Where and How to Perturb: On the Design of Perturbation Guidance in Diffusion and Flow Models

Donghoon Ahn[•1] Jiwon Kang[•1]
Sanghyun Lee[○1] Minjae Kim[○2] Jaewon Min[1] Wooseok Jang[1]
Sangwu Lee[3] Sayak Paul[4] Seungryong Kim[†1]

[1]KAIST AI [2]Korea University [3]Krea AI [4]Hugging Face

Abstract

Recent guidance methods in diffusion models steer reverse sampling by *perturbing* the model to construct an implicit weak model and guide generation away from it. Among these approaches, attention perturbation has demonstrated strong empirical performance in unconditional scenarios where classifier-free guidance is not applicable. However, existing attention perturbation methods lack principled approaches for determining where perturbations should be applied, particularly in Diffusion Transformer (DiT) architectures where quality-relevant computations are distributed across layers. In this paper, we investigate the granularity of attention perturbations, ranging from the layer level down to individual attention heads, and discover that specific heads govern distinct visual concepts such as structure, style, and texture quality. Building on this insight, we propose “HeadHunter", a systematic framework for iteratively selecting attention heads that align with user-centric objectives, enabling fine-grained control over generation quality and visual attributes. In addition, we introduce SoftPAG, which linearly interpolates each selected head’s attention map toward an identity matrix, providing a continuous knob to tune perturbation strength and suppress artifacts. Our approach not only mitigates the oversmoothing issues of existing layer-level perturbation but also enables targeted manipulation of specific visual styles through compositional head selection. We validate our method on modern large-scale DiT-based text-to-image models including Stable Diffusion 3 and FLUX.1, demonstrating superior performance in both general quality enhancement and style-specific guidance. Our work provides the first head-level analysis of attention perturbation in diffusion models, uncovering interpretable specialization within attention layers and enabling practical design of effective perturbation strategies. Our project page is available at: https://cvlab-kaist.github.io/HeadHunter/. The latest version of this paper is available at https://arxiv.org/abs/2506.10978.

1 Introduction

Diffusion models [19, 54, 53, 56, 58, 11, 45, 42] and flow-matching models [34, 37, 14] have gained great popularity in visual generation tasks, including images [45, 42, 41, 5, 14], videos [21, 18, 3, 63], 3D [43, 32, 60, 50], and 4D content [51, 64, 61]. The key behind the success of diffusion models lies in classifier-free guidance (CFG) [20], which substantially enhances image generation quality during conditional inference. Despite its effectiveness, CFG has two key limitations. First, it applies only to conditional generation, limiting its applicability in unconditional settings such as inverse

•: Equally contributed as first author, ○: Equally contributed as second author, †: Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

problems [25, 8, 55, 9]. Second, CFG often reduces sample diversity and leads to over-saturated or overly simplified outputs [47, 28, 48, 26].

To address the limitations of CFG in unconditional scenarios, alternative guidance strategies [1, 23, 22, 26, 24] have been proposed. These methods steer the denoising trajectory by perturbing the input or the model itself, or by training a weaker model, thereby guiding samples away from low-quality regions and toward the high-quality data manifold. Among these strategies, attention-layer perturbation approaches [1, 22] continue to be a practical and widely explored strategy, as they can generate well-aligned weak models without additional training. Some papers explain these guidance methods from an energy perspective [22] and draw connections to Hopfield networks [10, 44], which empirically work well with specific blocks (medium blocks of U-Net [46] architecture) of attention.

However, there is still limited understanding of where perturbations should be applied. In particular, Diffusion Transformer (DiT) [41] architectures lack localized blocks responsible for global semantics, unlike U-Nets [46], and instead distribute this functionality more evenly across all layers [2]. This structural difference makes it even more critical to carefully select perturbation targets to achieve effective guidance.

To determine suitable perturbation targets, we first consider the underlying structure of DiT. In DiT, multi-head self-attention [59] plays a central role in modeling global dependencies. Each head attends to different aspects of the input, and prior work has shown that heads often specialize in distinct semantic features [59, 40, 12, 16]. This suggests that attention heads, rather than entire layers, may serve as more precise and effective targets for perturbation.

Motivated by this, we explore finer-grained yet semantically meaningful computational units within attention layers: *attention heads*. Interestingly, we observe that individual head-level perturbation guidance often capture interpretable visual concepts and specialize in tasks such as enhancing structural fidelity or injecting stylistic elements. Moreover, these functional roles can be composed by combining multiple heads.

Building on these observations, we (i) analyze key properties of head-level perturbation guidance, including composability and controllability, (ii) propose **HeadHunter**, a systematic framework for retrieving heads that align with arbitrary objectives, and (iii) introduce **SoftPAG**, a variant of PAG [1] that enables fine-grained, continuous control over guidance strength, mitigating over-smoothness and oversimplification caused by overly aggressive perturbations.

We demonstrate that HeadHunter not only outperforms layer-level perturbation guidance in general quality improvement but also enables targeted manipulation of visual styles, as supported by strong qualitative and quantitative results.

In summary, our main contributions are as follows:

- To the best of our knowledge, we are the first to apply perturbations at the level of individual attention heads, enabling fine-grained and concept-specific control.
- We analyze the properties of head-level perturbation guidance in Diffusion Transformers (DiT), providing insights into the specialization and combination of such heads.
- We propose HeadHunter, a systematic head selection framework for arbitrary objectives, and demonstrate its effectiveness in both general quality enhancement and style-specific guidance.
- We introduce SoftPAG, which interpolates each selected head's attention map toward the identity matrix, providing a continuous knob to tune perturbation strength and to mitigate oversaturation and oversimplification.

2 Preliminaries

Diffusion and flow matching models. Diffusion models [19, 52, 57] define a generative process as iterative denoising of a sample drawn from a Gaussian prior. The forward process gradually perturbs a clean data point $\mathbf{x}_0 \sim p_{\text{data}}$ through the noise schedule (α_t, σ_t), as $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $t \in [0, 1]$. A neural network $\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t)$ learns to predict the added noise $\boldsymbol{\epsilon}$, enabling reverse-time sampling from noise to data. On the other hand, flow matching [35, 36] provides a deterministic alternative by learning a continuous-time velocity field that guides a linear interpolation from noise to data: $\mathbf{x}_t = (1 - t)\mathbf{x}_0 + t\,\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and the associated velocity field is $\mathbf{u}(\mathbf{x}_t, t) = \boldsymbol{\epsilon} - \mathbf{x}_0$. A neural network $\hat{\mathbf{u}}_\theta(\mathbf{x}_t, t)$ is trained to approximate this target velocity.

Multi-head attention mechanism. Vaswani et al. [59] introduced the multi-head attention mechanism, which projects queries, keys, and values into multiple subspaces and computes attention in parallel. At the l-th layer of the denoising network $\hat{\boldsymbol{\epsilon}}_\theta$, each head $h \in \{1, \ldots, H_l\}$ applies distinct linear projections:

$$\mathbf{Q}_{l,h} = \mathbf{Q}_l \mathbf{W}^Q_{l,h}, \ \mathbf{K}_{l,h} = \mathbf{K}_l \mathbf{W}^K_{l,h}, \ \mathbf{V}_{l,h} = \mathbf{V}_l \mathbf{W}^V_{l,h}, \tag{1}$$

where $\mathbf{Q}_l, \mathbf{K}_l, \mathbf{V}_l \in \mathbb{R}^{N \times d}$ are the query, key, and value matrices, $\mathbf{W}^{Q,K,V}_{l,h} \in \mathbb{R}^{d \times \bar{d}}$ are the head-specific projections, $\bar{d}$ is the per-head dimensionality, and H_l is the number of attention heads in layer l. The attention map and output for each head are

$$\mathbf{A}_{l,h} = \text{Softmax}\left(\frac{\mathbf{Q}_{l,h}\mathbf{K}^\top_{l,h}}{\sqrt{\bar{d}}}\right), \quad \mathbf{O}_{l,h} = \mathbf{A}_{l,h}\mathbf{V}_{l,h}, \tag{2}$$

and all head outputs are concatenated and projected:

$$\text{MultiHead}(\mathbf{Q}_l, \mathbf{K}_l, \mathbf{V}_l) = \text{Concat}(\mathbf{O}_{l,1}, \ldots, \mathbf{O}_{l,H_l})\mathbf{W}^O_l, \tag{3}$$

where $\mathbf{W}^O_l \in \mathbb{R}^{(H_l \bar{d}) \times d}$ is the output projection matrix. Multi-head attention enables the model to capture diverse semantic relations across tokens [12, 40, 16].

Classifier-free guidance. Classifier-free guidance (CFG) [20] enhances conditional generation by extrapolating predictions from conditional and unconditional models:

$$\hat{\boldsymbol{\epsilon}}_{\text{CFG}} = (1 + w)\hat{\boldsymbol{\epsilon}}_{\text{cond}} - w\hat{\boldsymbol{\epsilon}}_{\text{uncond}}, \tag{4}$$

where w is the guidance scale, and $\hat{\boldsymbol{\epsilon}}_{\text{cond}}$, $\hat{\boldsymbol{\epsilon}}_{\text{uncond}}$ denote the predicted noise with and without conditioning, respectively.

Attention perturbation guidance. As attention maps encode spatial and semantic structures [4, 39], perturbing them can guide the denoising trajectory away from suboptimal predictions, improving structural fidelity at test time [1, 22, 26]. Analogous to classifier-free guidance [20], attention perturbation guidance extrapolates between original and perturbed predictions:

$$\hat{\boldsymbol{\epsilon}}_{\text{guided}} = (1 + w)\hat{\boldsymbol{\epsilon}}_{\text{original}} - w\hat{\boldsymbol{\epsilon}}_{\text{perturbed}}, \tag{5}$$

where $\hat{\boldsymbol{\epsilon}}_{\text{perturbed}}$ is obtained by modifying attention maps $\mathbf{A}_{l,h}$ in Eq. 2 during the forward pass.

Let $\mathcal{L}$ denote the perturbed layers and H_l the number of heads in layer l. Perturbations are applied to $\mathbf{A}_{l,h}$ for all heads $h \in \{1, \ldots, H_l\}$ and $l \in \mathcal{L}$. In perturbed-attention guidance (PAG) [1], the attention maps are replaced with the identity matrix, disabling contextual aggregation:

$$\mathbf{A}^{(\text{PAG})}_{l,h} = \mathbf{I} \in \mathbb{R}^{N \times N}.$$

3 Motivation

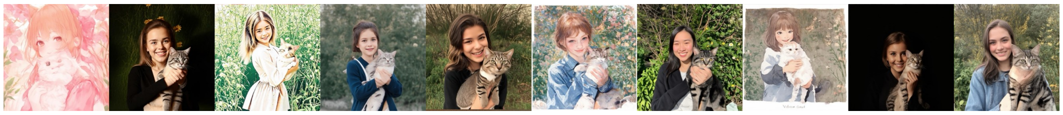


Figure 1: **Motivating example.** Each image is generated with PAG [1], where perturbation is applied to a single attention head within DiTs. Guiding with different perturbed *attention heads* produces notably distinct results. Results for additional heads are provided in Appendix E.1. All images are generated with the prompt *"smiling girl holding a cat, in a flower garden"* using `stable-diffusion-3-medium`. Each row corresponds to a single layer, with different heads perturbed across columns.

Attention perturbation guidance [1, 22, 26] steers generation away from weaker predictions by slightly altering the model's forward pass. Thus, deciding what to weaken becomes critical, but principled methods for determining where in the diffusion network to apply perturbation remain underexplored. As a result, prior works often rely on heuristic selection of perturbation targets, such

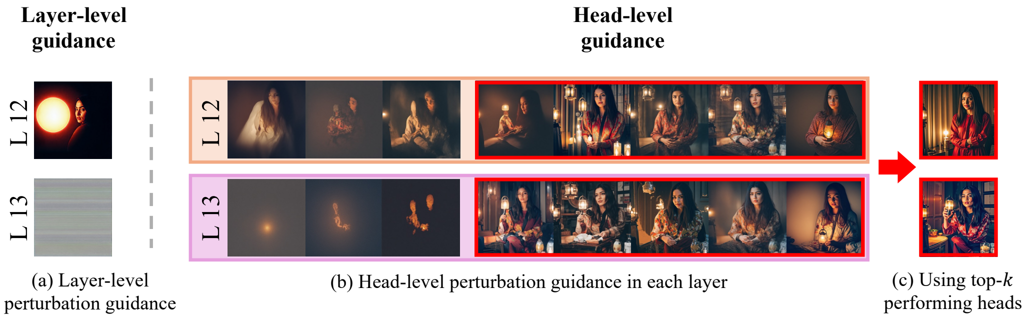

Layer-level guidance
Head-level guidance
L 12
L 13
L 12
L 13
(a) Layer-level perturbation guidance
(b) Head-level perturbation guidance in each layer
(c) Using top-k performing heads

Figure 2: **Generated images from head- and layer-level perturbation guidance.** (a) Results of layer-level perturbation guidance, where perturbation is applied to all heads in the layer. (b) Results of head-level perturbation guidance, where each result is obtained by independently applying perturbation to a single head of the layer. Red boxes indicate high-performing heads in terms of PickScore [27]. (c) Perturbation guidance using only the high-performing heads identified in (b) yields higher-quality generations across both low-performing layers (L13) and well-performing ones (L12). The prompt *"Turkish girl with lantern, dark room"* is used.

as the middle block of a U-Net [46], which is known to process high-level semantic information with global self-attention [23, 1].

