival supervised baselines [9, 10, 24]. Collectively, these works show that diffusion models encode rich priors transferable to perception. However, most approaches treat diffusion feature as static representations, leaving the role of internal attention mechanism underexplored, particularly in architectures like MM-DiT.

Diffusion models for segmentation. Diffusion models [4, 15] excel at photorealistic image synthesis by aligning semantic cues between images and text, which has spurred efforts to repurpose their learned representations for segmentation. For example, OVDiff [28] synthesizes class-specific support images and extracts their features to form prototypes that guide open-vocabulary segmentation, while DiffSegmenter [60] mines self- and cross-attention maps from pretrained text-to-image diffusion models to localize arbitrary objects. Building on these insights, iSeg [54] applies an entropy-reduction step to self-attention maps and iteratively fuses them with cross-attention to progressively sharpen segmentation, and Diffseg [57] demonstrates that self-attention features alone can be converted directly into masks. Meanwhile, DiffCut [13] frames unsupervised segmentation as a graph-partitioning problem over diffusion features. Our work not only probes the latent representations learned by diffusion models but also leverages them to jointly enhance both image generation and segmentation.

3 Seg4Diff: Segmentation for Diffusion

3.1 Motivation and Overview

The remarkable success of diffusion models in text-guided image generation [46, 43] is primarily attributed to their attention mechanisms [46, 20], with recent architectures like multi-modal diffusion transformer (MM-DiT) [15] further enhancing performance through joint attention over image and text modalities. Despite these advancements, the precise interactions within these multi-modal attention mechanisms remain largely underexplored. In this regard, our work aims to deeply investigate these internal interactions to elucidate the principles underlying their success, and based on these insights, we propose a simple yet effective training scheme to transfer the learned representations to the semantic segmentation task.

We begin by briefly reviewing rectified flow framework [36] and multi-modal attention mechanism in MM-DiT in Sec. 3.2. Then, in Sec. 3.3, we delve into how text and image tokens interact in attention mechanism of MM-DiT. Building on this, we identify key layers critical for image-text alignment and thereby results in faithful image generation, where we introduce zero-shot segmentation framework to leverage this semantic grounding capability. Subsequently, we propose learning strategy that jointly enhances both segmentation and generative capabilities in Sec. 3.5.

3.2 Preliminaries

Rectified flow framework. While diffusion models reverse a predefined process using simulated data [22, 52], conditional flow matching (CFM) [35] trains continuous normalizing flows without simulation by framing generative modeling as regression between analytic conditional vector fields and a learnable velocity field. Rectified Flow [36] further simplifies this framework by adopting a deterministic linear interpolation between data and noise distributions. Specifically, given a data sample $x_0 \sim p_{\text{data}}$ and a noise sample $\epsilon \sim \mathcal{N}(0, I)$, the interpolation at time $t \in [0, 1]$ is defined as:

$$x_t = (1 - t)x_0 + t\epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \quad (1)$$

The corresponding ground-truth velocity field along this path is constant and given by

$$u_t(x_t) = \epsilon - x_0.$$

Let $v_t(x_t)$ denote the predicted velocity field parameterized by the model. The training objective minimizes the discrepancy between $v_t(x_t)$ and $u_t(x_t)$, leading to the flow matching (FM) loss

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t \sim \mathcal{U}(0,1), x_0 \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(0, I)} \|u_t(x_t) - v_t(x_t)\|^2. \quad (2)$$

Multi-modal attention. Building upon this formulation, transformer-based architecture known as multi-modal diffusion transformer (MM-DiT), which is exemplified by Stable Diffusion 3 (SD3) [15], results in significant improvements in generation capability. At its core, MM-DiT employs a multi-modal attention (MM-Attn) mechanism that simultaneously processes image and text modalities.

Given input image I and text T , let $x_{\text{img}} \in \mathbb{R}^{hw \times d}$, and $x_{\text{text}} \in \mathbb{R}^{l \times d}$ be the image and text embeddings, where h and w is the height and width of image latent, and l is the text token length. For each modality $m \in \{I, T\}$, query, key, and value embeddings Q_m , K_m , and V_m are channel-wise split to Q_m^h , K_m^h , and V_m^h for each attention head $h \in \{1, 2, \dots, H\}$, where H is the number of heads. These embeddings are concatenated into Q^h , K^h and $V^h \in \mathbb{R}^{(hw+l) \times (d_k)}$ respectively:

$$Q^h = [Q_I^h; Q_T^h], \quad K^h = [K_I^h; K_T^h], \quad V^h = [V_I^h; V_T^h], \quad (3)$$

where d_k is the attention embedding dimension and $[\cdot; \cdot]$ denotes token-wise concatenation. Subsequently, these results are leveraged to produce multi-head attention score A^h , which is then multiplied with value tokens to produce attention output O^h . The final MM-Attn($x_{\text{img}}, x_{\text{text}}$) after output projection $\mathcal{P}_O$ is then split back into image and text embeddings for downstream processing:

$$A^h = \text{softmax}(Q^h (K^h)^\top / \sqrt{d_k}), \quad (4)$$

$$O^h = A^h V^h, \quad (5)$$

$$\text{MM-Attn}(x_{\text{img}}, x_{\text{text}}) = \mathcal{P}_O([A^1 V^1, \dots, A^H V^H]), \quad (6)$$

where $[\cdot, \cdot]$ denotes channel-wise concatenation. The MM-Attn module is designed to handle four distinct types of query-to-key interaction: image-to-image (I2I), image-to-text (I2T), text-to-image (T2I), and text-to-text (T2T), which we denote A_{I2I} , A_{I2T} , A_{T2I} and A_{T2T} respectively. These four pathways are illustrated in Fig. 2(a), where we particularly focus on A_{I2T} for analysis.

3.3 Emergent Semantic Alignment

In this section, we investigate the interactions between x_{img} and x_{text} in DiT. To do so, we decompose and analyze attention scores and the resulting features across all heads and layers. For all the analysis, we sample 50 text prompts consisting of diverse linguistic and stylistic nature from DrawBench [49] and then generate corresponding images using the pretrained model. For clarity, we report results at timestep $t = 8$ and $t = 28$; additional analyses over other timesteps are provided in Appx. B.1. In addition, we focus on Stable Diffusion 3 architecture in our main analysis, where analysis on other MM-DiT variants such as Stable Diffusion 3.5 [53] and Flux [3] can be found in Appx. C.

Decomposing joint image-text attention. We start by visualizing and quantifying the attention scores to assess interaction strengths between image and text tokens in MM-Attn. As shown in Fig. 2 (b) and (c), we compute layer-wise attention scores and aggregate them by interaction type, equivalent to Eq. 4. Fig. 2 (b) reveals that I2T scores are disproportionately higher than I2I, even though the I2T region in (a) is roughly 40x smaller than I2I, indicating that I2T dominates the overall attention budget. Conversely, Fig. 2 (c) shows that T2I interactions are relatively smaller than T2T, suggesting that text tokens primarily self-attend to preserve semantic anchors. Based on these findings, we focus our subsequent analysis on I2T interactions, where further discussion can be found in Appx. B.2.

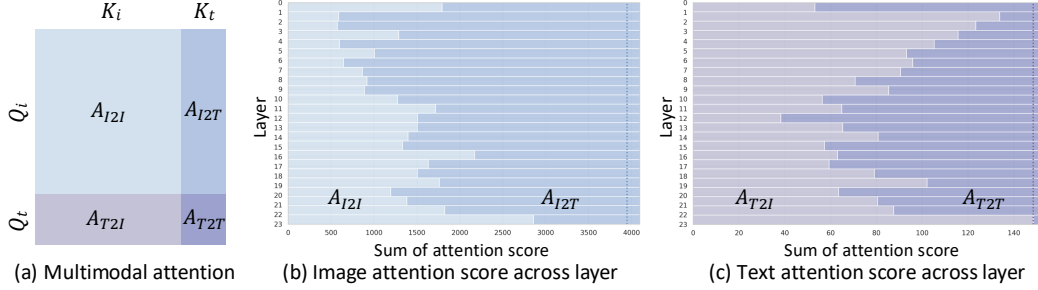


Figure 2: **Multi-modal attention mechanism.** (a) Conceptual visualization of the attention map. (b–c) Ratios of attention assigned to image vs. text tokens. The dotted line denotes the ratio under uniform attention. Higher cross-modal proportions are observed in A_{I2T} and A_{T2I} .

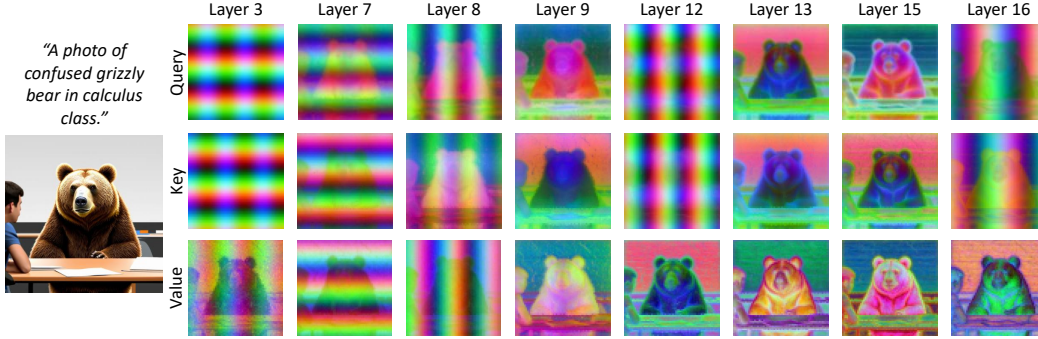


Figure 3: **PCA visualization of query, key and value projections.** PCA results demonstrate that some layers exhibit strong positional bias, whereas some layers show clear semantic groups.

Attention feature analysis. To understand how MM-DiT’s strong image-text interaction operates, we analyze feature space of MM-Attn. Feature PCA visualized in Fig. 3 highlights several layers indicate that query and key features are semantically well aligned, whereas other layers display stronger positional biases. We further analyze the average norm of the value-projected features at each layer, which is computed as $\frac{1}{N} \sum_{i=1}^N \|V^i\|_2$ for image, text and actual prompt tokens, which excludes following padding tokens. The attention norm serves as a proxy for the dominance of aggregated information, i.e., how strongly each key token influences a query token [32]. As shown in Fig. 4, similar layers—particularly the 9th MM-DiT block—exhibit notably high attention norms for text tokens compared to that of image tokens. This pattern suggests that text information is injected primarily into semantically aligned image-token regions at these specific layers.

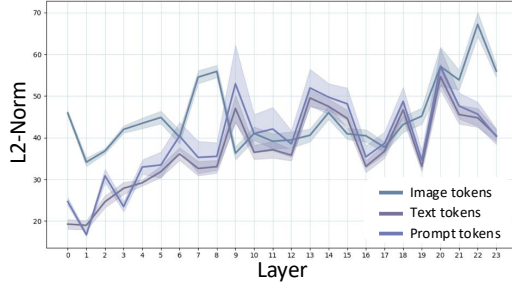


Figure 4: **Attention feature analysis.** The L2 norm of the value projection for image and text tokens reveals that certain layers exhibit significantly stronger value magnitudes for text tokens compared to image tokens.

Attention perturbation for image-text alignment. Building on this observation, we test whether these layers *causally* drive image synthesis and image-text alignment. Following SEG [25], we apply Gaussian blur to the image-to-text (I2T) attention maps of selected layers and regenerate images, which is shown in Fig. 5. If these layers mediate alignment, attenuating their attention should reduce semantic fidelity. Indeed, perturbing specific layers—which matches with semantically well-aligned layers in our previous analysis—induces clear mismatches between text and content, confirming their critical role in aligning image and text.

We then convert this perturbation into a lightweight guidance mechanism to both validate and exploit the effect. Rather than directly perturbing the sample, we *guide* the standard denoising trajectory

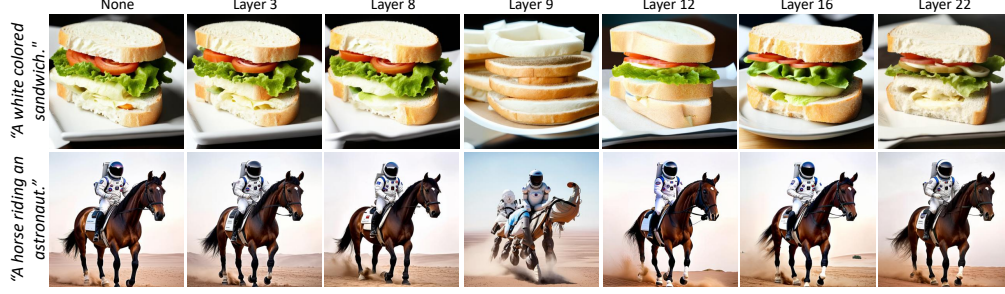


Figure 5: **Effect of layer-wise I2T attention perturbation.** Perturbing specific layer causes pronounced structural degradation, whereas other layers yields minor changes.