However, unlike U-Net, transformer-based architectures as in DiTs [41, 5, 14, 29, 15] lack a coarse-to-fine synthesis structure [2]. This absence of an explicit bottleneck makes it challenging to directly identify layers responsible for high-level semantics. Instead, the model distributes semantic processing more uniformly across layers. In this setting, attention, particularly multi-head self-attention [59], plays a central role in modeling global dependencies without convolutions.

Each attention head processes different aspects of the input, and prior works in large language models and vision transformers have shown that heads specialize in capturing distinct semantic attributes [59, 40, 12, 16]. This suggests that attention heads, rather than entire layers, may serve as more effective units for applying perturbation.

Motivated by this observation, we increase the granularity of attention perturbation from entire layers to individual *attention heads*. In conventional approaches, referred to as **layer-level perturbation guidance** (or simply layer-level guidance), perturbations are applied uniformly to all attention heads within selected layers. In contrast, our proposed **head-level perturbation guidance** (or simply head-level guidance) selectively perturbs specific attention heads, enabling finer and more targeted control. As shown in Fig. 1, perturbing different heads leads to clearly distinct effects in the generated images, indicating that heads act as *semantically meaningful substructures*. This highlights the limitations of layer-level perturbation, which may be overly coarse and suboptimal. Even with carefully chosen layers, layer-level perturbation fails to leverage the modularity and functional diversity inherent in multi-head attention, motivating our focus on head-level guidance.

3.1 Analysis of head-level perturbation guidance

Definition. We define head-level perturbation guidance as applying perturbations selectively to a subset of *attention heads* during the forward pass. Specifically, given a set $\mathcal{S} = \{(l_1, h_1), \dots, (l_m, h_m)\}$ of m selected (layer, head) pairs, we replace their attention maps with identity matrices:

$$\mathbf{A}_{l,h}^{(\text{PAG})} = \mathbf{I} \quad \text{for } (l,h) \in \mathcal{S}, \tag{6}$$

where $\mathbf{A}_{l,h}^{(\text{PAG})}$ denotes the perturbed attention map of head h at layer l, and $\mathbf{I}$ is the identity matrix used in PAG [1]. While we illustrate this with PAG, the formulation also applies to other perturbation methods such as SEG [22], with each head perturbed independently. See Appendix F for details.

Inefficiency of layer-level guidance due to intra-layer diversity. Even when we perturb the attention map of a single layer within diffusion models, it exhibits highly polysemantic attentive behavior, as the computation in individual heads occurs independently and diversely [13, 40]. Therefore, applying the same perturbations across all heads in a layer fails to account for this intra-layer diversity, potentially disrupting the output from the guidance and leading to unintended side effects. As shown

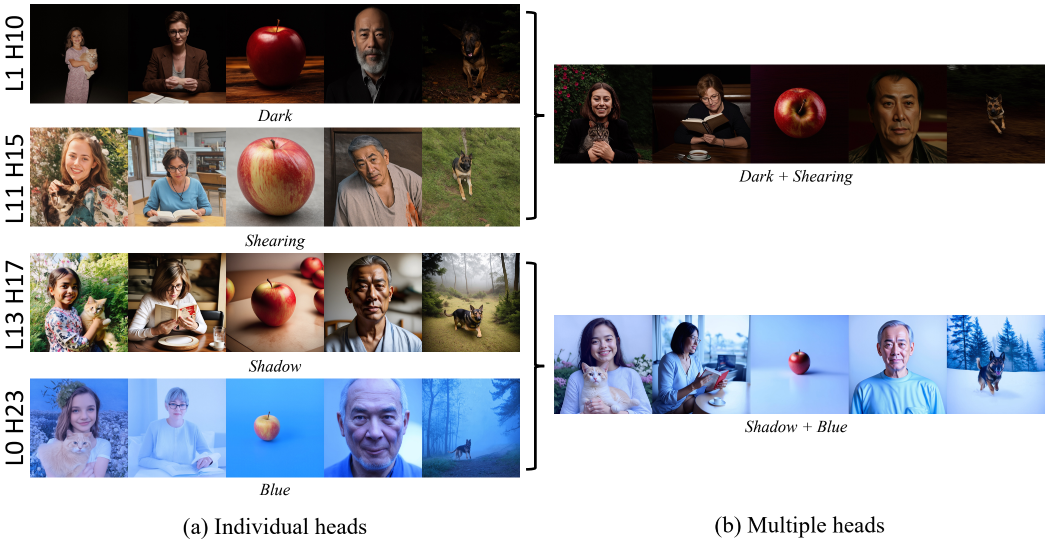


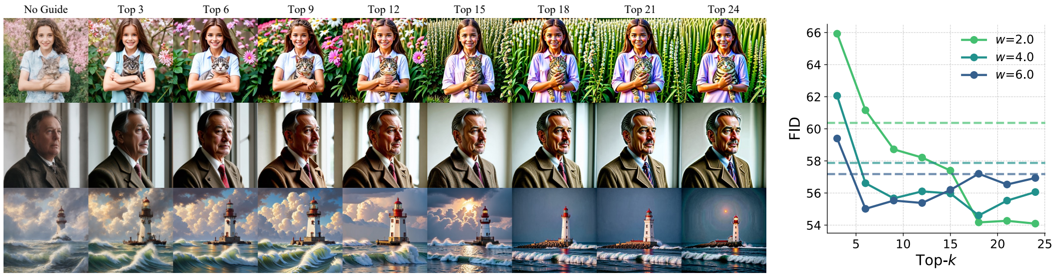
Figure 3: **Effect of head-level guidance on concept amplification and combination.** (a) Guiding with individual heads amplifies specific visual concepts such as darkness, geometry, shadow, or color. (b) Guiding with two heads simultaneously combines their effects in the output.

in Fig. 2, some heads from supposedly poor layers (`L13` of Fig. 2 (a)) still produce high-quality results (Fig. 2 (b)). This suggests that overall quality may degrade due to the effects of a few heads dominating the outcome, indicating further room for improvement. To test this, we filtered and perturbed only high-scoring heads based on preference scores. The resulting samples (Fig. 2 (c)) show clear improvements over standard layer-level guidance. These results demonstrate the inefficiency of coarse layer-level guidance and motivate us to *analyze head-level guidance and its properties*.

Individual head-level guidance occasionally reveals interpretable concepts. As can be seen in Fig. 1, we observe that perturbing individual heads for guidance leads to highly diverse behaviors in terms of their influence on generation. Beyond quality, we find that certain heads amplify specific visual attributes when used for guidance. These include texture, color tone, geometry, and lighting. We provide some examples in Fig. 3. `(L11,H15)` consistently induces shearing, while other heads amplify darkness or blue tone. Additional interpretable heads and their effects are shown in Appendix E.2. We further analyze these effects in depth in the discussion section, where we examine how individual heads contribute to image generation, how their roles differ across layers, and how they interact when composed together.

Concepts can be composed via head combinations. Interestingly, combining multiple heads for head-level guidance results in the composition of their associated visual concepts. As shown in Fig. 3 (b), the generated images exhibit hybrid effects—such as the combination of lighting (`L1, H10`) and shearing distortions oriented from top-left to bottom-right (`L11, H15`)—suggesting that head-level guidance can serve as a mechanism for concept-level control. More examples of concept composition are provided in Appendix E.3.

Composing heads increases image quality but may lead to over-saturation and over-simplification. Perturbing more heads strengthens the guidance effect but can also introduce undesirable artifacts such as oversaturation or oversimplification. As shown in Fig. 4 (a), adding more perturbed heads results in over-perturbation, leading to over-smoothed or overly simplified output. Fig. 4 (b) further illustrates this effect quantitatively: the quality metric (FID [27]) using 1K prompts from the MS COCO dataset begins to degenerate as more heads are added after some threshold. This suggests that, in addition to the choice of perturbation method, the number of perturbed heads itself is a key factor for controlling the overall perturbation strength.



(a) Samples guided by the top-k heads selected by HeadHunter.

(b) Qualitative results.

Figure 4: **Results of HeadHunter for general image quality improvement.** Performance improves as more top-ranked heads are added, demonstrating the effectiveness of compositional head selection via HeadHunter. The dotted horizontal line in (b) indicates the best score achieved by layer-level guidance, which is surpassed by a compact set of top-k heads for $k < 10$. Dashed lines indicate the FID of the best-performing layer-level perturbation for each guidance scale w in Eq. 5.

4 Controlling Head-Level Attention Perturbation

4.1 HeadHunter: An Iterative Framework for Retrieving Effective Attention Heads

The analysis in Sec. 3.1 revealed that individual attention heads exert distinct and often complementary influences via head-level guidance—some modulate lighting, others control color or geometry. Moreover, these effects can be *composed*, meaning that combinations of heads can yield richer and more controllable outputs than any single head alone. This raises a key question: Can we guide image generation toward a *user-defined* objective by selectively perturbing attention heads that are *automatically* identified based on their individual and compositional effects?

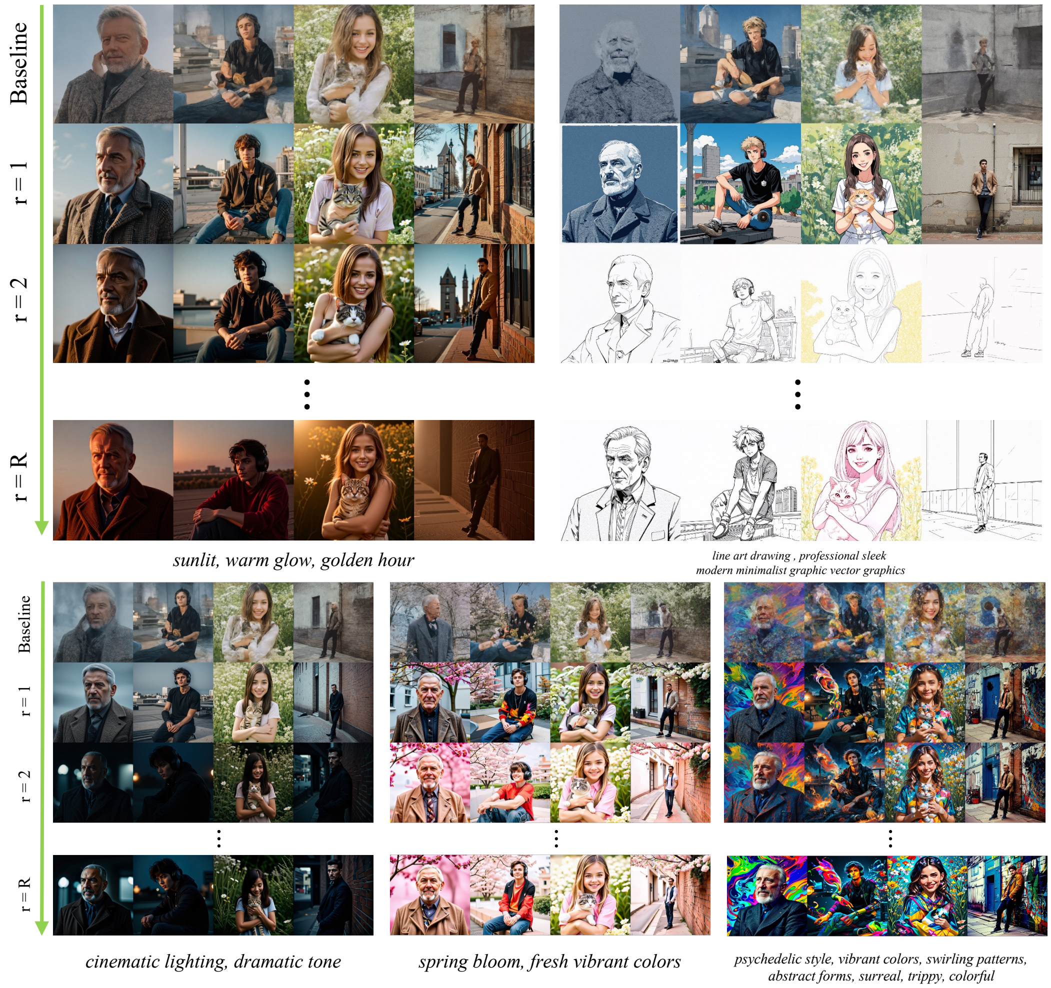
To address this, we introduce **HeadHunter**, a framework that optimizes an arbitrary objective function by iteratively selecting a subset of attention heads to perturb. Each round consists of three stages: generation, evaluation, and expansion. In the generation stage, the framework perturbs attention maps for each candidate head $(l, h) \in \mathcal{S}$ and generates samples using attention perturbation guidance with multiple prompt–seed pairs $\mathcal{Q} = \{(p_1, s_1), \ldots, (p_M, s_M)\}$. During the evaluation stage, it computes the average objective score using the user-given objective function $\mathcal{O}$ over the generated samples for each candidate. In the expansion stage, the top-k performing heads are added to the selected head set $\mathcal{S}_{\text{final}}$. This process is repeated for a fixed number of rounds R. The full procedure is described in Algorithm 1. We validate HeadHunter both quantitatively and qualitatively across tasks such as general quality and style enhancement.