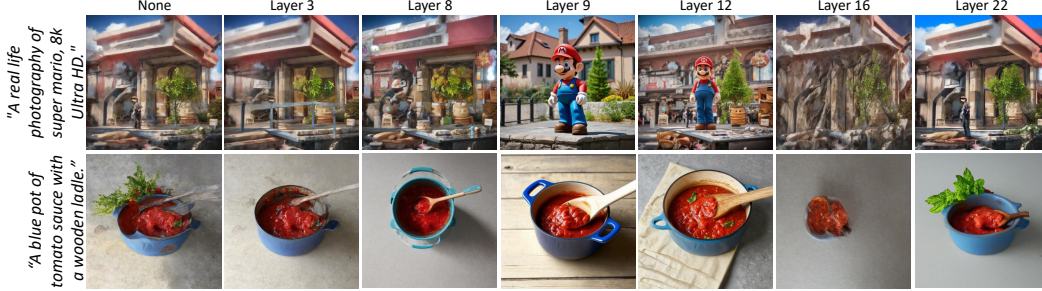


Figure 6: **Effect of perturbed I2T guidance.** Guiding specific layer with perturbed I2T attention score significantly enhances image quality.

with a negatively perturbed companion sample [1, 25]. Specifically,

$$\tilde{\epsilon}_{\theta}(x_t) = \epsilon_{\theta}(x_t; A_{I2T}) + s(\epsilon_{\theta}(x_t; A_{I2T}) - \hat{\epsilon}_{\theta}(x_t; \tilde{A}_{I2T})), \quad (7)$$

where $\tilde{\epsilon}_{\theta}$ is the guided noise estimate, ϵ_{θ} the standard estimate, $\hat{\epsilon}_{\theta}$ is the perturbed noise estimate with $\tilde{A}_{I2T}$, which is the perturbed version of A_{I2T} , and $s > 0$ a guidance scale. As shown in Fig. 6, this procedure substantially improves image quality while modifying only the specific layer’s I2T attention regions.

Together, these findings indicate that (i) semantically well-aligned layers are essential for injecting text conditioning into the image, and (ii) overall synthesis quality hinges disproportionately on these layers—consistent with [2], which likewise identifies particular MM-DiT layers as crucial for effective image editing.

3.4 Emergent Semantic Grouping

From the previous analysis, we identified MM-DiT layers that exhibit stronger, more critical multi-modal interactions than others and encode rich semantic structure. We now examine whether this semantic grounding manifests from a segmentation perspective.

Specifically, we focus on the semantic grounding capability within the I2T attention maps. Obtaining segmentation prediction begins by encoding the input image I into a latent representation x_{img} using a VAE encoder $\mathcal{E}_{\text{VAE}}$. To ensure reliable segmentation fidelity, we use an intermediate noisy latent, generated by linearly interpolating the clean latent $\mathcal{E}_{\text{VAE}}(I)$ with random noise $\epsilon \sim \mathcal{N}(0, 1)$ at a denoising timestep t , following the rectified flow formulation [15, 36]. This technique preserves spatial structure while retaining semantic content, where the ablation on timestep choice can be found in Appx. B.1. Meanwhile, a text prompt T is formed by concatenating the ground-truth classnames existing in the input image (e.g., "car mountain sky water") and encoded into text embeddings x_{text} via a text encoder $\mathcal{E}_T$:

$$x_{\text{img}} = t \cdot \mathcal{E}_{\text{VAE}}(I) + (1 - t) \cdot \epsilon, \quad (8)$$

$$x_{\text{text}} = \mathcal{E}_T(T). \quad (9)$$

Both embeddings, x_{img} and x_{text} , are fed into the MM-DiT blocks. In each layer, let A_{I2T}^h be the attention map from image tokens to a specific text token for head h . We compute the mean I2T

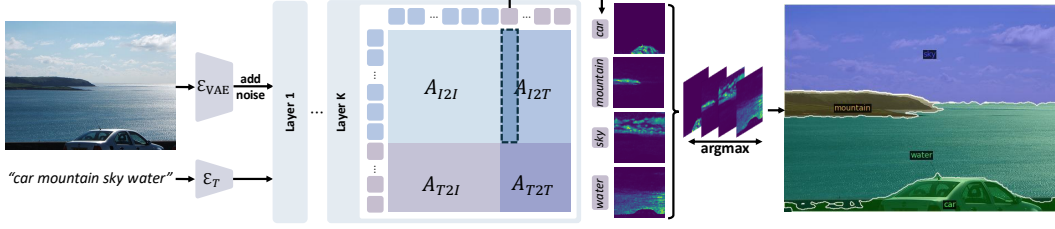


Figure 7: **Open-vocabulary semantic segmentation scheme in our framework.** We generate segmentation masks by interpreting the I2T attention scores, where the score map for each text token serves as a direct measure of image-text similarity to produce the final prediction.

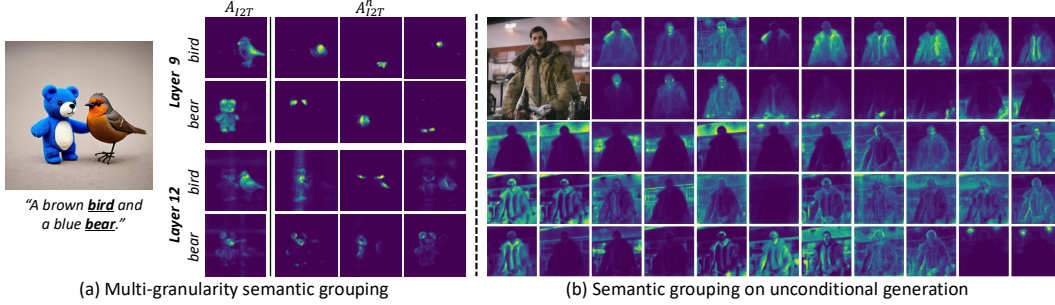


Figure 9: **Deeper analysis on multi-modal attention mechanism.** (a) multi-granularity behavior of token-level and head-level attention, and (b) emergent semantic grouping on <pad> tokens in unconditional generation scenario.

attention map, $\bar{A}_{I2T} \in \mathbb{R}^{hw \times l \times Hd_k}$, by averaging I2T attention $A_{I2T}^h \in \mathbb{R}^{hw \times l \times d_k}$ across all heads, which is then reshaped to construct a mask logit $M^{(j)} \in \mathbb{R}^{h \times w \times l}$ corresponding to each text token index j . Formally, the overall process is summarized by the following equations:

$$\bar{A}_{I2T} = \frac{1}{H} \sum_{k=1}^H A_{I2T}^k, \quad (10)$$

$$M^{(j)} = \text{Reshape}(\bar{A}_{I2T}^{(j)}) \in \mathbb{R}^{h \times w}, \quad \text{for } j = 1, \dots, l. \quad (11)$$

Since a single classname may correspond to multiple text tokens, we average their respective attention maps to produce a single logit map per class. This results in a final logit tensor $P \in \mathbb{R}^{h \times w \times C}$, where C is the number of entire classes. The final prediction is obtained by applying an argmax operation at each pixel location to assign the most likely class. [71, 11, 51]

We visualize the qualitative results in Fig. 7 and quantitative results in Fig. 8. Detailed explanation regarding evaluation metrics, pACC, mACC and mIoU, can be found in Appx. A.4. The result shows that specific layers are responsible for semantic grounding, where 9th layer performs the best, which aligns with previously identified in above. We refer to layers as *semantic grounding expert* layers.

Multi-granularity semantic grouping in multi-head attention. We further decompose attention map of semantic grounding expert layer into individual heads to see how each text token captures complete semantic region. Fig. 9 (a) visualizes head- and token-level attention map of for the highlighted text token, A_{I2T}^h and A_{I2T} respectively. The attention heads focus on a distinct semantic part of the object, such as ears and legs of a bear. Subsequently, token-level attention map precisely delineates the target semantics, demonstrating an inherent multi-granular grouping capability.

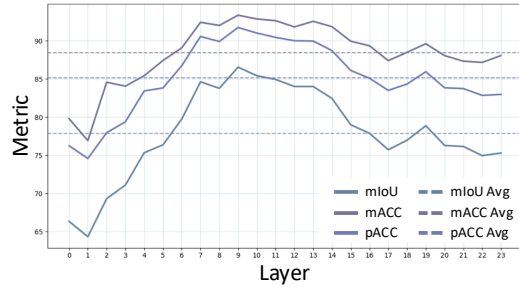


Figure 8: **Segmentation performance across layers.** Semantic grounding quality varies across MM-DiT layers, peaking in the middle blocks and specifically at the 9th layer.

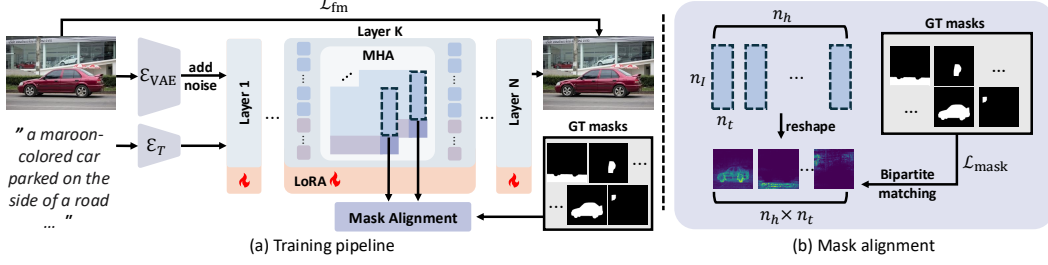


Figure 10: **Lightweight fine-tuning pipeline via mask alignment.** We introduce a simple yet effective *mask alignment* strategy that strengthens the I2T attention maps in the semantic grounding expert layer during additional diffusion fine-tuning with a LoRA adapter.

Unsupervised semantic grouping in $\langle \text{pad} \rangle$ tokens. This grouping behavior also emerges during unconditional generation, where the entire text sequence is set to $\langle \text{pad} \rangle$. As shown in Fig. 9 (b), attention maps from individual $\langle \text{pad} \rangle$ tokens—despite lacking explicit semantics—consistently attend to coherent regions, with different token positions specializing in different objects or parts. This allows the model to discover and segment meaningful groups, and these attention maps can thus serve as effective proposals for training-free unsupervised segmentation. To leverage this, we adapt the segmentation scheme in Fig. 7. Instead of using the attention map of a classname token, we select the best-matching mask from the proposals generated by the $\langle \text{pad} \rangle$ tokens.

3.5 Boosting Segmentation and Generation

Motivated by our layer-wise analysis and the causal perturbation study in Sec. 3.3, which identified a specific semantic grounding expert layer in MM-DiT as the key mediator of image-text alignment, we introduce a simple, lightweight LoRA fine-tuning scheme, called *mask alignment for segmentation and generation (MAGNET)*, that directly strengthens alignment in this layer. The goal is to boost semantic grouping in its I2T attention while preserving the model’s generation quality; empirically, this improves both zero-shot segmentation and image synthesis.

Specifically, we optimize two complementary losses. First, for each prediction $v_t(x_t)$ we apply a flow-matching loss $\mathcal{L}_{\text{FM}}$ to supervise the diffusion process. Second, to enhance the semantic grouping expressed by the expert layer’s I2T attention, we add a mask loss $\mathcal{L}_{\text{mask}}$.

Specifically, to compute $\mathcal{L}_{\text{mask}}$, we extract the $l \cdot H$ I2T attention maps from the semantic grounding expert layer for l text tokens and H heads and normalize each map to form candidate mask logits. We then perform bipartite matching between these candidates and the ground-truth masks to obtain one-to-one assignments [8]. Since we do not consider classification, we match masks solely based on the mask loss. For each matched pair, the loss is computed similarly as bipartite matching criterion:

$$\mathcal{L}_{\text{mask}} = \lambda_{\text{focal}} \mathcal{L}_{\text{focal}} + \lambda_{\text{dice}} \mathcal{L}_{\text{dice}}, \quad (12)$$

where $\mathcal{L}_{\text{focal}}$ is focal loss, $\mathcal{L}_{\text{dice}}$ is Dice loss, and $\lambda_{\text{focal}}, \lambda_{\text{dice}}$ weight their contributions.

Finally, the total objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{FM}} + \lambda_{\text{mask}} \mathcal{L}_{\text{mask}}, \quad (13)$$

with λ_{mask} controlling the strength of mask alignment.

4 Experiments

4.1 Implementation Details

For zero-shot inference, the diffusion process is fixed at timestep $t = 8$ of 28 using the flow-matching Euler discrete scheduler. Training uses 10k images from either SA-1B [30] or COCO [34], with captions generated by CogVLM [61] following SD3’s procedure [15]. Head-level attention maps are used for SA-1B and token-level maps for COCO to match mask granularity. Images are processed at 1024×1024 resolution, and each transformer layer is equipped with a LoRA module of rank $r = 16$, trained using AdamW with $\text{lr} = 1 \times 10^{-5}$, default β parameters, and weight decay. Training runs on two NVIDIA A6000 GPUs with per-device batch size 4 and gradient accumulation for an effective batch size of 16.

Model	Arch.	Train.	VOC20	Object	PC59	ADE	City
ProxyCLIP [33]	CLIP-H/14	-	83.3	49.8	39.6	24.2	42.0
CorrCLIP [66]	CLIP-H/14	-	91.8	52.7	47.9	28.8	49.9
DiffSegmter [60]	SD1.5	-	66.4	40.0	45.9	24.2	12.4
iSeg [54]	SD1.5	-	82.9	37.3	39.2	24.2	24.8
Seg4Diff	SD3	-	89.2	62.0	49.0	34.2	26.5
Seg4Diff	SD3.5	-	86.1	57.8	43.4	30.7	23.8
Seg4Diff	Flux.1-dev	-	83.1	50.6	38.2	23.9	17.1
Seg4Diff + MAGNET	SD3	SA-1B	89.1	62.0	49.1	34.7	25.4
Seg4Diff + MAGNET	SD3	COCO	89.8	62.9	51.2	35.2	26.0

Table 1: **Open-vocabulary semantic segmentation performance.** Cross-modal alignment in the I2T attention maps of the semantic grounding expert layer yields competitive results, further enhanced by mask alignment.