This iterative design greedily selects attention heads in the order that most increases the objective. As a result, it achieves rapid improvements during the early rounds (Fig. 6). Later rounds enable the selection of heads that strengthen specific styles, such as imposing a particular tone on the image (Fig. 8), even if they do not improve overall quality. We present detailed experimental results and analysis in later sections.

4.1.1 Improving general image quality

In Sec. 3.1, we observe that even layers considered ineffective under layer-level guidance can improve quality when selectively activating appropriate heads. This suggests that *fine-grained head-level selection across the entire set of heads* may yield stronger enhancements in image quality. To explore this, we apply HeadHunter to search the head selection space.

Experiments. Various metrics can be used to assess image quality, including vision-language models and learned reward functions. Following recent trends in diffusion models that optimize image preference scores [49, 27, 62, 65], we demonstrate our method using PickScore [27] as the objective $\mathcal{O}$. HeadHunter is applied with 20 prompt-seed pairs $\mathcal{Q}$, one round ($R = 1$), and a selection of $k = 24$ heads per round. Additional implementation details are provided in Appendix D. Qualitative results in Fig. 4 (a) show that image quality progressively improves as more HeadHunter-selected heads are included in the guidance process, supporting the compositional nature of head-level perturbations observed earlier. Fig. 4 (b) presents quantitative results in terms of FID, evaluated on



sunlit, warm glow, golden hour

line art drawing , professional sleek modern minimalist graphic vector graphics

cinematic lighting, dramatic tone

spring bloom, fresh vibrant colors

psychedelic style, vibrant colors, swirling patterns, abstract forms, surreal, trippy, colorful

Figure 5: **Qualitative results of HeadHunter for style-oriented quality improvement.** The first row presents the unguided result. In each subsequent row, we apply head-level guidance using additional attention heads that are incrementally selected by the HeadHunter framework. As more heads are accumulated over rounds, the generated images exhibit progressively stronger alignment with the target style while improving visual quality. Original figures and additional results conducted in FLUX.1-Dev [29] can be found in Appendix D.2.

1K prompts from the MS COCO [33] validation set. The performance improves as top-ranked heads are incrementally added, except in the high guidance scale case ($w = 6.0$), where oversaturation may occur. Remarkably, using only 25% of the heads ($k = 6$ out of 24) achieves comparable to or even better performance than full layer-level perturbation (shown as dotted lines for each guidance scale) and can be boosted much further by using additional heads. These results suggest that HeadHunter can identify a compact yet effective subset of attention heads, outperforming heuristic layer-based selection strategies.

4.1.2 Improving style-oriented quality

While Section 4.1.1 demonstrated HeadHunter's effectiveness in improving overall fidelity of samples, many real-world applications require more targeted control over *specific* visual styles, such as evoking a particular mood, or mimicking classical art techniques. Inspired by the findings of Sec. 3.1 and Fig. 3, which reveal that certain attention heads are responsible for distinct stylistic or geometric attributes, we investigate whether HeadHunter can selectively enhance a target style through head-level guidance—while preserving or even enhancing overall image quality. We refer to this approach as style-oriented quality improvement.

Experiments. To evaluate HeadHunter's ability to enhance specific styles, we prepare two prompt sets: one for *style* (e.g., "warm golden hour glow") and one for *content* (e.g., "portrait of a violinist"). Each trial uses a composite prompt of the form "`style, content`." We run HeadHunter with $R = 5$ and $k = 3$. Fig. 5 shows that, as more heads are selected, the generated images increasingly exhibit the intended style while maintaining structural integrity.

Fig. 6 reports quantitative metrics: PickScore [27] and the LAION Aesthetic Score (AES) [49] both improve steadily as k increases in the prompt-seed pairs $\mathcal{Q}$. These results confirm that HeadHunter effectively amplifies target styles in alignment with human preferences. For additional qualitative examples with SD3 [14] and FLUX.1-Dev [29], we refer readers to Figs. 24 to 29 in Appendix.

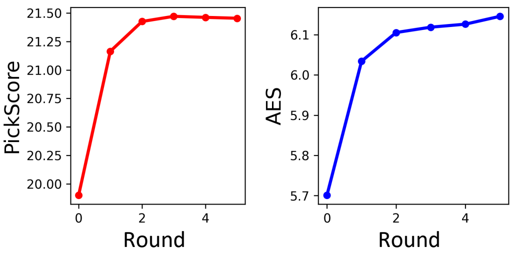


Figure 6: **Quantitative results of HeadHunter for style-oriented quality improvement.** As more heads accumulate, the generated images progressively align better with the target style and exhibit improved visual quality.

Assessing generalizability to unseen content prompts. To evaluate the generalizability of HeadHunter, we compare it against both the baseline and CFG using a set of 50 unseen content prompts. As shown in Tab. 1, HeadHunter achieves significantly higher human preference scores than the baseline and performs comparably to CFG, thereby validating its ability to generalize to novel content prompts. Notably, as illustrated in Fig. 7, **HeadHunter leads to a substantial enhancement of stylistic attributes even compared to CFG**. For the style prompt *sunlit, warm glow, and golden hour*, HeadHunter produces visibly intensified reddish tones and sunlight effects. Likewise, for the *line art drawing ...* style prompt, monotone line characteristics are markedly enhanced. These results suggest that HeadHunter can serve as an effective plug-and-play module within existing inference pipelines (including those employing CFG) to improve both stylistic fidelity and overall image quality without requiring additional training.

4.1.3 Discussion

Surprising utility of individually weak heads. Interestingly, some heads that individually produce poor outputs can still play an essential role in composition. In Fig. 8 (a), we observe a sudden enhancement in stylistic expression (e.g., intensified red) with inclusion of certain head set. By individually performing head-level guidance with each head in the set, we found that some head alone yields low-quality outputs but it reliably enforces a style trait. Such heads are unlikely to be selected through one-shot evaluation due to low quality, but HeadHunter's iterative nature allows them to be integrated later as their utility emerges in composition. This highlights HeadHunter's capacity to exploit *compositional synergies* among heads.

4.2 Controlling Guidance Intensity via Attention Map Interpolation

Although our head-level guidance provides fine-grained, objective-specific control, we propose an orthogonal approach that enables *continuous* modulation of the guidance effect. This allows for more precise adjustment of the perturbation intensity applied to selected heads or layers. Specifically, we introduce a simple interpolation strategy that interpolates the original attention map $\mathbf{A}$ with its perturbed counterpart. For example, the interpolated attention map for perturbed-attention guidance (PAG) [1] is defined as:

$$\mathbf{A}_{l,h}^{(\text{SoftPAG})} = (1 - u)\mathbf{A}_i + u\mathbf{I}, \quad \text{for } (l, h) \in \mathcal{S}, \quad u \in [0, 1]. \tag{7}$$

This formulation replaces $\mathbf{A}_{l,h}^{(\text{PAG})}$ in Eq. 6 with its interpolated version, enabling a smooth transition between the original attention map and the fully perturbed one.

We present qualitative results across varying u values in Fig. 9. These results show that increasing the interpolation parameter u initially enhances sample quality by rectifying structural artifacts in the samples. However, beyond a certain point, it leads to oversimplified images with smoothed backgrounds and exaggerated structures. The interpolation method provides a **"sweet spot"** where the samples exhibit improved structure while retaining visually realistic features. By treating the

Table 1: **Quantitative evaluation of generalizability to unseen content prompts.** Number in the parenthesis denotes guidance scale w. Applying HeadHunter (style-oriented quality setting) to unseen content prompts demonstrates strong generalization, yielding significantly higher human preference scores than the baseline and performance comparable to CFG.

Method	**PickScore ↑**	**AES ↑**	**HPS ↑**	**Imreward ↑**
Baseline	19.66	5.37	0.2147	-0.591
CFG (3.0)	20.87	5.71	0.2924	0.844
CFG (6.0)	20.92	5.80	0.3046	1.063
HeadHunter (3.0)	20.70	5.92	0.2901	0.470
CFG (3.0) + HeadHunter (3.0)	20.92	5.93	0.3036	0.845

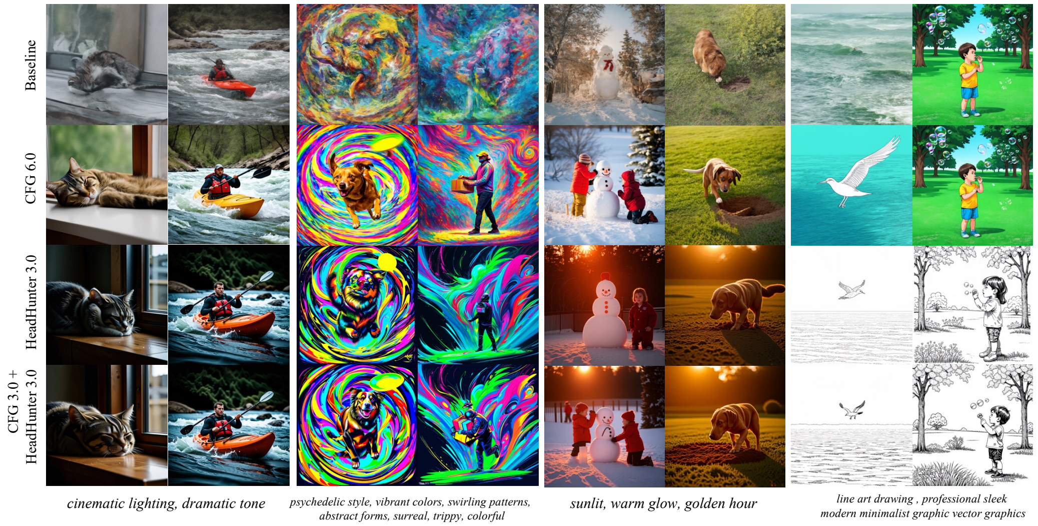
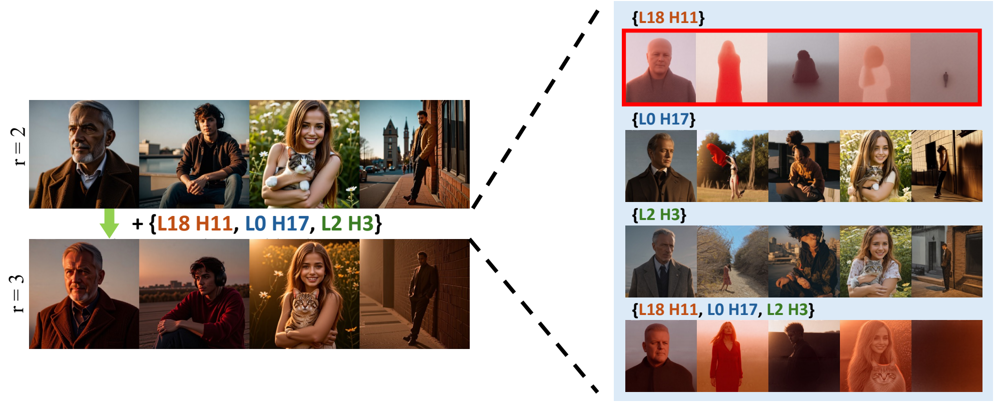

Baseline
CFG 6.0
HeadHunter 3.0
CFG 3.0 +
HeadHunter 3.0
cinematic lighting, dramatic tone
psychedelic style, vibrant colors, swirling patterns, abstract forms, surreal, trippy, colorful
sunlit, warm glow, golden hour
line art drawing , professional sleek modern minimalist graphic vector graphics

Figure 7: **Qualitative evaluation of generalizability to unseen content prompts.** Applying HeadHunter (style-oriented quality setting) results in substantial enhancement of stylistic attributes, outperforming not only the baseline but also CFG.



(a) "Sudden style change" moment and participating heads at that moment

(b) Detailed analysis of how each attention head individually contributes to guidance

Figure 8: **Role of individually weak heads.** For sudden stylistic transition moments (e.g., adding global warmth in (a)), we visualize the effect of newly added heads in (b). One of these heads generate blurry reddish outputs when used alone (see red box). Although they fail to produce meaningful content when used independently, they contribute effectively when composed with previously selected structural heads. This highlights the importance of iterative strategy since those heads are unlikely to be selected by one-shot evaluation due to low quality. Additional results can be found in Fig. 21

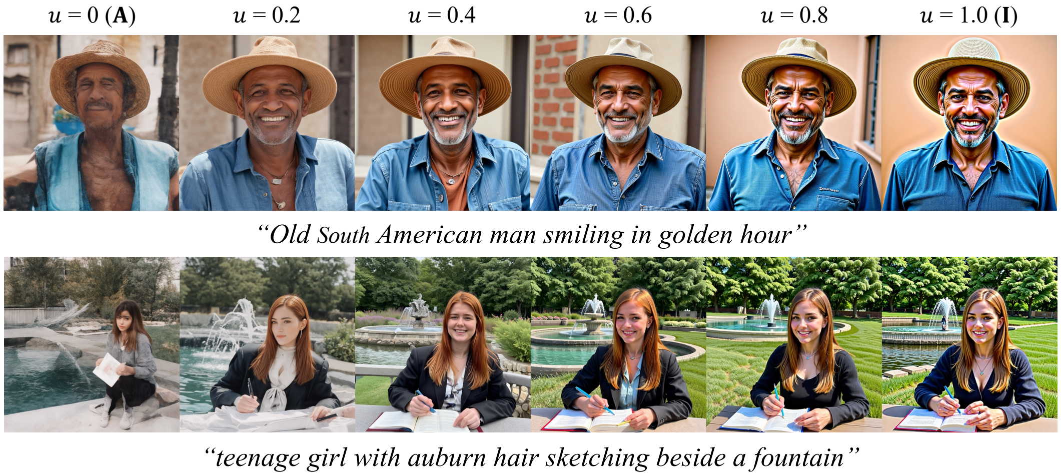

$u = 0$ (**A**)
$u = 0.2$
$u = 0.4$
$u = 0.6$
$u = 0.8$
$u = 1.0$ (**I**)
"Old South American man smiling in golden hour"
"teenage girl with auburn hair sketching beside a fountain"

Figure 9: **Generated Images with linear interpolation between attention map $\mathbf{A}$ and an identity matrix $\mathbf{I}$**. While increasing u enhances quality up to a point, it eventually results in over-saturation and structural over-simplification. Detailed hyperparameters can be found in Appendix G.

identity-matrix perturbation in PAG [1] as a specific point in probability distribution space, our formulation enables a smooth and principled interpolation between $\mathbf{A}$ and $\mathbf{I}$. The scalar $u \in [0, 1]$ provides an intuitive and interpretable control over perturbation strength. Additionally, since this method is a simple linear interpolation of self-attention maps , it can be implemented with a few lines of code. We refer to this controllable variant of PAG as **Soft Perturbed-Attention Guidance (SoftPAG)**.

Interpolation parameter u *vs.* guidance scale w. Both the guidance scale w and interpolation parameter u influence the strength of perturbation, often affecting saturation and structural fidelity. This raises the question of whether adjusting u offers any advantage over simply tuning w.

Fig. 10 in Appendix visualizes metric values over a grid of (w, u) pairs, where brighter regions indicate better performance and red boxes denote optimal configurations. **Across most metrics, the best-performing setting occurs at $u < 1.0$,** indicating that full replacement (i.e., $u = 1$ as in PAG) is often suboptimal. These results underscore the importance of explicitly controlling perturbation strength via interpolation, rather than relying solely on the guidance scale. Due to space limitations, we defer a detailed discussion of the effectiveness of SoftPAG and the selection of u and w to Appendix B.

5 Conclusion

In this work, we move beyond heuristic layer selection in attention perturbation guidance by identifying attention heads as a more meaningful and fine-grained unit of intervention. To the best of our knowledge, this is the first work to perform attention perturbation at the level of individual heads. Specifically, we present HeadHunter, an iterative framework for selecting semantically relevant attention heads aligned with arbitrary, user-defined objectives. Empirical results on state-of-the-art DiT-based models, including Stable Diffusion 3 and FLUX.1, show that head-level perturbation not only improves general image quality with a single search but also aesthetically enhances specific visual styles. In addition, we introduce SoftPAG, a lightweight yet powerful mechanism for continuously modulating perturbation strength via attention map interpolation. Together, these two approaches address key limitations of prior guidance methods, enabling more targeted, effective, and controllable inference-time interventions. We believe this work opens promising directions for interpretable and modular control in generative modeling.

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) and 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).

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] A. Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, and Dominik Lorenz. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, abs/2311.15127, 2023.

[4] 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.

[5] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024.

[6] Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. Pixart-δ: Fast and controllable image generation with latent consistency models. *arXiv preprint arXiv:2401.05252*, 2024.

[7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-$alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

[8] Hyungjin Chung, Jeongsol Kim, Michael T. McCann, M. Klasky, and J. C. Ye. Diffusion posterior sampling for general noisy inverse problems. *ArXiv*, abs/2209.14687, 2022.

[9] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and J. C. Ye. Improving diffusion models for inverse problems using manifold constraints. *Neural Information Processing Systems*, abs/2206.00941, 2022.

[10] Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168:288–299, 2017.

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural information processing systems*, 34:8780–8794, 2021.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[13] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

[14] 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.

[15] fal. Auraflow: A flow-based text-to-image generation model. `https://huggingface.co/fal/AuraFlow`, 2024. Accessed: 2025-05-14.

[16] Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. *arXiv preprint arXiv:2310.05916*, 2023.

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv (Cornell University)*, 2017.

[18] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, abs/2210.02303, 2022.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[21] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

[22] Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. *Advances in Neural Information Processing Systems*, 2024.

[23] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7462–7471, 2023.

[24] Junha Hyung, Kinam Kim, Susung Hong, Min-Jung Kim, and Jaegul Choo. Spatiotemporal skip guidance for enhanced video diffusion sampling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11006–11015, 2025.

[25] Zahra Kadkhodaie and Eero P. Simoncelli. Stochastic solutions for linear inverse problems using the prior implicit in a denoiser. *Neural Information Processing Systems*, 2021.

[26] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 2024.

[27] 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. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023.

[28] T. Kynkäänniemi, M. Aittala, Tero Karras, S. Laine, Timo Aila, and J. Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *Advances in Neural Information Processing Systems*, abs/2404.07724, 2024.

[29] Black Forest Labs. Flux.1 [dev]: A 12b parameter rectified flow transformer for text-to-image generation. `https://huggingface.co/black-forest-labs/FLUX.1-dev`, 2024. Accessed: 2025-05-14.

[30] Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Arsh Koneru, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. Reflect-dit: Inference-time scaling for text-to-image diffusion transformers via in-context reflection. *arXiv preprint arXiv:2503.12271*, 2025.

[31] Tiancheng Li, Weijian Luo, Zhiyang Chen, Liyuan Ma, and Guo-Jun Qi. Self-guidance: Boosting flow and diffusion generation on their own. *arXiv preprint arXiv:2412.05827*, 2024.

[32] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 300–309, 2023.

[33] 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.

[34] Y. Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *International Conference on Learning Representations*, abs/2210.02747, 2022.

[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] Nanye Ma, Larry B. Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. *European Conference on Computer Vision*, 2024.

[38] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025.

[39] Jisu Nam, Heesu Kim, DongJae Lee, Siyoon Jin, Seungryong Kim, and Seunggyu Chang. Dreammatcher: appearance matching self-attention for semantically-consistent text-to-image personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8100–8110, 2024.

[40] Jungwon Park, Jungmin Ko, Dongnam Byun, Jangwon Suh, and Wonjong Rhee. Cross-attention head position patterns can align with human visual concepts in text-to-image generative models. *arXiv preprint arXiv:2412.02237*, 2024.

[41] 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.

[42] 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 (Cornell University)*, 2023.

[43] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[44] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

[45] 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.

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[47] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradely, Otmar Hilliges, and Romann M. Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. *International Conference on Learning Representations*, 2024.

[48] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. S. Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, abs/2205.11487, 2022.

[49] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

[50] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.

[51] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, A. Vedaldi, Devi Parikh, Justin Johnson, and Yaniv Taigman. Text-to-4d dynamic scene generation. In *International Conference on Machine Learning*, 2023.

[52] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.

[53] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv (Cornell University)*, 2015.

[54] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ArXiv*, abs/2010.02502, 2020.

[55] Jiaming Song, Arash Vahdat, M. Mardani, and J. Kautz. Pseudoinverse-guided diffusion models for inverse problems. *International Conference on Learning Representations*, 2023.

[56] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Neural Information Processing Systems*, pages 11895–11907, 2019.

[57] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[58] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv (Cornell University)*, 2020.

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[60] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36:8406–8441, 2023.

[61] Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T. Barron, and Aleksander Holynski. Cat4d: Create anything in 4d with multi-view video diffusion models. *arXiv preprint arXiv:2411.18613*, 2025.

[62] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

[63] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. *International Conference on Learning Representations*, 2025.

[64] Heng Yu, Chaoyang Wang, Peiye Zhuang, Willi Menapace, Aliaksandr Siarohin, Junli Cao, László A. Jeni, S. Tulyakov, and Hsin-Ying Lee. 4real: Towards photorealistic 4d scene generation via video diffusion models. *Advances in Neural Information Processing Systems*, abs/2406.07472, 2024.

[65] Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8018–8027, 2024.

NeurIPS Paper Checklist

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 substantiate our claims in Sections 4 with results included in the section.

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss limitations in the separate section of Appendix.

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: [NA]

   Justification: Our paper does not include theoretical results.

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: We provide all the experimental details of our work. Our work relies on open models and public codebases. Since we don't conduct any training, our results are not tied to any specific datasets.

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 will make our code publicly available along with instructions.

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: Experimental settings, including network architecture, guidance scale, and sampling step are detailed in Section G.

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: [No]

   Justification: Although we conducted experiments using multiple evaluation metrics, we did not report statistical significance tests or error bars.

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 include this information in the Appendix.

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: The research involves algorithmic development and evaluation on standard optimization benchmarks. It does not involve human subjects or obviously ethically sensitive applications, and we assume it conforms to the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper focuses on technical contributions and does not include a specific discussion of broader positive or negative social impacts.

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: Our work is analysis-driven. We leverage open models and open codebases to guide and craft our experiments.

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: We make sure to credit all the sources of the assets we used in our work.

13. **New assets**

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

    Answer: [NA]

    Justification: Our paper do not provide new assets.

    Guidelines:

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]

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: The research does not involve human subjects, therefore IRB approval is not applicable.

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: We used LLM only for writing, editing and formatting purposes.

Appendix

This appendix provides supplementary material to support the main paper.

Contents

A Related Works

Diffusion models. Diffusion models [19, 54, 53, 56, 58, 11, 45, 42] have become the foundation of modern generative modeling, driving advances in both image [45, 42, 41, 5, 14] and video synthesis [21, 18, 3, 63]. Early methods rely on stochastic differential equations (SDEs) to learn a reverse denoising process from noise to data. More recently, deterministic samplers based on ordinary differential equations (ODEs), including rectified flow [37, 14] and flow matching [34], have emerged as efficient alternatives, learning continuous trajectories that transform noise into data. These methods accelerate convergence and improve stability, especially in large-scale models. In parallel, network architectures have shifted from U-Net backbones [42] to Diffusion Transformers (DiT) [41, 5, 14, 7, 6], enhancing scalability and representational capacity.

Attention perturbation guidance. In recent studies, various modifications and extensions of classifier-free guidance have been introduced. Among them, methods that directly perturb the attention layers for guidance work effectively. Self-Attention Guidance (SAG) [23] applies Gaussian blur to the self-attention layers of the UNet to introduce perturbations. Similarly, Perturbed-Attention Guidance (PAG) [1] generates perturbations by replacing the attention maps with an identity matrix. Spatiotemporal Skip Guidance (STG) [24] is a 3D extension of PAG that improves sample quality by selectively skipping spatiotemporal layers. Smoothed Energy Guidance (SEG) [22] defines the energy of self-attention and employs 2D Gaussian blur to reduce the curvature of the energy landscape, using the result as a form of perturbation. Autoguidance [26] enhances image quality by using a bad version of the same conditional model as a source of perturbation. Self-Guidance (SG) [31] leverages the model prediction at a noisy timestep to generate perturbations.

B Continuous control on perturbation strength via attention map interpolation

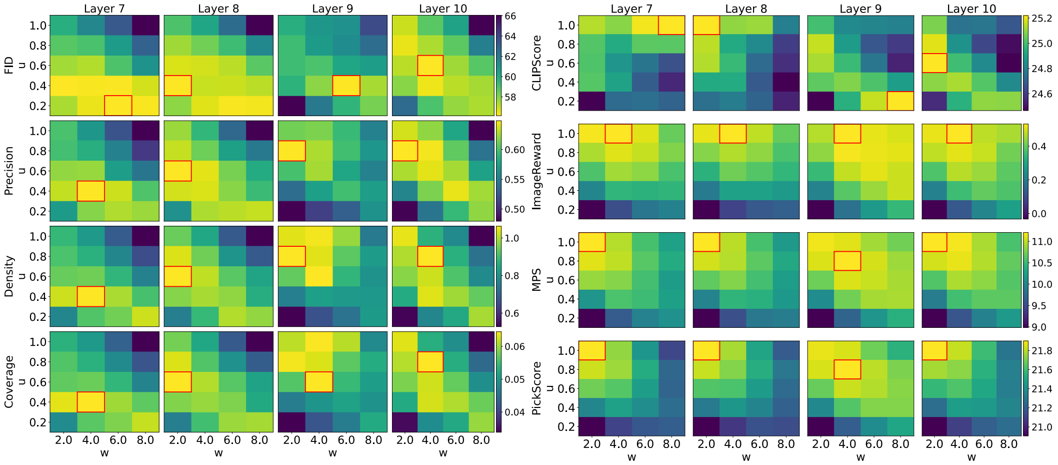


Figure 10: **Grid search for SoftPAG, on guidance scale w and interpolation parameter u with different metrics.**

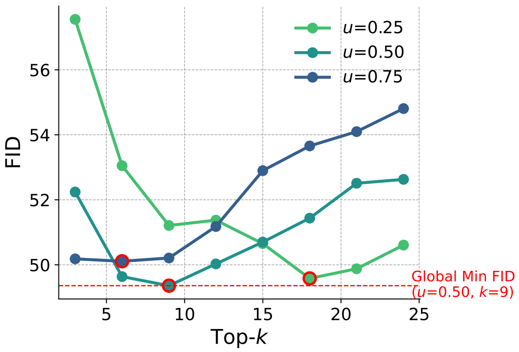


Figure 11: **Effect of interpolation parameter u on FID across PickScore top-k head selections.** As u decreases, the perturbation becomes milder, shifting the optimal point (lowest FID) toward larger k (i.e., more heads are needed). Notably, $u = 0.5$ yields the best FID (indicated by the red dashed line), highlighting the benefit of moderate perturbation strength.