Model	Arch.	Train.	VOC21	PC59	Object	Stuff-27	City	ADE
ReCO [50]	CLIP-L/14	-	25.1	19.9	15.7	26.3	19.3	11.2
MaskCLIP [14]	CLIP-B/16	-	38.8	23.6	20.6	19.6	10.0	9.8
MaskCut [62]	DINO-B/8	-	53.8	43.4	30.1	41.7	18.7	35.7
DiffSeg [57]	SD1.5	-	49.8	48.8	23.2	44.2	16.8	37.7
DiffCut [13]	SSD-1B [19]	-	62.0	54.1	32.0	46.1	28.4	42.4
Seg4Diff	SD3	-	54.9	52.6	38.5	49.7	24.2	44.9
Seg4Diff	SD3.5	-	52.3	52.9	36.8	47.1	24.2	41.5
Seg4Diff + MAGNET	SD3	SA-1B	55.1	52.8	39.0	50.8	24.2	45.0
Seg4Diff + MAGNET	SD3	COCO	56.1	53.5	38.8	53.5	24.4	45.4

Table 2: **Unsupervised segmentation performance.** Although not specifically designed for unsupervised semantic segmentation, exploiting the emergent semantic grouping of <pad> tokens in the I2T attention maps achieves competitive results.

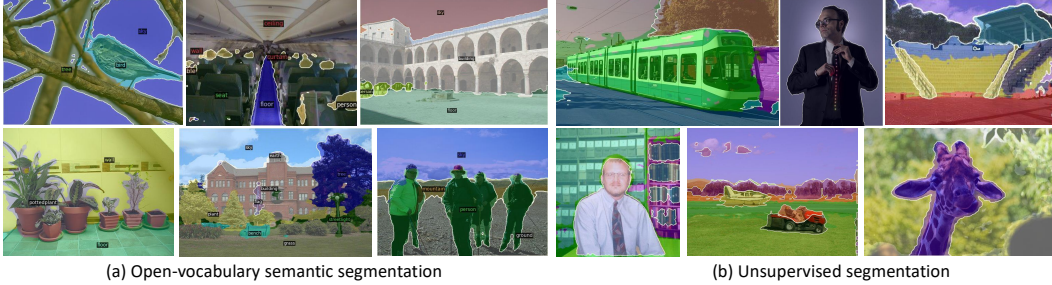


Figure 11: **Qualitative results on segmentation.** Through our Diff4Seg, MM-DiT demonstrates strong performance in (a) open-vocabulary semantic segmentation and (b) unsupervised segmentation. Additional results are provided in Appx. F.1.

4.2 Seg4Diff for Segmentation

Experimental setting. We evaluate our method on two tasks: open-vocabulary semantic segmentation and unsupervised segmentation. For open-vocabulary semantic segmentation, we report mIoU on the validation sets of PascalVOC [16], COCO-Object [34], Pascal Context-59 [16], and ADE20K [69], excluding the background class. We compare U-Net diffusion-based segmentation methods [60, 55] as well as CLIP-based approaches [33, 66]. We note that CLIP-based methods process entire classnames for prediction, which is not identical to our evaluation setting. For unsupervised segmentation, we report mIoU on the validation sets of PascalVOC, Pascal Context-59, COCO-Object, COCO-Stuff-27 [34], Cityscapes [12], and ADE20K. We adopt mask proposal evaluation protocol of DiffSeg [57], including the background class for PascalVOC, Pascal Context-59, and COCO-Object in line with DiffCut [13]. We compare with training-free methods built on diverse backbones, including CLIP [50, 70], DINO [62], and U-Net diffusion models [57, 13].

Results. Tab. 1 summarizes our open-vocabulary semantic segmentation results. Remarkably, using only this single layer without any refinement or postprocessing, our approach achieves competitive performance on Pascal VOC and COCO-Object dataset. Moreover, ours are robust on more complex datasets, where DiT architecture benefits from the larger and consistent spatial resolution throughout the entire layers. This result shows that MM-DiT inherently learned fine-grained semantic grounding capability during the generation process. On the other hand, as shown in Tab. 2, our method achieves competitive performance on unsupervised segmentation. This strong outcome indicates that an emergent knowledge of semantic grouping exists within the model’s multi-modal attention layers, which effectively function as learnable proxies for semantic classes—even when occupied by content-free <pad> tokens.

4.3 Seg4Diff for Boosting Segmentation and Generation

Experimental setting. We evaluate the image generation capability of Seg4Diff on three benchmark datasets: Pick-a-pic [31], MS-COCO [34], and SA-1B captions generated via CogVLM [61]. To assess text-image coherence, we utilize CLIPScore [21], which measures the semantic alignment between the text prompt and the generated image. We report T2I-CompBench++ [27] to further evaluate the fidelity and compositional quality of our generated images.

Method	Training	Pick-a-Pic	COCO	SA-1B	Mean
Baseline	–	27.0252	26.0638	28.3422	27.1437
+ MAGNET	SA-1B	27.0547	26.2318	28.4476	27.2447
+ MAGNET	COCO	27.0409	26.2319	28.5553	27.2760

Table 3: **CLIPScore on text-to-image generation benchmarks.** Mask alignment consistently improves alignment with text prompts across various datasets.

Method	Training	Attribute binding			Object relationships			Num.	Comp.
		Color	Shape	Texture	2D	3D	non		
Baseline	–	0.7864	0.5644	0.7200	0.2435	0.3318	0.3124	0.5566	0.3719
+ MAGNET	SA-1B	0.7836	0.5679	0.7252	0.2330	0.3151	0.3113	0.5460	0.3709
+ MAGNET	COCO	0.7919	0.5687	0.7260	0.2301	0.3234	0.3120	0.5584	0.3735

Table 4: **T2I-Compbench++ performance.** Mask alignment enhances attribute binding, object relationships, and compositional understanding compared to the baseline.

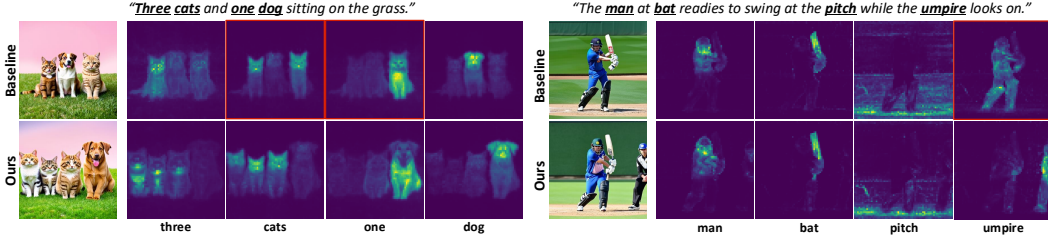


Figure 12: **Effects of the proposed mask alignment.** Mask alignment improves structural coherence and alignment between image and text.

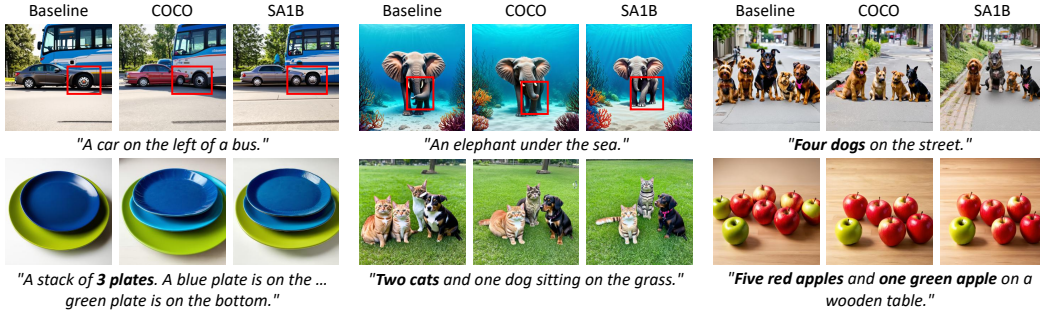


Figure 13: **Qualitative results on image generation.** Mask alignment improves compositional accuracy by reducing structural errors, correcting object counts, and refining colors and textures.

Results. As shown in Tab. 3 and 4, our fine-tuning method improves image generation quality over the baseline, achieving higher CLIPScores and better text-image alignment. On T2I-Compbench++, the COCO-trained model shows clear gains in attribute binding, numerical concepts, and complex scenes, which suggests our method effectively enhances the model’s ability to render objects with their correct attributes. However, because our training scheme do not target reasoning on object interactions, the baseline model maintains a marginal lead in rendering object relationships. Qualitative results in Fig.12 and 13 further confirm that ABoost corrects attention misalignments of the baseline, producing more accurate and plausible images. Overall, our strategy effectively enhances object-centric image generation as well as segmentation performance, as previously evidenced in Tab. 1 and 2.

5 Conclusion

This paper introduces Seg4Diff, a systematic framework that investigates the emergent capabilities of multi-modal diffusion transformers towards open-vocabulary semantic segmentation. Through an in-depth analysis of multi-modal attention mechanism, we identify semantic grounding expert layers that are critical for aligning textual semantics with corresponding image regions. We demonstrate that attention maps from these layers can be directly used to achieve competitive zero-shot segmentation performance. To further enhance this capability, we introduce a lightweight fine-tuning method that strengthens the semantic grouping in these expert layers. This targeted approach not only leads to meaningful improvements in segmentation quality but also enhances the model’s generative fidelity. Our findings reveal that semantic alignment is an intrinsic property of diffusion transformers that can be selectively amplified with minimal computational cost. This work paves the way toward unified models that bridge the gap between generation and perception, delivering both high-quality image synthesis and accurate semantic understanding.

Acknowledgments and Disclosure of Funding

This research was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190075, RS-2024-00509279, RS-2025-II212068, RS-2023-00227592, RS-202502214479, RS-2024-00457882, RS-2025-25441838, RS-2025-25441838, RS-2025-02214479, RS-2025-02217259), the Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (RS-2024-00345025, RS-2024-00333068, RS-2023-00222280, RS-2023-00266509), and National Research Foundation of Korea (RS-2024-00346597), and the ETH AI Center through an ETH AI Center postdoctoral fellowship to Sunghwan Hong.

References

- [1] Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *European Conference on Computer Vision*, pages 1–17. Springer, 2024.
- [2] Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. Stable flow: Vital layers for training-free image editing. *arXiv preprint arXiv:2411.14430*, 2024.
- [3] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024.
- [4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.
- [5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023.
- [6] Lingling Cai, Kang Zhao, Hangjie Yuan, Yingya Zhang, Shiwei Zhang, and Kejie Huang. Freemask: Rethinking the importance of attention masks for zero-shot video editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1898–1906, 2025.
- [7] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22560–22570, 2023.
- [8] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021.
- [9] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. Cats: Cost aggregation transformers for visual correspondence. *Advances in Neural Information Processing Systems*, 34:9011–9023, 2021.
- [10] Seokju Cho, Sunghwan Hong, and Seungryong Kim. Cats++: Boosting cost aggregation with convolutions and transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7174–7194, 2022.
- [11] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024.
- [12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.
- [13] Paul Couairon, Mustafa Shukor, Jean-Emmanuel Haugeard, Matthieu Cord, and Nicolas Thome. Diffcut: Catalyzing zero-shot semantic segmentation with diffusion features and recursive normalized cut. *arXiv preprint arXiv:2406.02842*, 2024.
- [14] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10995–11005, 2023.

- [15] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.
- [16] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.
- [17] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2024.
- [18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [19] Yatharth Gupta, Vishnu V Jaddipal, Harish Prabhala, Sayak Paul, and Patrick Von Platen. Progressive knowledge distillation of stable diffusion xl using layer level loss. *arXiv preprint arXiv:2401.02677*, 2024.
- [20] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.
- [21] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [23] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- [24] Sunghwan Hong, Jisu Nam, Seokju Cho, Susung Hong, Sangryul Jeon, Dongbo Min, and Seungryong Kim. Neural matching fields: Implicit representation of matching fields for visual correspondence. *Advances in Neural Information Processing Systems*, 35:13512–13526, 2022.
- [25] Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. *Advances in Neural Information Processing Systems*, 37:66743–66772, 2024.
- [26] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024.
- [27] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-Image Generation. *IEEE Transactions on Pattern Analysis Machine Intelligence*, pages 1–17, January 5555.
- [28] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for open-vocabulary segmentation. In *European Conference on Computer Vision*, pages 299–317. Springer, 2025.
- [29] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024.
- [30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.
- [31] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation, 2023.
- [32] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. *arXiv preprint arXiv:2004.10102*, 2020.
- [33] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. *arXiv preprint arXiv:2408.04883*, 2024.