Pre-selecting interpolation parameter u in HeadHunter search. As running HeadHunter multiple times with different interpolation parameters u is computationally expensive, it is generally sufficient to search for heads using the original PAG. Nonetheless, we can explore the effect of u through a single-round search to study general quality trends. To this end, we run HeadHunter with the same setup as in Sec. 4.1.1, replacing PAG with SoftPAG (Eq. 7) and varying u. The results, shown in Fig. 11, report FID [17] scores as a function of k for different u values.

We observe that lower u values (i.e., softer perturbations) require more heads to achieve optimal performance, but also tend to yield better overall results with a moderate setting. Empirically, we find that $u = 0.5$ and $k = 9$ offer a strong trade-off.

This supports the view that both u and k modulate the strength of guidance. While this paper introduces two axes of fine-grained controllability—head-level selection and attention map interpolation—tuning u is significantly more efficient than rerunning head selection. In practice, post-selecting u after a standard head search is typically sufficient. However, if more compute is available, one may rerun HeadHunter with a moderate u (e.g., $u = 0.5$) for a refined selection.

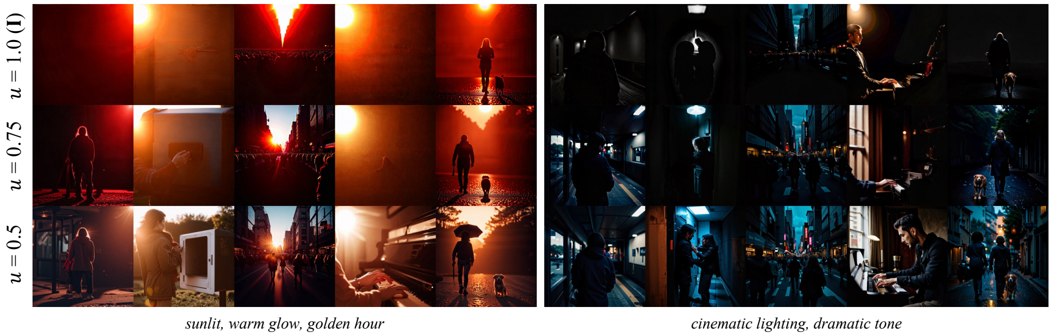

$u = 1.0$ (I)
$u = 0.75$
$u = 0.5$
sunlit, warm glow, golden hour
cinematic lighting, dramatic tone

Figure 12: **Incorporating SoftPAG with the final heads retrieved by HeadHunter for style-oriented quality improvement.** Setting the interpolation parameter $u = 1.0$ makes the effect of head-level guidance more visible, which helps in selecting effective heads via HeadHunter. However, it may also introduce unwanted artifacts such as oversaturation or oversimplification. Post-tuning u after head selection effectively reduces these artifacts while preserving the style enhancement, with minimal additional cost.

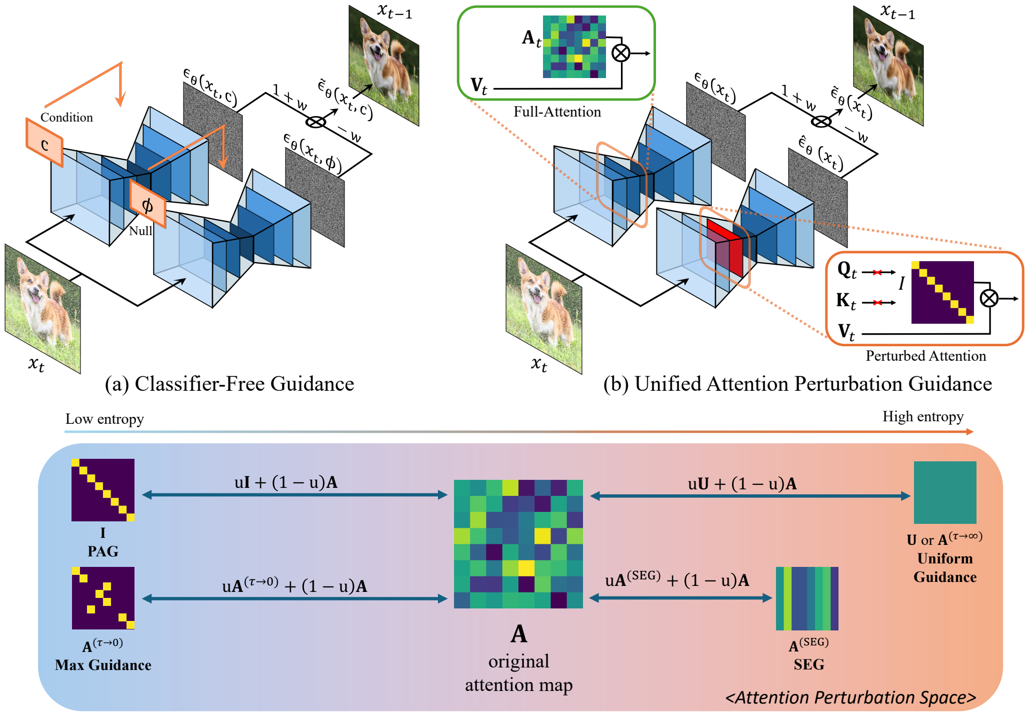
Post-selecting of u for artifact removal with retrieved heads. As discussed in Sec. 3.1, the number of heads used in head-level guidance affects how strong the perturbation is. When more heads are used, the intended effect such as style enhancement can be amplified. However, this may also lead to unwanted side effects including oversaturation or oversimplification due to bias of reward models [27, 49]. In the HeadHunter framework, heads are added step by step. This helps reinforce the target objective, but may also increase the risk of artifacts. To better control the strength of guidance, we can use SoftPAG, which allows us to mitigate the artifacts through the interpolation parameter u.

As tuning both the number of heads and the interpolation parameter u at the same time can be complex and computationally expensive, we adopt a simple two-stage approach. First, we use HeadHunter with $u = 1.0$ to select effective heads. A larger u helps make each head's effect more visible, which is helpful during selection. Then, in the second stage, we optionally reduce u to fine-tune the strength of the guidance and reduce any unwanted artifacts.

This approach offers a practical trade-off by eliminating complex tuning while retaining control over the final output. In Fig. 12, we show examples where retrieved heads with $u = 1.0$ introduce unnatural artifacts. Reducing the SoftPAG parameter to $u = 0.5$ effectively removes these artifacts while preserving the style enhancement provided by the selected heads.

C Other perturbation methods in our framework and comparison

In the main paper, we focus on a single representative perturbation method: identity matrix replacement (PAG). Within our framework, which interprets the attention map as a probability distribution, identity perturbation can be seen as one end of the entropy-modulating transform that minimally shifts the distribution's entropy. Fig. 13 provides a conceptual overview of this unified view, with other perturbations. Our key insight is twofold: **(1) attention perturbations can be understood as transformations over probability distributions**, and **(2) the strength of any perturbation can be smoothly controlled via interpolation with the original attention map**. This enables elegant integration of existing methods while offering controllability.


Condition
c
ϕ
Null
$\epsilon_\theta(x_t, c)$
$1+w$
$\tilde{\epsilon}_\theta(x_t, c)$
$-w$
$\epsilon_\theta(x_t, \phi)$
x_t
x_{t-1}
$\mathbf{A}_t$
$\mathbf{V}_t$
Full-Attention
$\epsilon_\theta(x_t)$
$1+w$
$\tilde{\epsilon}_\theta(x_t)$
$-w$
$\hat{\epsilon}_\theta(x_t)$
x_t
x_{t-1}
$\mathbf{Q}_t$
$\mathbf{K}_t$
$\mathbf{V}_t$
I
Perturbed Attention
Low entropy
High entropy
$\mathbf{uI} + (1 - \mathbf{u})\mathbf{A}$
I
PAG
$\mathbf{uU} + (1 - \mathbf{u})\mathbf{A}$
U or $\mathbf{A}^{(\tau\to\infty)}$
Uniform Guidance
$\mathbf{uA}^{(\tau\to 0)} + (1 - \mathbf{u})\mathbf{A}$
$\mathbf{A}^{(\tau\to 0)}$
Max Guidance
A
original attention map
$\mathbf{uA}^{(\mathrm{SEG})} + (1 - \mathbf{u})\mathbf{A}$
$\mathbf{A}^{(\mathrm{SEG})}$
SEG
<Attention Perturbation Space>

(a) Classifier-Free Guidance

(b) Unified Attention Perturbation Guidance

(c) Perturbation methods

Figure 13: **Unified attention perturbation guidance.** Our framework unifies a variety of perturbation strategies by interpreting attention maps as probability distributions and interpolating between the original and perturbed variants.

C.1 Uniform Guidance

From an information-theoretic perspective, the identity matrix attention map is a special case where each row of the attention map corresponds to a Dirac delta distribution, which is the lowest possible entropy case. At the opposite extreme of the entropy spectrum lies the uniform distribution $\mathbf{U}$, whose rows are all $[\frac{1}{N}, \ldots, \frac{1}{N}]$; it attains the *maximum* entropy, allowing each query to attend equally to every key. Replacing $\mathbf{A}$ by $\mathbf{U}$ in selected heads gives **Uniform Guidance (UG)**. Similar to the interpolation with the identity matrix in SoftPAG, we can define a distributional perturbation between the original attention matrix $\mathbf{A}$ and $\mathbf{U}$ via linear interpolation:

$$\mathbf{A}^{(\mathrm{SoftUG})} = (1 - u)\mathbf{A} + u\mathbf{U}, \quad u \in [0, 1]. \tag{8}$$

We show the interpolation results in Fig. 14. Interestingly, this perturbation strategy tends to preserve the original structure of objects in the unguided sample ($u = 0$) slightly better. However, it often over-restores artifacts present in the original structure, sometimes generating unintended elements (e.g., a framed portrait in the background of Row 1, or a swimming pool in Row 3). Both methods,

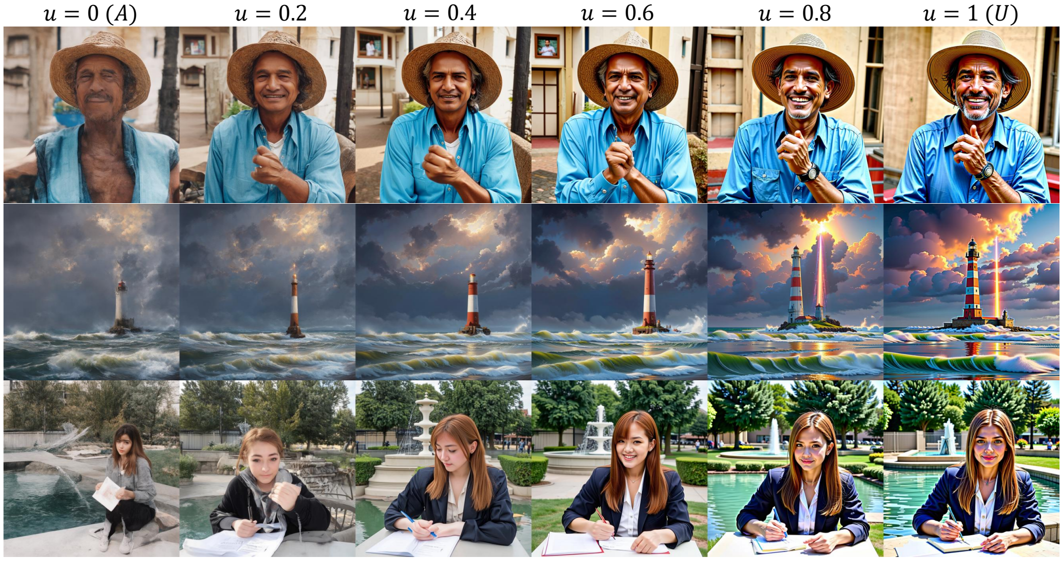


Figure 14: **Generated Images with linear interpolation between attention map $\mathbf{A}$ and a uniform matrix $\mathbf{U}$**. While increasing u enhances quality up to a point, it eventually results in over-saturation and structural over-simplification. Detailed hyperparameters can be found in Appendix G.

however, may exhibit oversaturation or oversimplification when overly strong perturbations are applied, highlighting the importance of a balanced interpolation with the original attention map $\mathbf{A}$.