- [34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- [35] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.
- [36] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.
- [37] Run Luo, Zikai Song, Lintao Ma, Jinlin Wei, Wei Yang, and Min Yang. Diffusiontrack: Diffusion model for multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3991–3999, 2024.
- [38] Benyuan Meng, Qianqian Xu, Zitai Wang, Xiaochun Cao, and Qingming Huang. Not all diffusion model activations have been evaluated as discriminative features. *Advances in Neural Information Processing Systems*, 37:55141–55177, 2024.
- [39] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models, 2022.
- [40] Jisu Nam, Gyuseong Lee, Sunwoo Kim, Hyeonsu Kim, Hyounghwon Cho, Seyeon Kim, and Seungryong Kim. Diffusion model for dense matching. *arXiv preprint arXiv:2305.19094*, 2023.
- [41] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.
- [42] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.
- [43] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- [44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- [45] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- [46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [48] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dream-booth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023.
- [49] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.
- [50] Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer. *Advances in Neural Information Processing Systems*, 35:33754–33767, 2022.
- [51] Heeseong Shin, Chaehyun Kim, Sunghwan Hong, Seokju Cho, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Towards open-vocabulary semantic segmentation without semantic labels. *arXiv preprint arXiv:2409.19846*, 2024.

- [52] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- [53] Stability AI. Stable diffusion 3.5. <https://github.com/Stability-AI/sd3.5>, 2024.
- [54] Lin Sun, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. iseg: An iterative refinement-based framework for training-free segmentation. *arXiv preprint arXiv:2409.03209*, 2024.
- [55] Lin Sun, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. iseg: An iterative refinement-based framework for training-free segmentation, 2024.
- [56] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023.
- [57] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3554–3563, June 2024.
- [58] Pedro Vélez, Luisa F Polanía, Yi Yang, Chuhan Zhang, Rishabh Kabra, Anurag Arnab, and Mehdi SM Sajjadi. From image to video: An empirical study of diffusion representations. In *ICCV*, 2025.
- [59] Chaoyang Wang, Xiangtai Li, Lu Qi, Henghui Ding, Yunhai Tong, and Ming-Hsuan Yang. Semflow: Binding semantic segmentation and image synthesis via rectified flow. *Advances in Neural Information Processing Systems*, 37:138981–139001, 2024.
- [60] Jinglong Wang, Xiawei Li, Jing Zhang, Qingyuan Xu, Qin Zhou, Qian Yu, Lu Sheng, and Dong Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter, 2024.
- [61] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024.
- [62] Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3124–3134, 2023.
- [63] Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. What matters when repurposing diffusion models for general dense perception tasks? *arXiv preprint arXiv:2403.06090*, 2024.
- [64] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023.
- [65] Heeji Yoon, Jaewoo Jung, Junwan Kim, Hyungyu Choi, Heeseong Shin, Sangbeom Lim, Honggyu An, Chaehyun Kim, Jisang Han, Donghyun Kim, et al. Visual representation alignment for multimodal large language models. *arXiv preprint arXiv:2509.07979*, 2025.
- [66] Dengke Zhang, Fagui Liu, and Quan Tang. Correlclip: Reconstructing correlations in clip with off-the-shelf foundation models for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2411.10086*, 2024.
- [67] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36:45533–45547, 2023.
- [68] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5729–5739, 2023.
- [69] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019.
- [70] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022.
- [71] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11175–11185, 2023.

Appendix

A Further implementation detail

- Model configuration
- Text prompts for analysis
- Attention perturbation
- Segmentation evaluation metrics
- Further details on image quality metrics

B Additional experiments

- Ablation on timestep choice
- Comparison on I2T and T2I attention maps
- Ablation on normalization method
- Segmentation performance per attention layer and head
- Quantitative results on attention perturbation

C Generalization to other baselines

- Stable Diffusion 3.5
- Flux.1-dev

D Additional visualizations

- Image-to-text (I2T) and text-to-image (T2I) attention map
- Per-head attention maps
- Statistics of attention score and norm
- PCA visualization of attention features
- Emergent behavior of <pad> tokens

E Additional qualitative results for generation

- Extended I2T attention perturbation results
- Extended I2T attention perturbation guidance results
- Image generation results of trained models

F Additional qualitative results for segmentation

- Open-vocabulary semantic segmentation results
- Unsupervised segmentation results

G Limitations of our method

A Further implementation detail

A.1 Model configuration

We utilize the Stable Diffusion 3 (SD3) [15] model for our main analysis. SD3 originally incorporates three vision-language text encoders: CLIP-G/14, CLIP-L/14 [44], and T5-XXL [45]. Due to memory constraints, we disable the T5 encoder and use only the first 77 tokens from the two CLIP encoders. We leverage the attention score from timestep $t = 8$ of 28, where the attention logits are semantically grouped well. Further ablation on timesteps can be found at Sec. B.1. We used classifier-free guidance [23] with scale 7.5 for generation if not specified.

For training and segmentation evaluation, the input images are center-cropped if non-square and resized to a resolution of 1024×1024 . After VAE encoding, the image is further downsampled to 64×64 latent. The resulting attention maps are bilinear upsampled back to the original image size for segmentation evaluation.

A.2 Text prompts for analysis

We borrowed text prompts from DrawBench [49], where diverse categories are included to assess the capability of generative models. From 200 prompts in total, we randomly sampled 50 prompts for our analysis. The full list of selected prompts is shown in Fig.14.

1. A red colored car.	27. An elephant is behind a tree. You can see the trunk on one side and the back legs on the other.
2. A black colored dog.	28. A pear cut into seven pieces arranged in a ring.
3. A blue colored dog.	29. Rbepfraigerator.
4. A red colored banana.	30. Dining table.
5. A white colored sandwich.	31. An instrument used for cutting cloth, paper, and other thin material, consisting of two blades laid one on top of the other and fastened in the middle so as to blow them to be opened and closed by thumb and finger inserted through rings on the end of the handles.
6. A yellow colored giraffe.	32. A bicycle on top of a boat.
7. A green cup and a blue cell phone.	33. A car on the left of a bus.
8. A horse riding an astronaut.	34. Acersecomicke.
9. A shark in the desert.	35. Artophagous.
10. Three cars on the street.	36. Backlotter.
11. One dog on the street.	37. A photo of a confused grizzly bear in calculus class.
12. Two dogs on the street.	38. Photo of an athlete cat explaining its latest scandal at a press conference to journalists.
13. One cat and one dog sitting on the grass.	39. Hyper-realistic photo of an abandoned industrial site during a storm.
14. Three cats and one dog sitting on the grass.	40. A real life photography of super mario, 8k Ultra HD.
15. Three cats and two dogs sitting on the grass.	41. Colouring page of large cats climbing the eifel tower in a cyberpunk future.
16. A triangular pink stop sign. A pink stop sign in the shape of a triangle.	42. Photo of a mega Lego space station inside a kid's bedroom.
17. An illustration of a small green elephant standing behind a large red mouse.	43. A spider with a moustache bidding an equally gentlemanly grasshopper a good day during his walk to work.
18. A small blue book sitting on a large red book.	44. A bridge connecting Europe and North America on the Atlantic Ocean, bird's eye view.
19. A stack of 3 cubes. A red cube is on the top, sitting on a red cube. The red cube is in the middle, sitting on a green cube. The green cube is on the bottom.	45. A magnifying glass over a page of a 1950s batman comic.
20. A stack of 3 books. A green book is on the top, sitting on a red book. The red book is in the middle, sitting on a blue book. The blue book is on the bottom.	46. A realistic photo of a Pomeranian dressed up like a 1980s professional wrestler with neon green and neon orange face paint and bright green wrestling tights with bright orange boots.
21. A small vessel propelled on water by oars, sails, or an engine.	47. A sign that says 'Hello World'.
22. A large plant-eating domesticated mammal with solid hoofs and a flowing mane and tail, used for riding, racing, and to carry and pull loads.	48. A sign that says 'Diffusion'.
23. An American multinational technology company that focuses on artificial intelligence, search engine, online advertising, cloud computing, computer software, quantum computing, e-commerce, and consumer electronics.	49. New York Skyline with 'NeurIPS' written with fireworks on the sky.
24. A large thick-skinned semiaquatic African mammal, with massive jaws and large tusks.	50. New York Skyline with 'Google Research Pizza Cafe' written with fireworks on the sky.
25. A machine resembling a human being and able to replicate certain human movements and functions automatically.	
26. A grocery store refrigerator has pint cartons of milk on the top shelf, quart cartons on the middle shelf, and gallon plastic jugs on the bottom shelf.	

Figure 14: Selected prompts for our analysis.

A.3 Attention perturbation

To assess the importance of image-to-text (I2T) attention alignment across layers, we applied a Gaussian blur along the text token dimension of the I2T attention map. We used 1D Gaussian kernel with standard deviation $\sigma = 9$ and kernel size $k = 5$ to accommodate the typical text token length. The blur was applied to the attention logits after softmax, using reflective padding to preserve the total attention mass per image token.

A.4 Segmentation evaluation metrics

To assess layer-wise segmentation performance in Sec. 3.4, we use three standard metrics: pixel accuracy (pACC), mean accuracy (mACC), and mean Intersection-over-Union (mIoU). Each provides a progressively more rigorous evaluation based on pixel-level true positives (TP), false positives (FP), and false negatives (FN) for the number classes N . pACC reports the fraction of correctly classified pixels over entire classes, which can be easily skewed by large classes like background. mACC averages per-class accuracy, $\frac{TP}{TP+FN}$, treating all classes equally yet still ignoring FP; mIoU is the most comprehensive, computing intersection-over-union, $\frac{TP}{TP+FP+FN}$.

$$\text{pACC} = \frac{\sum TP}{\sum TP + FN} \quad \text{mACC} = \frac{1}{N} \sum \frac{TP}{TP + FN} \quad \text{mIoU} = \frac{1}{N} \sum \frac{TP}{TP + FP + FN} \quad (14)$$

For unsupervised segmentation, we forward a noised input image with a null prompt to obtain the <pad> token attention maps. These maps are then greedily merged based on KL-divergence, following the procedure in [57]. The resulting mask proposals are evaluated via bipartite matching with ground-truth masks, while unmatched proposals are treated as false negatives.

A.5 Further details on image quality metrics.

For Pick-a-Pic, we generate 500 images for five random seeds respectively. We generated 5,000 images for MS-COCO and 1,000 images for SA-1B captions.

B Additional experiments

B.1 Ablation on timestep choice

Fig. 15 shows segmentation performance throughout different timesteps applied to the input image. Both PascalVOC and COCO-Object demonstrates the best performance on $t = 8$ of 28, where we report the segmentation performance.

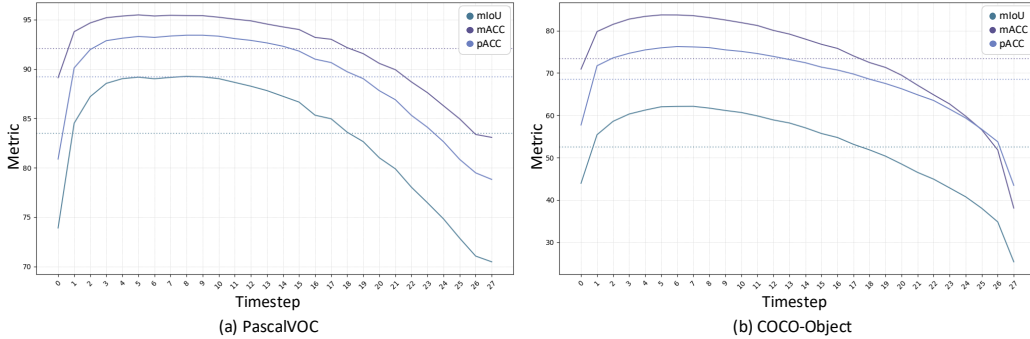


Figure 15: Segmentation performance across denoising timesteps.

B.2 Comparison on I2T and T2I attention maps

Fig. 16 presents attention maps for I2T and T2I directions corresponding to the highlighted keywords. The I2T maps exhibit more complete and contiguous object masks, whereas T2I tends to capture only partial or attenuated regions. This discrepancy likely stems from the distinct aggregation roles: I2T attention directly updates image tokens based on textual queries, while T2I updates text tokens conditioned on visual features. Given that the diffusion process ultimately operates on image tokens to synthesize outputs, it is reasonable that I2T attention aligns more strongly with semantically grounded image regions. Although this observation does not constitute a formal proof, the consistency of the qualitative patterns, which is further visualized in Appx. D.1 across prompts supports this interpretation.

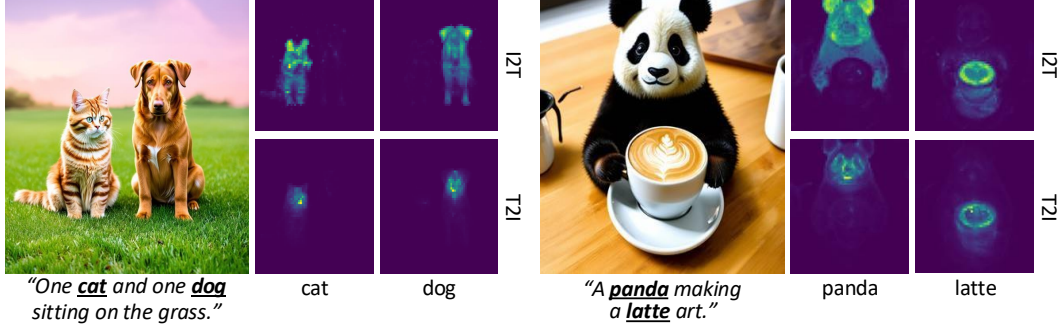


Figure 16: Comparison on I2T and T2I attention maps.

B.3 Ablation on normalization method

We conduct an ablation study to determine the optimal method for normalizing attention scores. As shown in Table 5, we evaluate the open-vocabulary semantic segmentation performance across four configurations: using scores before or after the softmax function, with and without additional min-max normalization to scale the logit between 0 to 1. The results indicate that using the raw scores directly after the softmax function yields the best performance. Consequently, we adopt this scheme for all subsequent experiments.

softmax	min-max	VOC	Object
X	X	83.2	55.1
X	✓	85.2	53.5
✓	X	89.0	61.8
✓	✓	<u>88.3</u>	<u>57.0</u>

Table 5: Ablation on attention score normalization methods.

B.4 Segmentation performance per attention layer and head

We present semantic segmentation results obtained by leveraging the attention scores from individual layers and heads in Fig. 17. For a comprehensive evaluation, we measured mIoU on the PASCAL VOC dataset [16], including the background class. Our analysis reveals that the middle layers consistently exhibit superior segmentation performance compared to early or late layers.

Segmentation Performance per Head																								
Layer 0	11.0	18.5	11.7	16.0	17.8	16.4	15.2	10.2	17.8	15.9	15.7	16.3	8.8	21.2	15.2	17.8	20.5	17.6	16.6	17.9	13.4	17.7	19.6	18.2
Layer 1	16.5	12.9	15.1	19.2	16.7	0.0	0.0	21.1	15.7	15.7	13.3	8.3	19.8	0.0	13.0	5.3	18.3	17.3	17.5	10.6	13.6	14.5	16.3	9.2
Layer 2	21.5	16.9	15.7	14.2	20.8	17.0	18.1	13.5	18.3	13.5	24.7	14.8	15.3	15.1	17.8	12.8	14.8	19.5	17.1	19.1	17.7	17.1	17.1	17.5
Layer 3	0.0	17.2	19.6	18.8	20.5	0.0	0.1	23.5	21.9	12.4	30.5	19.2	17.5	32.4	0.0	17.0	19.4	0.1	27.2	16.5	15.4	18.1	0.1	23.5
Layer 4	15.3	11.8	20.5	17.9	24.1	21.2	22.3	14.7	23.1	20.4	23.2	23.0	13.3	21.8	22.9	20.1	42.4	21.1	17.3	19.0	16.5	20.5	14.2	16.7
Layer 5	30.2	34.7	32.2	18.3	26.3	17.8	26.4	36.4	22.8	23.4	23.4	21.5	24.6	32.8	29.6	18.9	19.6	14.6	25.8	21.9	29.9	26.8	22.8	22.4
Layer 6	26.8	33.2	28.6	28.8	25.5	19.2	28.1	23.1	34.2	34.6	21.4	25.0	24.9	28.2	24.0	18.1	30.1	34.9	24.6	28.3	26.7	27.6	32.2	27.8
Layer 7	38.5	51.8	26.0	40.7	42.2	33.8	35.9	28.9	21.6	24.1	31.4	33.0	17.5	34.7	31.5	45.5	30.1	24.6	29.9	16.5	31.9	36.3	29.0	32.8
Layer 8	43.5	35.9	32.0	16.2	43.1	34.4	18.2	36.4	34.3	40.2	38.3	40.5	44.5	26.1	25.3	33.6	21.4	44.1	41.9	37.2	29.3	32.5	39.0	26.3
Layer 9	23.9	37.2	25.1	31.1	29.3	28.7	32.4	22.8	28.0	28.8	26.0	24.0	22.9	26.4	27.2	26.8	32.2	32.4	37.8	32.0	31.7	25.7	37.1	24.5
Layer 10	28.8	31.6	34.3	35.8	36.3	20.9	36.4	18.3	36.6	36.7	29.8	42.4	29.9	31.3	24.1	27.6	30.9	16.0	34.4	31.8	31.5	40.9	37.1	27.4
Layer 11	32.4	30.1	25.7	30.4	33.3	26.2	19.6	25.0	26.1	25.8	21.2	31.6	37.8	31.6	28.4	34.8	33.4	28.2	31.8	22.4	33.7	25.7	29.2	27.2
Layer 12	37.6	15.8	28.4	19.8	36.5	30.0	22.9	30.7	22.1	38.7	35.3	19.5	14.4	18.5	28.7	13.1	27.4	26.5	37.0	39.6	32.2	16.0	32.5	23.6
Layer 13	34.8	33.0	25.1	34.6	31.6	26.2	35.0	29.1	31.6	33.8	35.0	31.8	35.2	39.2	29.9	30.1	25.7	31.7	31.4	35.5	32.1	28.3	27.0	34.0
Layer 14	31.5	25.3	23.3	33.2	30.6	34.1	31.9	35.6	30.1	33.3	29.5	24.9	35.7	19.1	32.9	33.9	32.2	35.6	29.2	34.1	31.7	32.5	34.0	36.7
Layer 15	23.6	23.1	27.3	30.1	25.9	26.6	29.7	29.8	24.0	23.5	30.8	29.2	27.6	29.6	31.1	29.6	28.8	26.6	28.8	27.9	25.3	25.2	28.0	24.3
Layer 16	24.1	18.4	28.6	22.9	25.1	24.5	25.7	26.2	27.4	25.3	24.2	23.0	28.9	25.4	16.4	24.5	27.7	27.3	25.2	24.4	25.2	25.9	25.1	25.8
Layer 17	30.4	24.8	24.2	20.1	26.6	25.7	26.2	20.3	17.6	25.5	24.8	26.8	25.4	26.5	25.1	21.1	24.1	21.8	27.9	26.8	26.5	25.8	26.0	25.2
Layer 18	29.6	30.1	27.2	19.8	28.7	27.4	29.2	22.4	30.8	30.6	28.5	28.2	32.5	25.2	29.9	30.6	26.7	26.1	27.6	26.4	26.3	27.4	26.8	29.1
Layer 19	28.7	28.5	29.0	26.7	32.6	14.9	29.7	24.4	20.1	29.1	27.1	28.4	23.9	14.5	29.9	23.8	30.9	16.4	30.8	30.1	24.0	25.2	25.9	30.3
Layer 20	25.5	25.4	23.7	25.1	24.9	28.4	18.7	24.2	25.4	25.2	30.3	25.7	24.7	24.9	24.3	24.4	25.5	28.1	24.1	25.5	22.4	27.0	19.0	16.5
Layer 21	27.4	23.9	25.5	27.3	26.2	20.5	27.1	26.5	21.0	24.9	25.9	23.6	24.4	29.2	26.9	27.9	25.7	25.6	23.8	26.7	26.6	19.8	23.8	19.3
Layer 22	20.2	15.9	12.9	24.6	6.7	22.4	17.9	24.5	23.3	25.8	28.1	17.4	15.5	26.8	11.7	20.7	24.3	26.1	23.8	18.1	27.5	28.3	28.2	20.2
Layer 23	25.3	24.1	21.7	18.8	28.2	18.8	21.7	21.8	21.8	19.1	17.8	20.7	23.4	28.1	24.8	21.2	23.7	26.4	24.8	28.8	20.2	24.2	23.8	26.8
	Head 0	Head 1	Head 2	Head 3	Head 4	Head 5	Head 6	Head 7	Head 8	Head 9	Head 10	Head 11	Head 12	Head 13	Head 14	Head 15	Head 16	Head 17	Head 18	Head 19	Head 20	Head 21	Head 22	Head 23

Figure 17: mIoU score of each head.

B.5 Quantitative results on attention perturbation

We use CLIP-I, CLIP-T, and DINO scores to evaluate generation quality and alignment. CLIP-I measures cosine similarity between CLIP embeddings of generated and reference images, reflecting high-level perceptual fidelity. CLIP-T compares the CLIP embedding of a generated image with that of the conditioning text, assessing image–text alignment. DINO instead computes cosine similarity between self-supervised vision embeddings of generated and reference images, offering a language-free measure of semantic similarity. Higher values indicate better fidelity or alignment, with CLIP-I and DINO focusing on image–image consistency and CLIP-T on image–text consistency.

Aligned with the qualitative analysis in Sec. 3.3, 9th layer, which we designated as a semantic grounding expert in SD3, shows a noticeable drop on image fidelity score when perturbed.

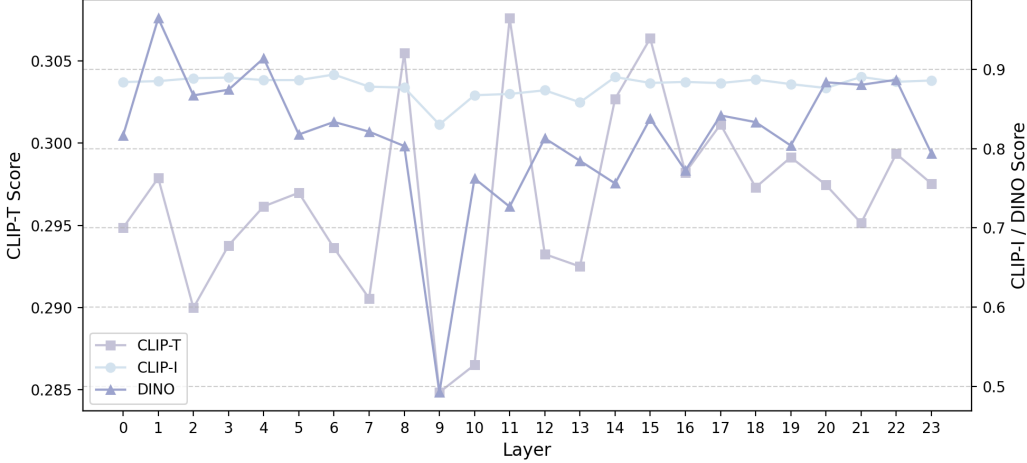


Figure 18: Image fidelity scores under layer-wise perturbations on SD3.

C Generalization to other baselines

While we mainly leverage Stable Diffusion 3 (SD3) [15] in our main analysis, we also apply our analysis to the other MM-DiT variants, Stable Diffusion 3.5 (SD3.5) [53] and Flux.1-dev [3]. We can similarly observe the correlation between value norm and segmentation performance among layers for both models. While SD3.5 appears to have layer 9, identical to SD3, to exhibit strong semantic grounding, Flux shows layer 12 and 17 to have a similar tendency. This hints that our observation and methodology are not proprietary for SD3, but can be applied to other DiT-based diffusion models with multi-modal attention, highlighting the generalizability of our insights and findings.

C.1 Stable Diffusion 3.5

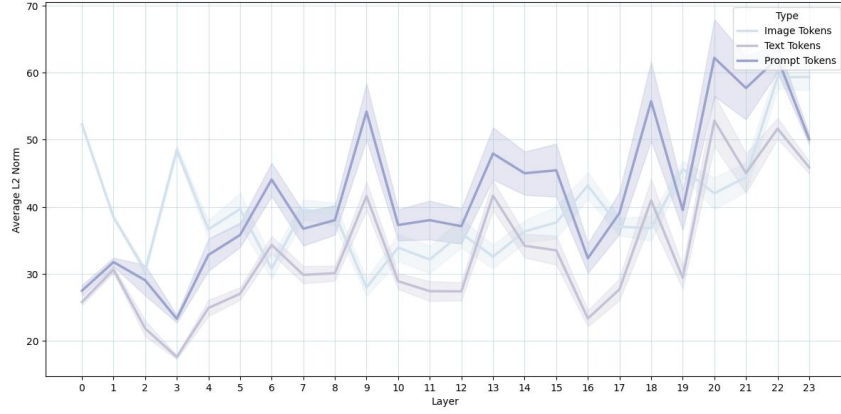


Figure 19: Average L2-Norm of values across layers on SD3.5.

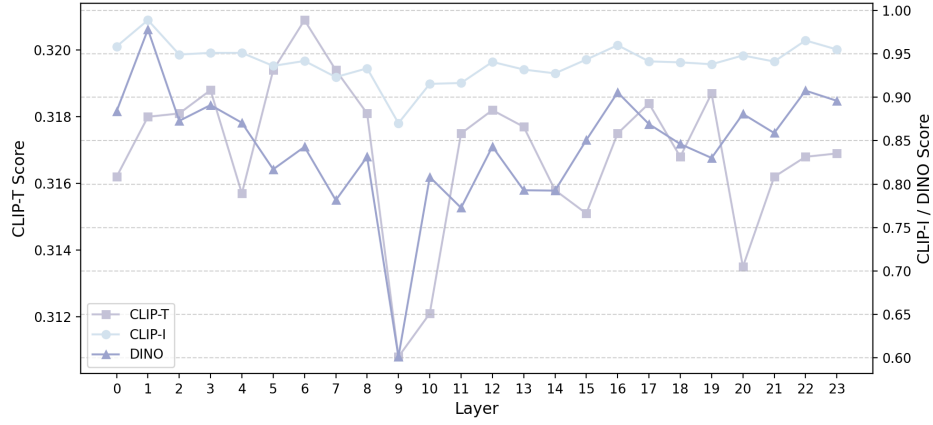


Figure 20: Image fidelity scores under layer-wise perturbations on SD3.5.