C.2 Smoothed Energy Guidance

Smoothed Energy Guidance (SEG) [22] applies a 2D Gaussian blur along the query axis to smooth the attention logits. By adjusting the kernel width σ, users can naturally control the strength of the perturbation. In the extreme case where $\sigma \to \infty$, SEG reduces to averaging all query features and applying the result globally.

SEG is not directly applicable to MMDiT. However, SEG is not directly compatible with MMDiT [14], where both image and text tokens participate in attention. Applying a 2D Gaussian blur over the query axis becomes nontrivial when attention spans across modalities, as in recent text-to-video models where multiple frames attend jointly. This issue becomes more pronounced with the rise of multimodal attention (MM-attention).

Restoring controllability to SEG using our framework. To address this, we propose a simple yet widely applicable solution: interpolate between the original attention map $\mathbf{A}$ and the one computed from the mean query feature—corresponding to the limiting case of SEG with $\sigma \to \infty$.

Specifically, we interpolate between the original attention map $\mathbf{A}$ and the SEG attention map $\mathbf{A}^{(\text{SEG})}$, computed using a mean query vector, as follows:

$$\mathbf{A}^{(\text{SoftSEG})} = (1-u)\mathbf{A} + u\mathbf{A}^{(\text{SEG})}, \quad u \in [0,1], \tag{9}$$

where

$$\mathbf{A}^{(\text{SEG})} = \text{Softmax}\left(\frac{\bar{\mathbf{Q}}\mathbf{K}^\top}{\sqrt{d}}\right), \quad \bar{\mathbf{Q}} := \mathbf{1}_N\bar{\mathbf{q}}^\top \in \mathbb{R}^{N\times d}, \quad \bar{\mathbf{q}} := \frac{1}{N}\sum_{i=1}^{N}\mathbf{Q}_i \in \mathbb{R}^d. \tag{10}$$

Here, $\mathbf{Q} \in \mathbb{R}^{N\times d}$ is the query matrix, and $\bar{\mathbf{Q}} \in \mathbb{R}^{N\times d}$ replicates the mean query vector across all N tokens. This allows for a smooth transition from the token-specific attention map to a globally-averaged variant, effectively reintroducing controllability.

This provides a controllable and modality-agnostic extension of SEG to architectures like MMDiT, where traditional Gaussian blurring becomes infeasible. We show the interpolation results in Fig. 15.

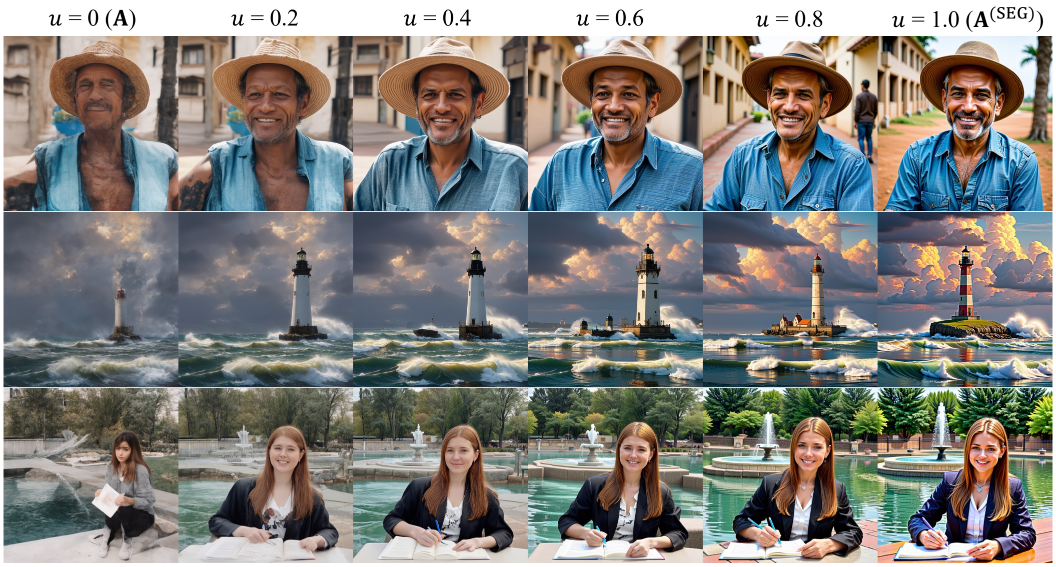


Figure 15: **Generated images with linear interpolation between attention map A and smoothed attention map $\mathbf{A}^{(\text{SEG})}$.**

C.3 Softmax temperature scaling

Attention perturbation using softmax temperature scaling. One intuitive way to increase or decrease entropy is by introducing a temperature parameter τ into the Softmax operation of attention, as follows:

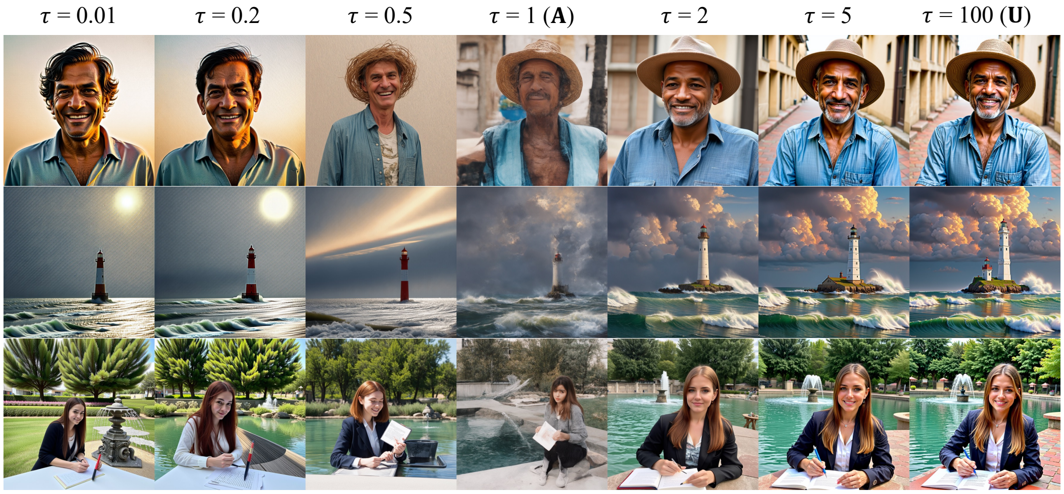


Figure 16: **Generated images with temperature scaling perturbation.**

$$\mathbf{A}^{(\text{temp})} = \text{softmax}\left(\frac{\log \mathbf{A}}{\tau}\right), \quad \tau > 0. \tag{11}$$

Lower temperatures ($\tau \to 0$) sharpen the distributions, driving them toward the Dirac delta distributions, while higher temperatures ($\tau \to \infty$) flatten the distribution towards uniform. This formulation provides a unified mechanism to traverse both entropy-increasing and entropy-decreasing directions within the attention perturbation space.

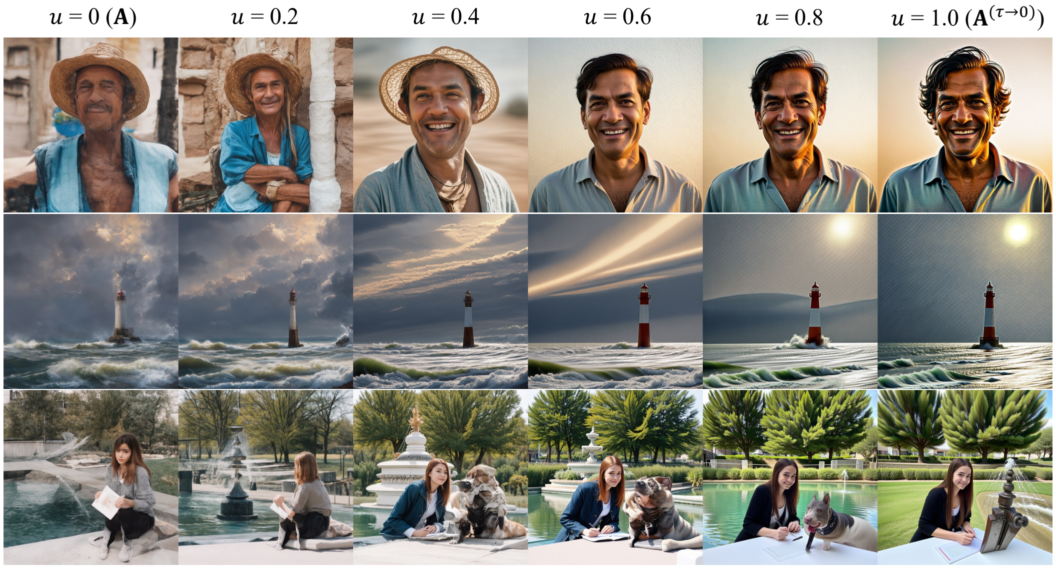


Figure 17: **Generated images with linear interpolation between attention map $\mathbf{A}$ and smoothed attention map $\mathbf{A}^{(\tau\to 0)}$.**

Perturbation $\tau \to 0$ results in different perturbation with PAG (I). As $\tau \to \infty$, the distributions approach uniform distributions, following an exponential-geodesic path. In contrast, as $\tau \to 0$, the distribution converges to a Dirac delta distribution. However, unlike the identity matrix **I**, where all probability mass is placed at each query's own index, this low-temperature limit concentrates mass on the maximum entry of each row, potentially differing from the diagonal. We illustrate the effect of varying τ in Fig. 16. Notably, increasing and decreasing τ leads to meaningful quality improvements, suggesting that modulation of entropy in either direction can enhance generation.

Intuitive control parameter using interpolation framework. Softmax temperature scaling provides a simple way to modulate perturbation strength by adjusting the temperature $\tau \in (0, \infty)$. However, it introduces a practically unintuitive control parameter τ, and the limiting distribution as $\tau \to 0$ is theoretically unreachable in practice due to numerical instability.

Our interpolation framework offers a compelling alternative: it allows for control using a normalized and intuitive parameter $u \in [0, 1]$, while enabling us to directly realize the theoretical limit of the attention map as $\tau \to 0$, denoted as $\mathbf{A}^{\tau\to 0}$, without the numerical issues associated with temperature scaling.

Concretely, in the limit $\tau \to 0$, the softmax attention converges to a one-hot distribution where, for each row, the position of the maximum logit becomes 1 and all others become 0:

$$\mathbf{A}_{ij}^{(\tau\to 0)} := \begin{cases} 1 & \text{if } j = \arg\max_k \mathbf{A}_{ik} \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Then the perturbation methods is as follows:

$$\mathbf{A}^{(\text{SoftMG})} = (1-u)\mathbf{A} + u\mathbf{A}^{(\tau\to 0)}, \quad u \in [0, 1]. \tag{13}$$

As shown in Fig. 17, our framework can replicate this limiting behavior while preserving controllability through a continuous interpolation parameter. Compared to softmax temperature scaling (Fig. 16), this approach offers a more intuitive and stable mechanism for navigating the space of attention perturbations. We refer to this perturbation-based guidance as **Max Guidance (MG)**.

Quantitative comparisons. We discuss various perturbation strategies and present their qualitative results in Fig. 9, 14, 16, 15, 17. To complement these findings, we provide quantitative comparisons in Tab. 2.

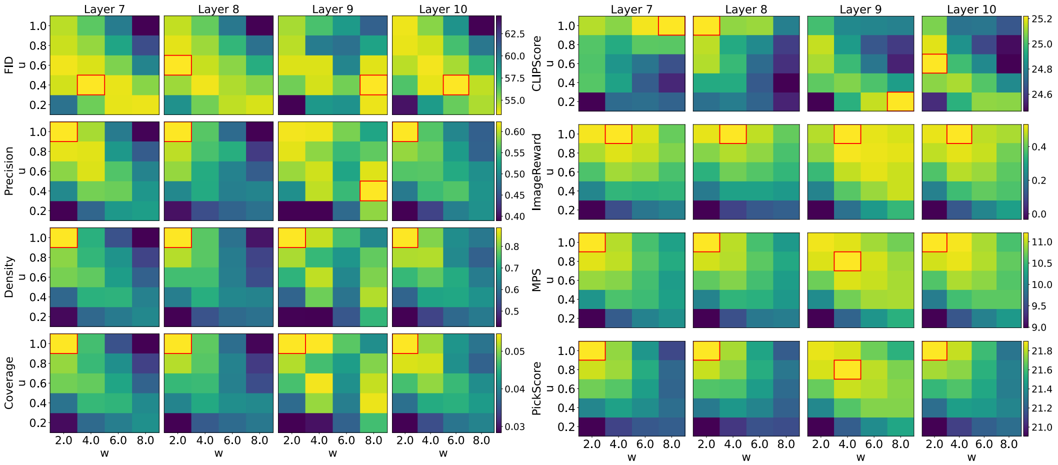


Figure 18: **Grid search for UG, on guidance scale w and interpolation parameter u with different metrics.**

Note that each perturbation method may benefit from different optimal hyperparameter configurations. To ensure a fair evaluation, we define a parameter pool and report the best-performing setting for each method and each metric. Specifically, we consider perturbation layers in the early-to-mid range, namely `layer 7`, `layer 8`, `layer 9`, and `layer 10`; guidance scales $w \in \{2.0, 4.0, 6.0, 8.0\}$; and interpolation parameters $u \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$.