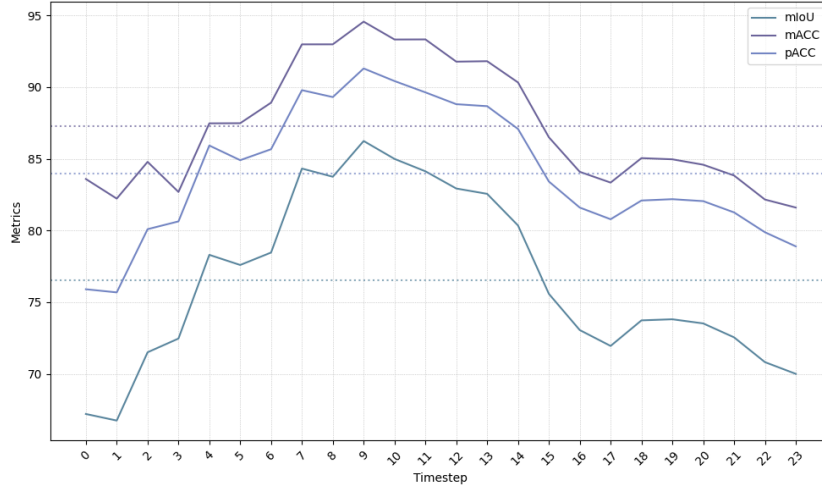


Figure 21: Segmentation performance across layers on SD3.5.

C.2 Flux.1-dev

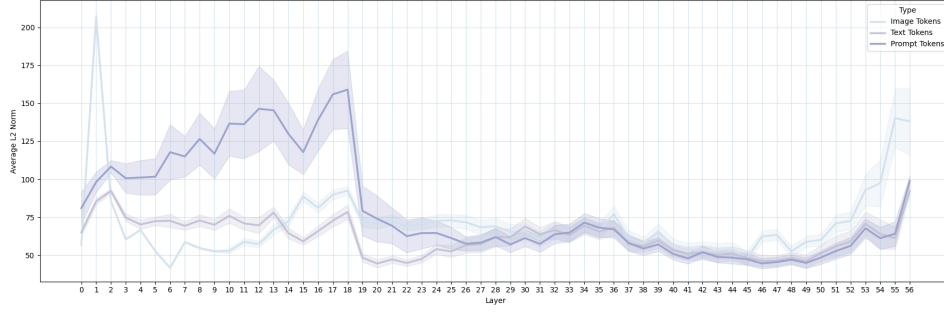


Figure 22: Average L2-Norm of values across layers on Flux.

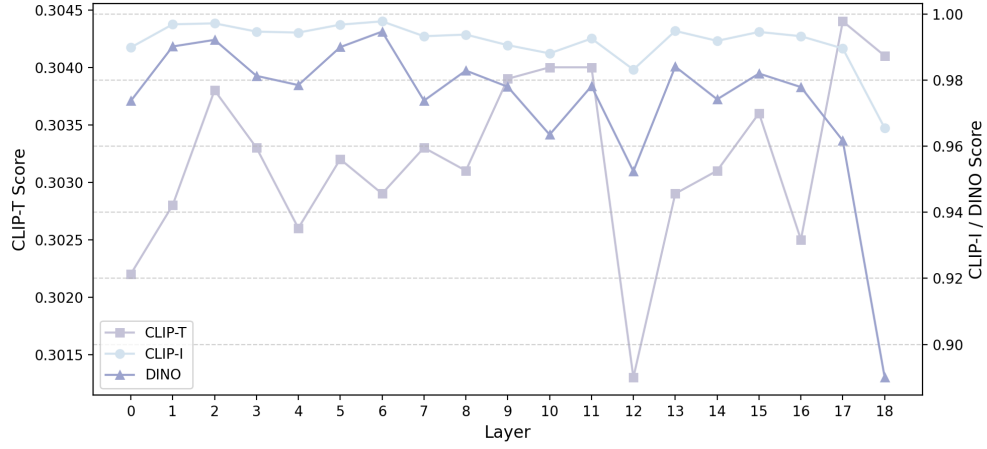


Figure 23: Image fidelity scores under layer-wise perturbations on Flux. We evaluate only for the first 19 layers, which employs MM-DiT architecture.

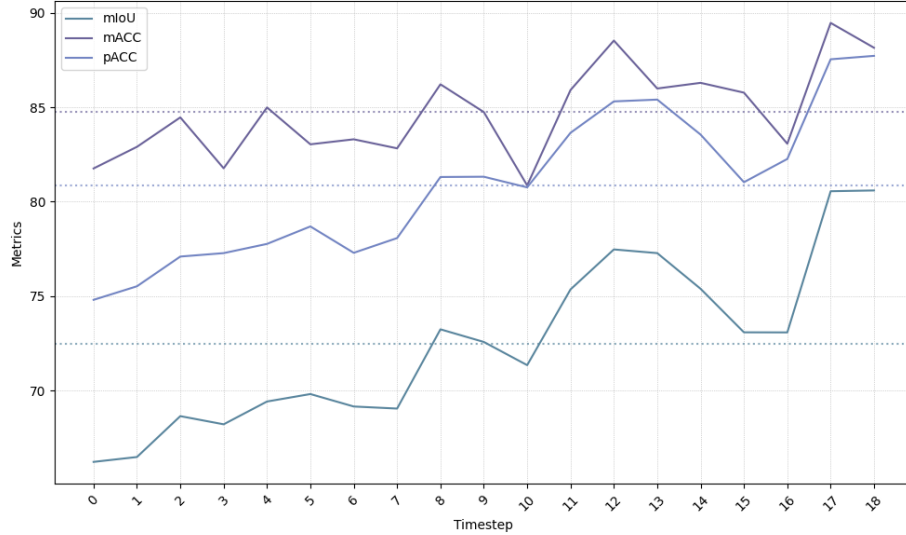


Figure 24: Segmentation performance across layers on Flux.

D Additional visualizations

D.1 Image-to-text (I2T) and text-to-image (T2I) attention map

We provide extended visualizations of image-to-text (I2T) and text-to-image (T2I) attention maps in Fig. 25. The maps are taken from layer 9, where we observe strong semantic alignment between visual and textual modalities. These results offer insight into the emergent cross-modal grounding dynamics of the multi-modal diffusion transformer (MM-DiT).

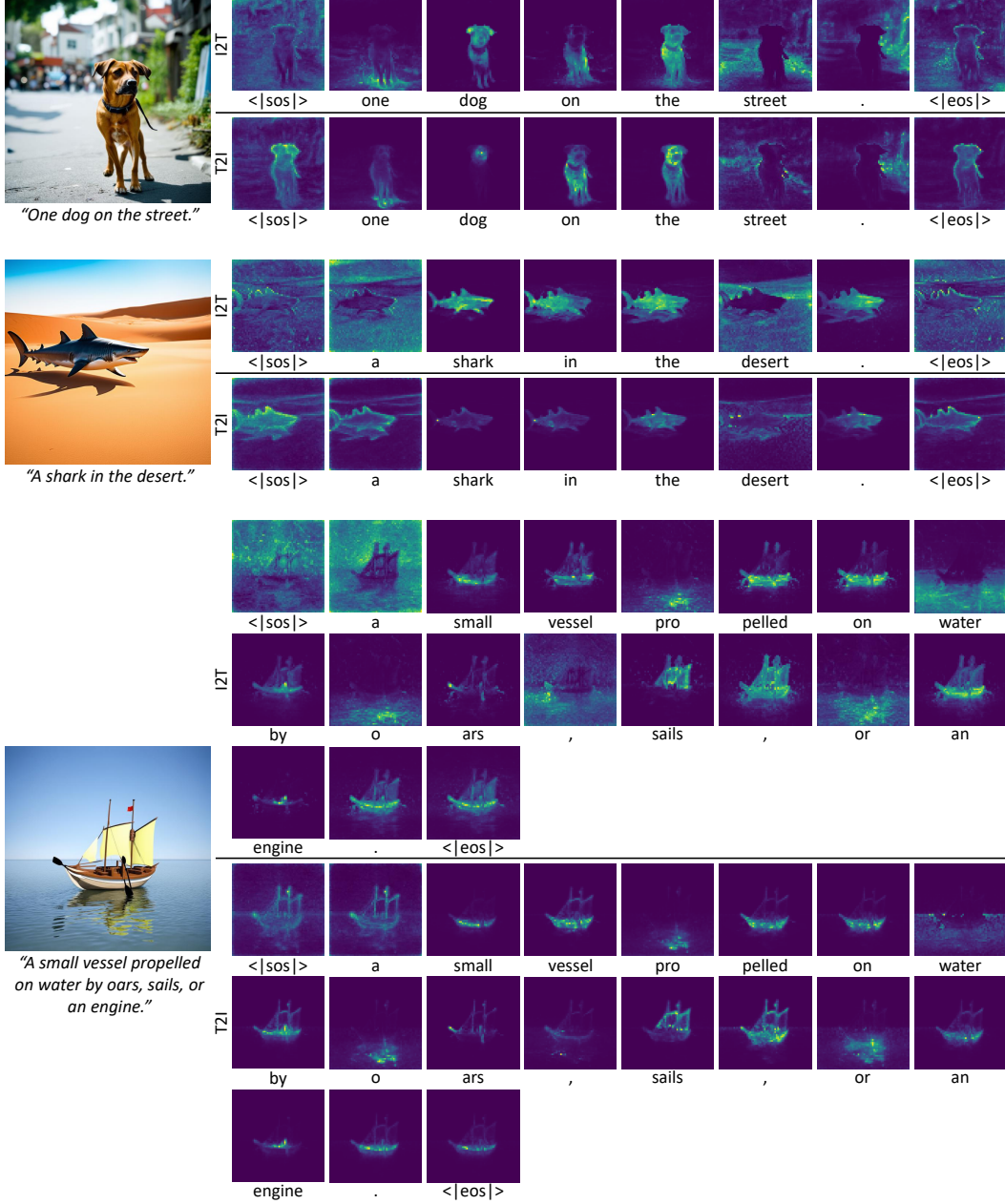


Figure 25: **Image-to-Text** attention map visualization of all prompt tokens.

D.2 Per-head attention maps

We present comprehensive visualizations of attention maps for individual heads in selected layers, as shown in Fig.26 and Fig.27. These results illustrate that each attention head focuses on distinct image regions. This effect is particularly pronounced in layer 9, where individual heads attend to different parts of the corresponding semantic region.

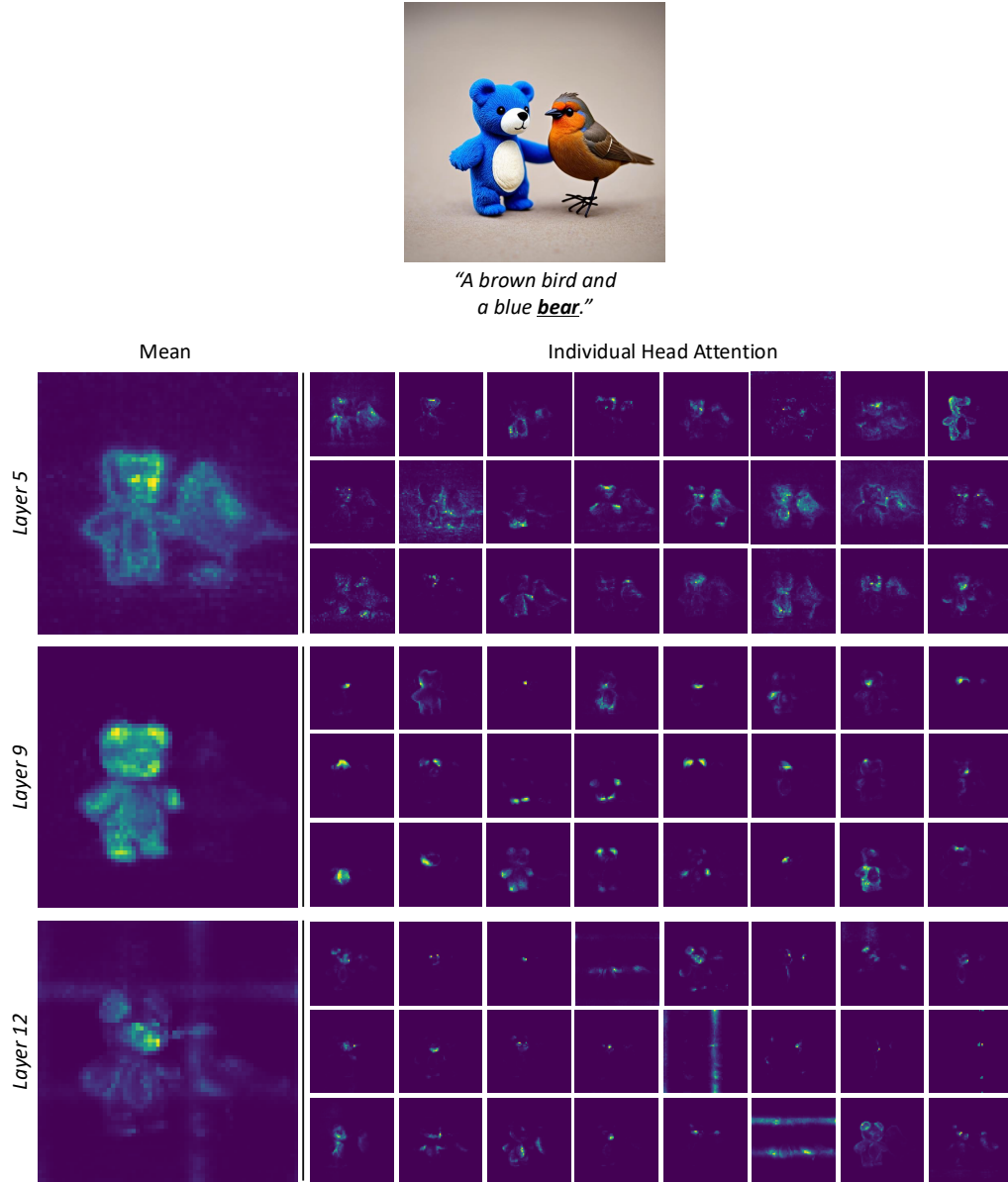


Figure 26: Visualization of attention scores for all heads.