As shown in the Tab. 2, the **I** perturbation consistently yields strong performance across most settings. Based on this robustness, we adopt it as our primary analysis tool in the main paper (Sec. 3, 4).

Method	FID ↓	Precision ↑	Recall ↑	Density ↑	Coverage ↑	PickScore ↑	ImageReward ↑	MPS ↑
Uniform Matrix **U**	53.37	0.62	**0.55**	0.89	0.05	21.90	0.53	11.21
Query Mean $\mathbf{A}^{(\mathrm{SEG})}$	**53.33**	0.62	0.53	0.91	**0.06**	21.89	0.50	11.13
Identity Matrix **I**	56.15	**0.65**	0.48	**1.06**	**0.06**	**22.19**	**0.65**	**11.70**

Table 2: **Comparison of perturbation strategies across different metrics.** Each method shows strengths in different metrics. Overall, Identity matrix perturbation (SoftPAG) achieves strong results across most quality, diversityand human preference metrics. Notably, identity matrix perturbation outperforms others in preference-based scores such as PickScore, indicating better alignment with human perception and more visually pleasing results. Note that FID does not always correlate with human judgment.

More quantitative results. We measured the quality of images generated by layers 7, 8, 9, and 10, which produce the highest-quality outputs. Fig. 19 depicts the results of SoftPAG, which interpolates between the identity matrix and the attention maps, while Fig. 20 depicts the results of UG, which interpolates between the uniform matrix and the attention maps.

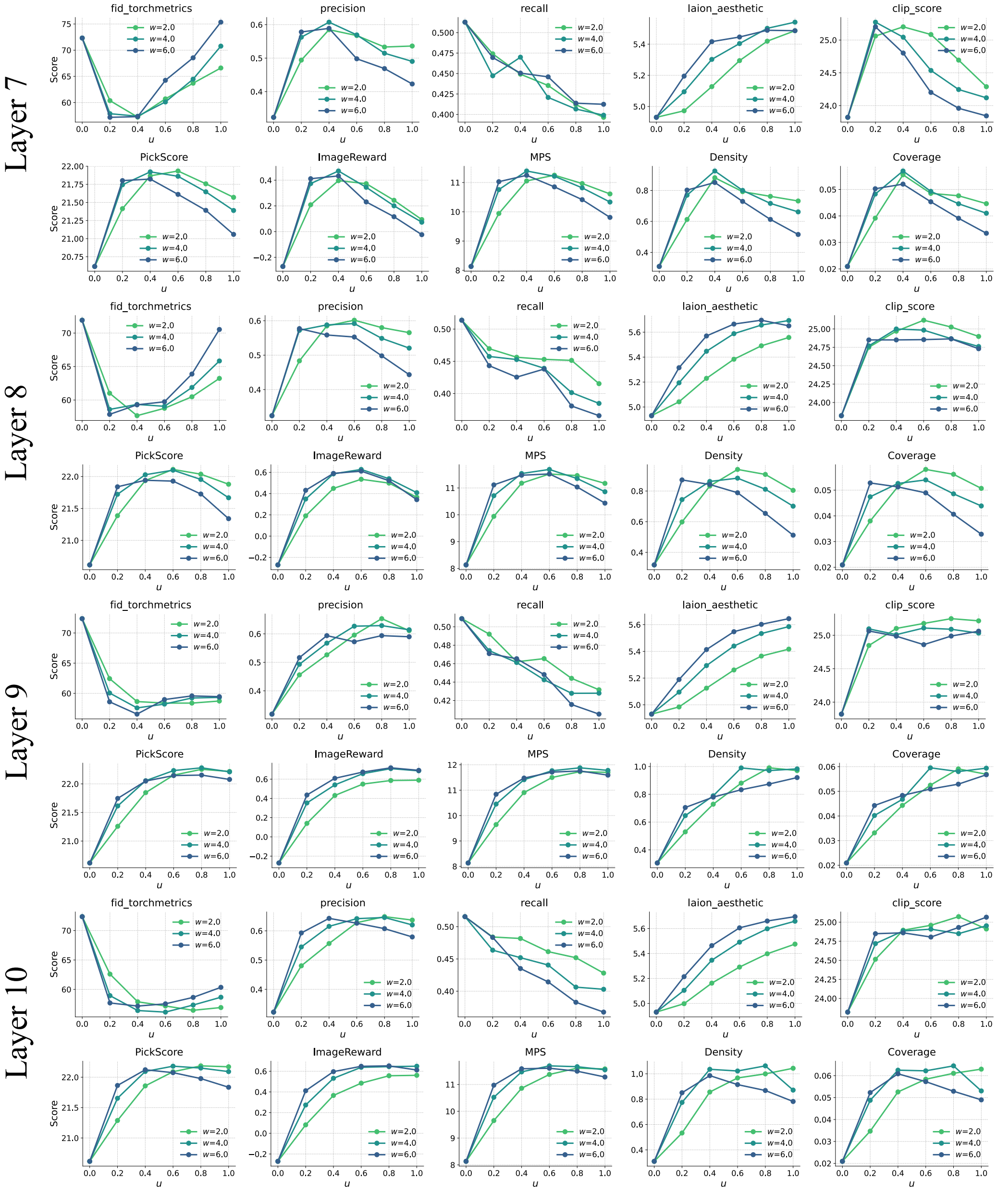


Figure 19: **Quantitative results of interpolation in SoftPAG, on layer** $7, 8, 9, 10$ **with different metrics.**

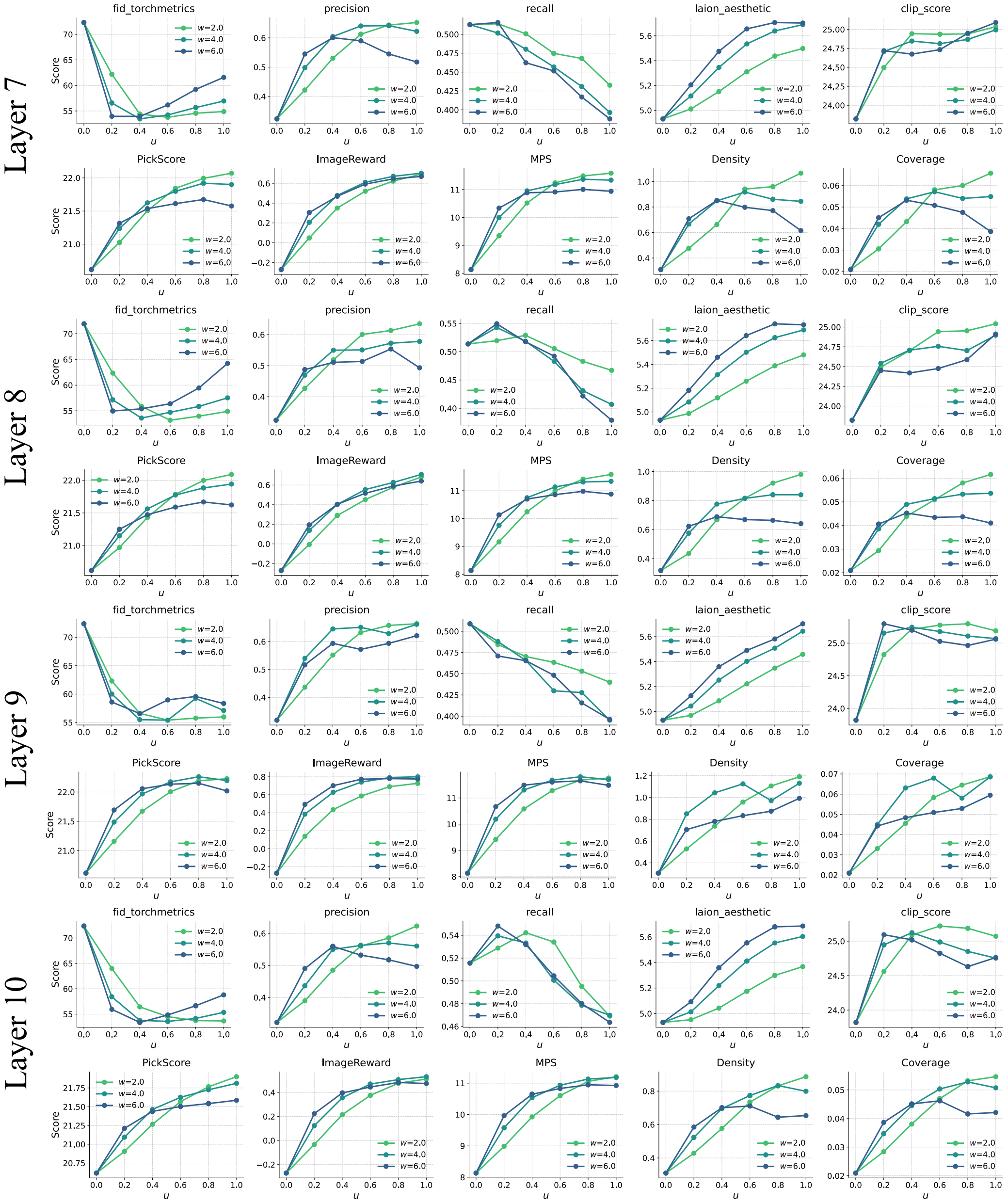


Figure 20: **Quantitative results of interpolation in UG, on layer $7, 8, 9, 10$ with different metrics.**

D HeadHunter: Implementation and additional analysis

D.1 Implementation details

Algorithm 1 HeadHunter: Iterative Objective-Aware Head Selection

```
Require: Diffusion model $\mathcal{M}$, objective function $\mathcal{O}$, prompt-seed pairs $\mathcal{Q} = \{(p_1, s_1), \ldots, (p_M, s_M)\}$, attention head set $\mathcal{S} = \{(l_1, h_1), \ldots, (l_L, h_{H_L})\}$, number of heads per round $k$, number of rounds $R$
Ensure: Selected heads $\mathcal{S}_{\text{final}}$
$\mathcal{S}_{\text{final}} \leftarrow \emptyset$                                   ▷ Initialize selected head set
$\mathcal{R} \leftarrow \mathcal{S}$                                              ▷ Initialize remaining head pool
for $t = 1$ to $R$ do
    $\mathcal{C} \leftarrow [\,]$                                           ▷ Initialize temporary score list
    for each $(l, h) \in \mathcal{R}$ do
        # Generation
        $\mathcal{S}_{\text{target}} \leftarrow \mathcal{S}_{\text{final}} \cup \{(l, h)\}$                 ▷ Define perturbation set
        Generate $\{\hat{x}_j\}_{j=1}^{M}$ using $\mathcal{M}$ with perturbation guidance on $\mathcal{S}_{\text{target}}$ and $(p_j, s_j) \in \mathcal{Q}$
        # Evaluation
        $s_{(l,h)} \leftarrow \frac{1}{M} \sum_{j=1}^{M} \mathcal{O}(\hat{x}_j, p_j)$              ▷ Compute average objective score
        Append $((l, h), s_{(l,h)})$ to $\mathcal{C}$
    end for
    # Expansion
    Sort $\mathcal{C}$ by $s_{(l,h)}$ in descending order
    $\mathcal{S}_{\text{new}} \leftarrow$ top-$k$ heads from $\mathcal{C}$
    $\mathcal{S}_{\text{final}} \leftarrow \mathcal{S}_{\text{final}} \cup \mathcal{S}_{\text{new}}$
    $\mathcal{R} \leftarrow \mathcal{R} \setminus \mathcal{S}_{\text{new}}$
end for
return $\mathcal{S}_{\text{final}}$
```

Improving general image quality. To run HeadHunter, we used 20 prompts ($M = 20$), with $k = 24$ and $R = 1$. All prompts and the corresponding seeds are listed in Tab. 3. Since SD3 contains 24 layers with 24 attention heads per layer, we consider the full set of $N = 24 \times 24 = 576$ attention heads. For head-level guidance, we set the guidance scale to 3.0 and the number of inference steps to 20. Additionally, we set the interpolation parameter $u = 0.25$ in Eq. 7 to avoid excessive over-smoothing effects.

While it is possible to set $R > 1$ to run HeadHunter iteratively, we found that a single round suffices in practice. This is based on the compositional nature of head-level guidance. Heads that individually improve quality tend to combine well, yielding consistent improvements in overall generation quality. As shown in Fig. 4, even a compact set of heads selected in a single round can outperform the best-performing layer-level guidance.

Improving style-oriented quality. For SD3, we evaluated a total of 55 style prompts for both quantitative and qualitative analysis. Since SD3 consists of 24 layers with 24 attention heads per layer, we considered all $N = 24 \times 24 = 576$ attention heads. For head-level guidance, we set the guidance scale to 3.0 and used 20 inference steps.

For FLUX.1-Dev, we used a total of 23 style prompts. The model comprises 57 layers with 24 attention heads per layer, resulting in $N = 57 \times 24 = 1368$ attention heads. Head-level guidance was applied with a guidance scale of 8.0 and 15 inference steps.

In both cases, we used 5 content prompts ($M = 5$) per style. The full list of content prompts is provided in Tab. 4.