*"Illustration of a mouse using
a **mushroom** as an umbrella."*

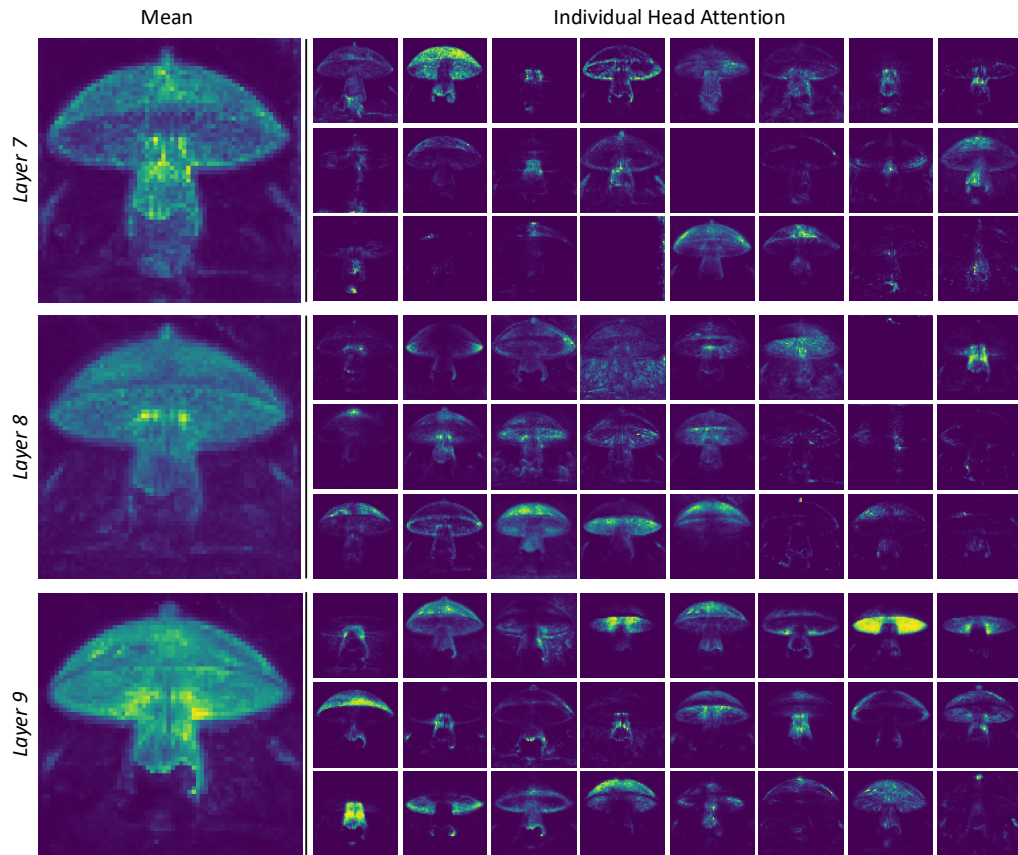


Figure 27: Visualization of attention scores for all heads.

D.3 Statistics of attention score and norm

We present the statistics of L2 norms of value-projected features across all layers and heads in Fig.28. Notably, layer 9 exhibits consistently large norms for text tokens, suggesting that text features strongly dominate this layer. When restricting the analysis to the actual prompt tokens only, this trend becomes even more pronounced.

On the other hand, we observe large image token norms in specific heads within layers 7 and 8, which coincide with the heads that show strong segmentation performance in Fig. 17. This suggests that these heads are highly responsive to image-specific features and may play a crucial role in localizing visual semantics prior to cross-modal alignment with text.

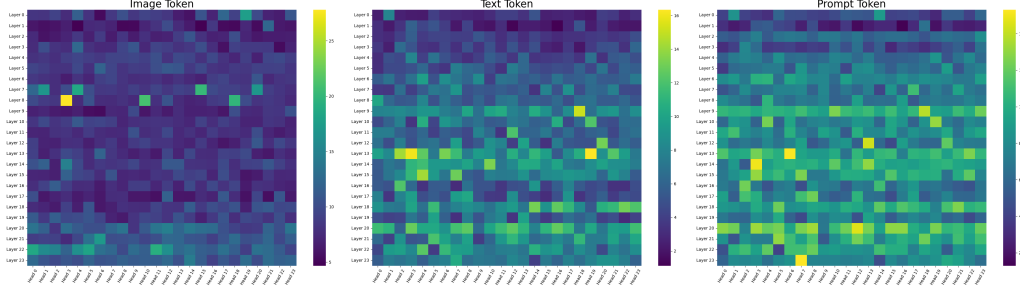


Figure 28: Average L2 norm of value tokens by layer and head.

D.4 PCA visualization of attention features

We provide PCA visualizations of the query-, key-, and value-projected image features across all layers in Fig. 29. Most layers exhibit a strong positional bias, whereas layer 9 reveals distinct semantic grouping. This suggests that image features become semantically well-grounded at layer 9, enabling more meaningful cross-modal interactions.

D.5 Emergent behavior of <pad> tokens

In Fig. 30, we visualize all 77 tokens under the unconditional generation setting described in Sec. 3.4, which includes <sos>, <eos>, and 75 <pad> tokens. Remarkably, we observe that individual <pad> tokens attend to distinct semantic regions, despite the absence of explicit semantic information in the text prompt.



"A bicycle on top of the boat"

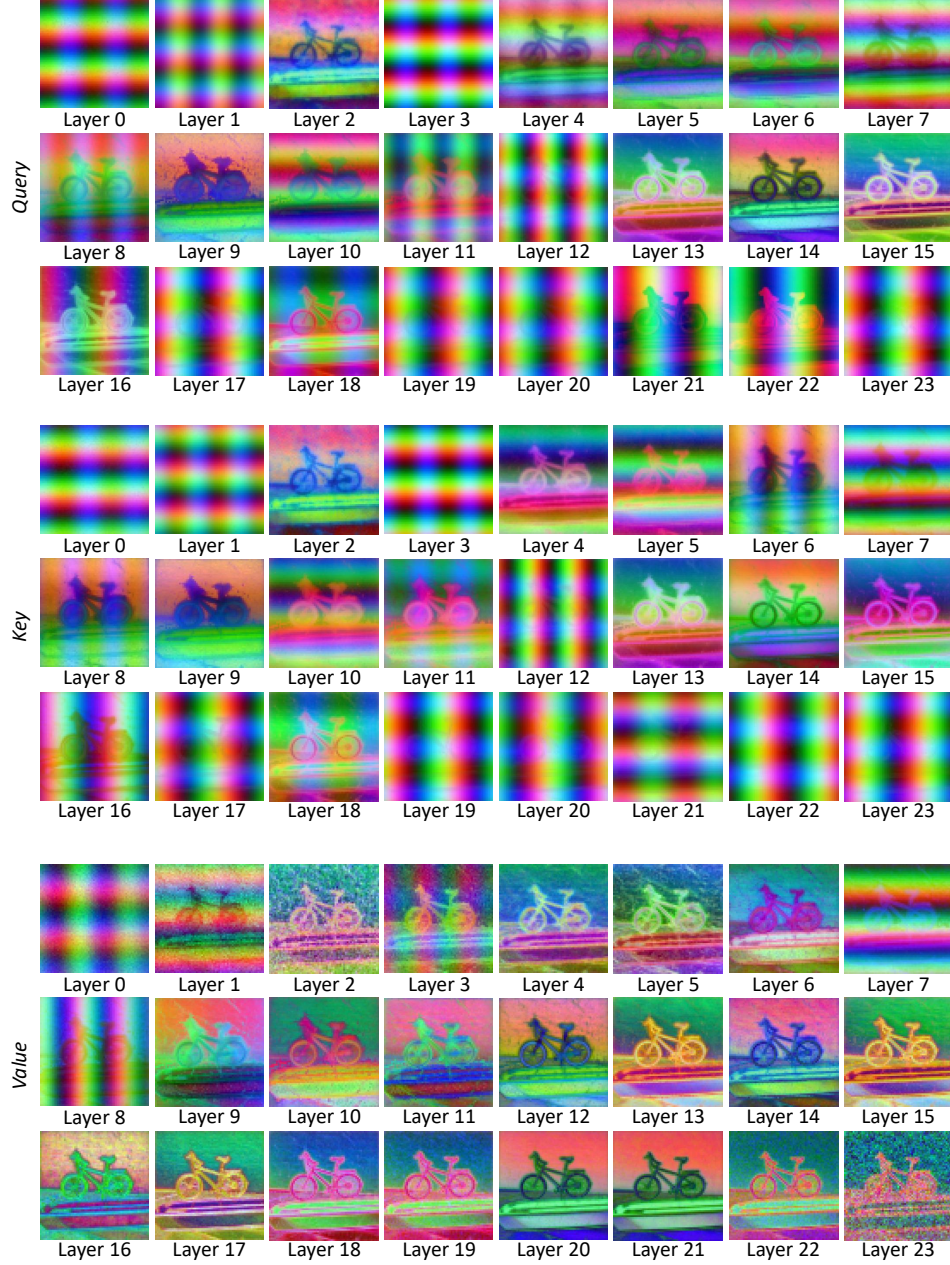


Figure 29: PCA visualization of query-, key-, and value-projected feature.



Figure 30: Emergent behavior of <pad> tokens in unconditional generation.

E Additional qualitative results for generation

We present extended results from attention perturbation experiments in Sec 3.3. Perturbations applied to layers other than layer 9 result in minor degradation of image fidelity or structure while largely preserving semantic content. In contrast, perturbing layer 9 leads to the generation of semantically irrelevant images. Conversely, if we leverage this perturbed sample as a negative sample for the guidance, we observe a substantial gain on the image quality. This strongly supports our claim that layer 9 plays a critical role in cross-modal interaction, with a particular emphasis on aligning with the text modality.

E.1 Extended I2T attention perturbation results

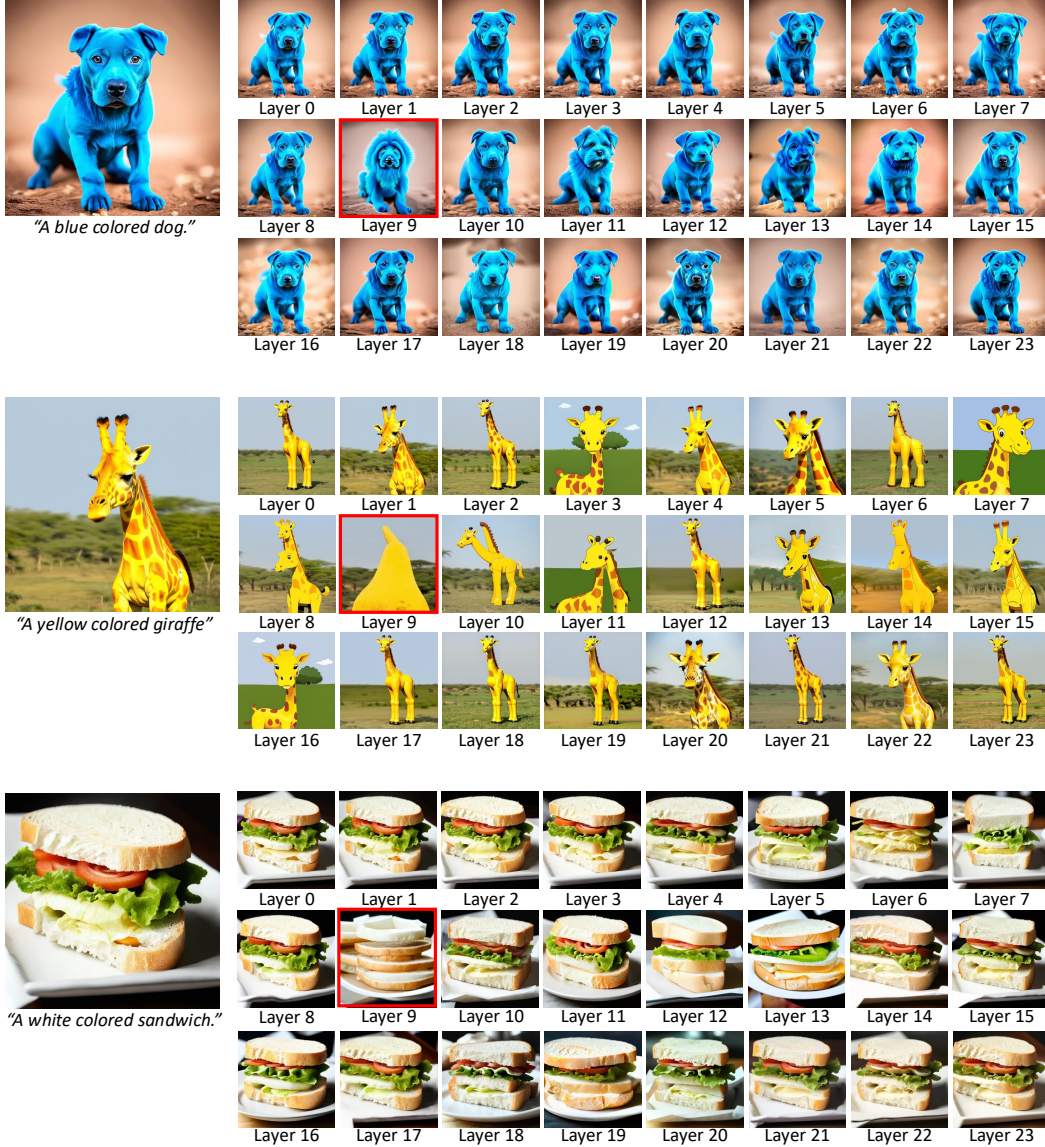


Figure 31: Effect of applying attention perturbation at different layers during image generation.

E.2 Extended I2T attention perturbation guidance results

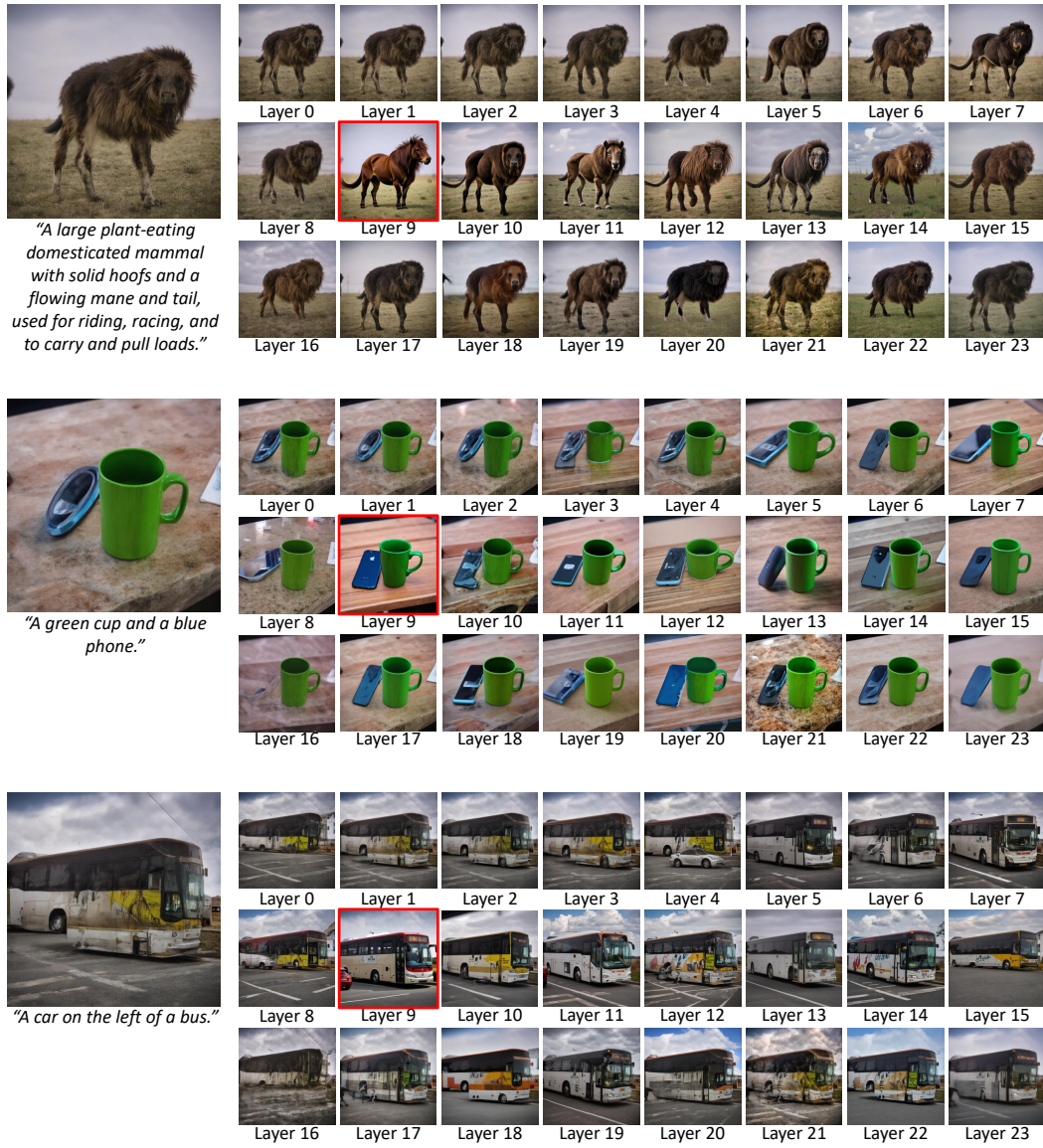


Figure 32: Effect of applying attention perturbation guidance at different layers during image generation.

E.3 Image generation results of our trained models



Figure 33: Qualitative results of trained models.

F Additional qualitative results for segmentation

F.1 Open-vocabulary semantic segmentation results

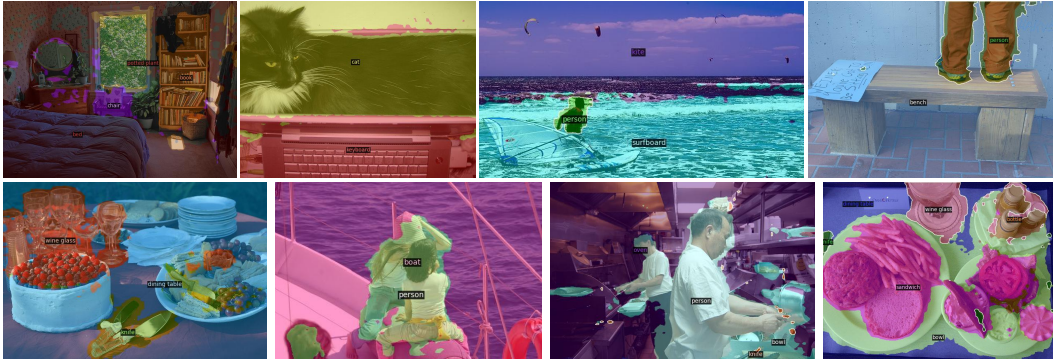
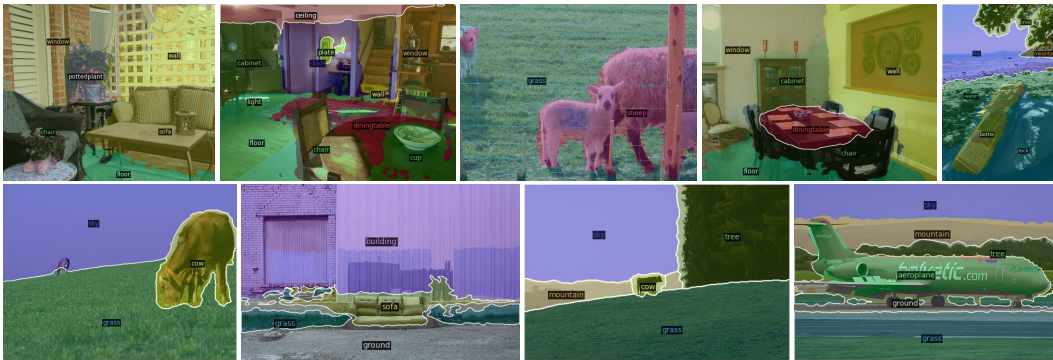
Figure 34: **Open-vocabulary semantic segmentation results of Seg4Diff on COCO-Object.**

Figure 35: Open-vocabulary semantic segmentation results of Seg4Diff on Pascal-Context59.

Figure 36: **Open-vocabulary semantic segmentation results of Seg4Diff on ADE20K.**

F.2 Unsupervised segmentation results

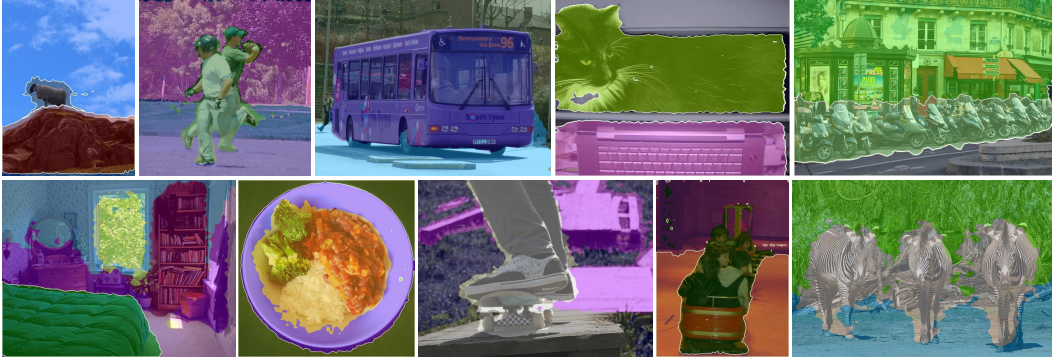


Figure 37: Unsupervised segmentation results of Seg4Diff on COCO.

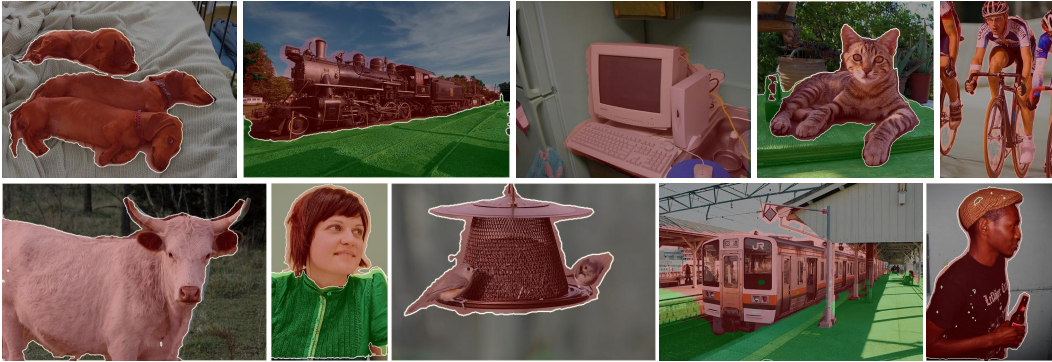


Figure 38: Unsupervised segmentation results of Seg4Diff on VOC.

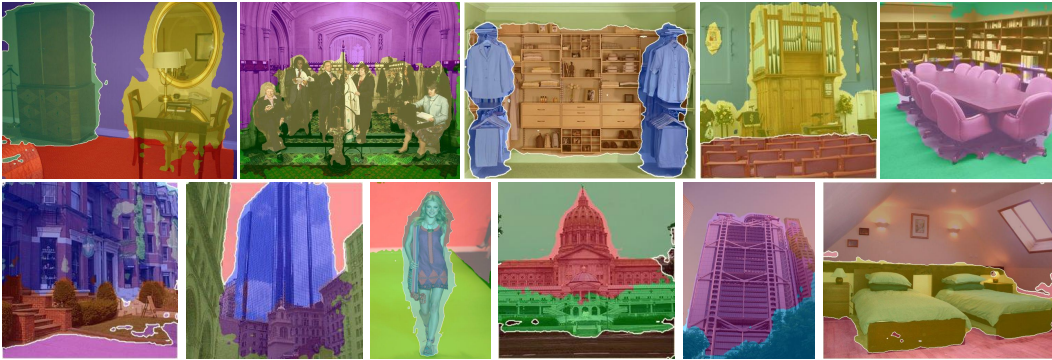


Figure 39: Unsupervised segmentation results of Seg4Diff on ADE20K.

G Limitations of our method

Our method evaluates segmentation without postprocessing or upsampling, which limits performance on extremely small objects. In addition, diffusion models may ground class names differently from the ground-truth annotations, introducing an inherent mismatch that impacts accuracy. Addressing this representation–annotation gap is an important direction for future work.

We focus on dense perception and image generation tasks, deferring reasoning-centric evaluations [65] (e.g., action recognition, spatial-relations QA) to future work, as our supervision targets segmentation semantics rather than high-level reasoning. For simplicity, the loss is applied to a single “sweet-spot” layer per backbone; broader layer/timestep exploration or auxiliary heads (e.g., optical flow, pose) could yield further gains without changing our core findings.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading “NeurIPS Paper Checklist”,**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers.**

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: We state the claims and contributions in abstract and introduction, Section 1.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The limitations are not discussed in the main paper but will be included in the supplementary material.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [No]

Justification: The paper includes mathematical formulations and algorithmic descriptions; however, it does not present any formally stated theorems or lemmas accompanied by a complete set of assumptions and rigorous proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides sufficient details to reproduce the main experimental results, including descriptions of the model architecture, training procedures, datasets used, and evaluation metrics.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have released the code and data to support further reproducibility.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide experimental setting/details in implementation detail section.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We provide statistical value in analysis.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide information on the computer resources in implementation detail section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: This research fully conforms to the NeurIPS Code of Ethics. It does not involve any personally identifiable information, sensitive attributes, or human subjects, and no ethical concerns were identified in the process of conducting this work.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).
- 10. Broader impacts**
- Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
- Answer: [NA]
- Justification: This work does not directly target any specific application domain. As such, the work does not have any immediate or foreseeable impacts.
- Guidelines:
- The answer NA means that there is no societal impact of the work performed.
 - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
 - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
 - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
 - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
 - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).
- 11. Safeguards**
- Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?
- Answer: [NA]
- Justification: This paper poses no such risks for misuse.
- Guidelines:
- The answer NA means that the paper poses no such risks.
 - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
 - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
 - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.
- 12. Licenses for existing assets**
- Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
- Answer: [Yes]
- Justification: The paper uses publicly available datasets and pretrained models, all of which are properly cited with their original sources.
- Guidelines:
- The answer NA means that the paper does not use existing assets.
 - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.