D.2 Additional results and analysis

D.2.1 Additional results

We present additional qualitative results for style-oriented quality improvement using HeadHunter in Figs. 24 to 26. Each row corresponds to the result obtained by perturbing the head set selected at round r of Alg.1. The top row shows unguided generation, and the lower rows show results with an increasing number of selected heads applied. As more heads are progressively added, the generated images align better with the target style while maintaining visual plausibility.

For example, for the style prompt *"sunlit, warm glow, golden hour"*, we observe an increasing glow effect, culminating in a global golden-hour tone from the fourth row onward. We provide a more detailed analysis of the contributing heads in Sec. 4.1.3 and Fig. 8, 21.

Interestingly, the framework is effective not only for photorealistic styles such as *"cinematic lighting"*, but also for abstract or simplified styles like *"line art"* or *"flat paper cut"*. Since this style adaptation is achieved without modifying any model parameters, it enables a reusable and lightweight preference tuning.

HeadHunter can benefit from with model size. Since larger models have more expressive attention heads, they can be more effectively leveraged in head-level perturbation guidance. In Fig. 24-29, we provide qualitative results of style-oriented quality improvement conducted on FLUX.1-dev [29].

D.2.2 Additional analysis

Distribution of selected heads across layers. We investigates the layer distribution of selected attention heads within a model's architecture, specifically examining whether impactful heads are localized to particular layers and if their distribution varies by objective. For the general quality setting, we select the top-15 heads per prompt and aggregate them, excluding duplicates (Fig. 22 (a)). For style-oriented quality setting, we repeat the process per style (Fig. 22 (b)).

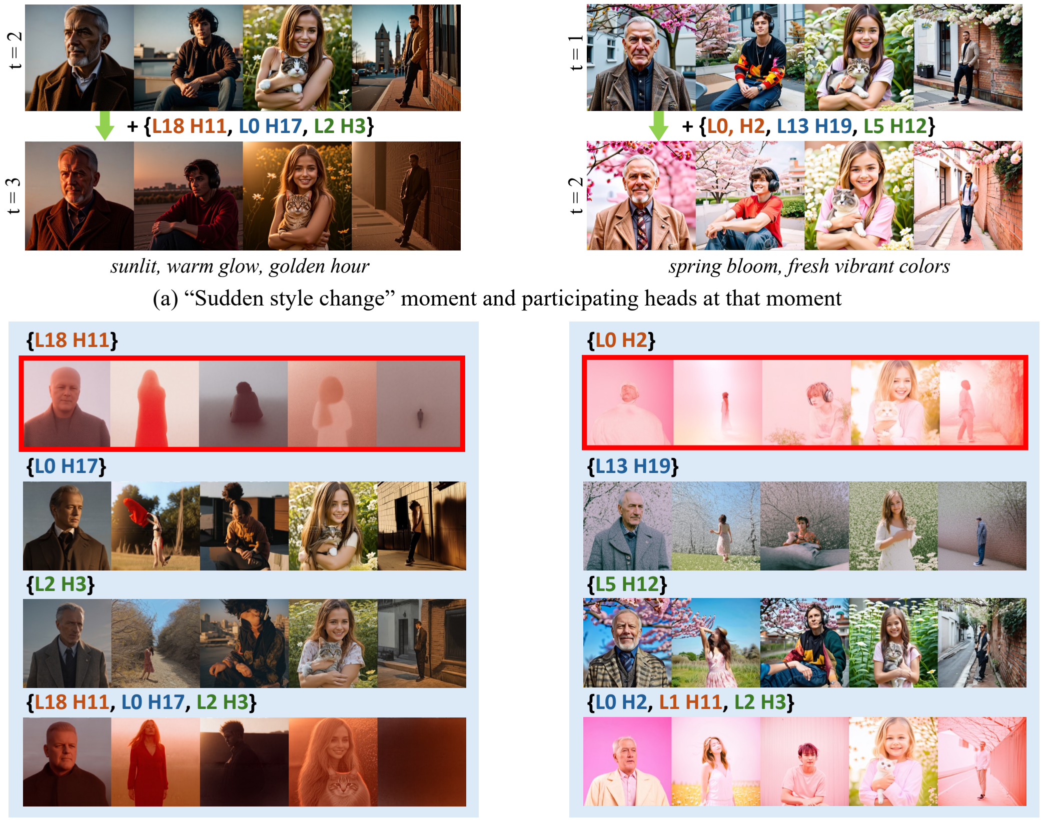
Our findings reveal that effective heads are not concentrated in any single layer. Even the most frequently selected layer accounts for less than 12% of all chosen heads across both objectives (Fig. 22 (a), (b)). This dispersed distribution suggests that a layer-wise guidance approach, which uniformly perturbs all heads within a layer, may be suboptimal. Such an approach risks overlooking critical heads in other layers while simultaneously applying modifications to irrelevant heads.

Furthermore, we observed distinct distribution patterns dependent on the objective. Heads identified for improving general image quality predominantly reside in early to mid-layers, with a mere 2.9% located in layers $\geq$16. In contrast, heads contributing to style-oriented quality improvement exhibit a significantly broader distribution, with 24.8% found in deeper layers ($\geq$16).

We further examine the evolution of head distribution during HeadHunter's iterative refinement for style-oriented quality setting in Fig. 22 (c)). Initially, selected heads are concentrated in mid-layers, mirroring the pattern observed for general quality setting. However, as rounds progress, the distribution become notably flatter, indicating a broader engagement of heads across layers for stylistic refinement. This quantitative shift aligns with qualitative observations.As shown in Fig. 24 (spring bloom, fresh vibrant colors), Fig. 25 (cinematic lighting, dramatic tone), and Fig. 26 (line art drawing ...), the second row (Round 1) shows improved image quality but little stylistic traits. However, from round 2 onward, the style becomes much more pronounced..

In conclusion, our research indicates that (1) effective heads are not localized to specific layers, thereby challenging the efficacy of layer-level guidance, and (2) head distributions are objective-dependent, underscoring the necessity of objective-specific selection. These findings collectively emphasize the critical role of targeted head-retrieval and guidance frameworks, such as HeadHunter, in fully harnessing the diverse functionalities of attention mechanisms.

Does the style enhancement occur even without style prompts? In our style-oriented quality improvement experiments, we search for heads that consistently enhance a given style by using composite prompts of the form "`style, content`". Then a natural question arises: *Do these heads still express their stylistic traits when the style is not explicitly specified in the prompt or even under unconditional generation?*


t = 2
+ {L18 H11, L0 H17, L2 H3}
t = 3
sunlit, warm glow, golden hour
t = 1
+ {L0, H2, L13 H19, L5 H12}
t = 2
spring bloom, fresh vibrant colors
{L18 H11}
{L0 H17}
{L2 H3}
{L18 H11, L0 H17, L2 H3}
{L0 H2}
{L13 H19}
{L5 H12}
{L0 H2, L1 H11, L2 H3}

(a) "Sudden style change" moment and participating heads at that moment

(b) Detailed analysis of how each attention head individually contributes to guidance

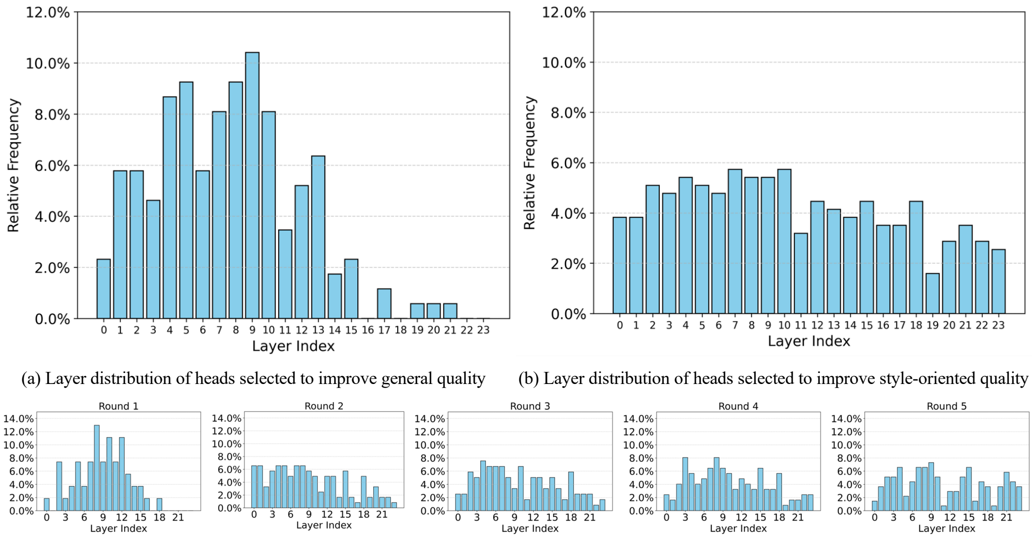
Figure 21: **Role of individually weak heads.** For sudden stylistic transition moments (e.g., adding global warmth or pink hues as shown in (a)), we visualize the effect of newly added heads in (b). One of these heads generate blurry reddish or pinkish outputs when used alone (see red box). Although they fail to produce meaningful content when used independently, they contribute effectively when composed with previously selected structural heads. This highlights the importance of iterative strategy since those heads are unlikely to be selected by one-shot evaluation due to low quality.

To explore this, we apply the heads selected by HeadHunter in an unconditional setting (i.e., with a null prompt, ""). As shown in Fig. 23 (a), we find that even with a null prompt, using these heads can still induce visual traits that align with the original style to some extent. This suggests that certain heads may encode intrinsic stylistic properties that influence generation regardless of prompt or initial noise.

Yet, this effect is not guaranteed across all styles. As can be seen in Fig. 23 (b), in some cases, while style traits are strongly expressed with style prompts, they become weak or ambiguous when only content prompts are used or under unconditional generation. This may be due to polysemanticity in head behavior. This emerging style consistency when using style prompts indirectly supports the validity of including style prompts during HeadHunter's search process in style-oriented quality setting.

Computational efficiency. HeadHunter requires a one-time search process to identify effective attention heads for a given objective. If we denote the number of prompt–seed pairs as M, total attention heads as N, number of rounds as R, and the costs for generation and evaluation as C_{gen} and C_{eval}, the overall cost is approximately $MN(C_{\text{gen}} + C_{\text{eval}})R$. In our experiments using `stable-diffusion-3-medium` with 20-step Euler sampling and PickScore [27] as the objective, this process takes roughly 3 hours on 8×NVIDIA H100 GPUs.

However, this cost is amortized over repeated use. Once a set of heads is selected for a given model and configuration, it can be reused across different prompts and latent seeds—unlike test-time scaling approaches [38, 30] that require per-sample optimization. In the previous section, we further



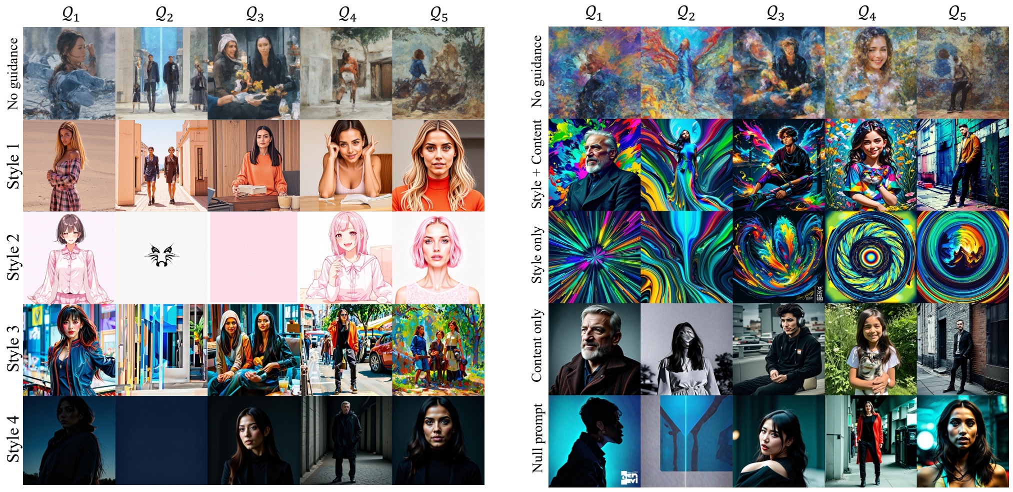
(a) Layer distribution of heads selected to improve general quality

(b) Layer distribution of heads selected to improve style-oriented quality

(c) Layer distribution of heads selected for each round, to improve style-oriented quality

Figure 22: **Layer distribution of heads selected by HeadHunter for different objectives.** Effective heads are not confined to specific layers and the distribution is different with each objective, highlighting the limitations of layer-level guidance and the necessity of objective-specific head selection. (a) In general image quality setting, selections concentrate in early–mid layers (2.9% in layers ≥ 16). (b) In style-oriented quality setting, selections spread broadly, with 24.8% in layers ≥ 16. (c) Round-wise distributions of (b) begin with an early-layer bias similar to (a), but gradually flatten over rounds, mirroring (b). This trend aligns with qualitative observations. Round 1 primarily improves general quality, whereas stylization becomes more prominent in subsequent rounds.

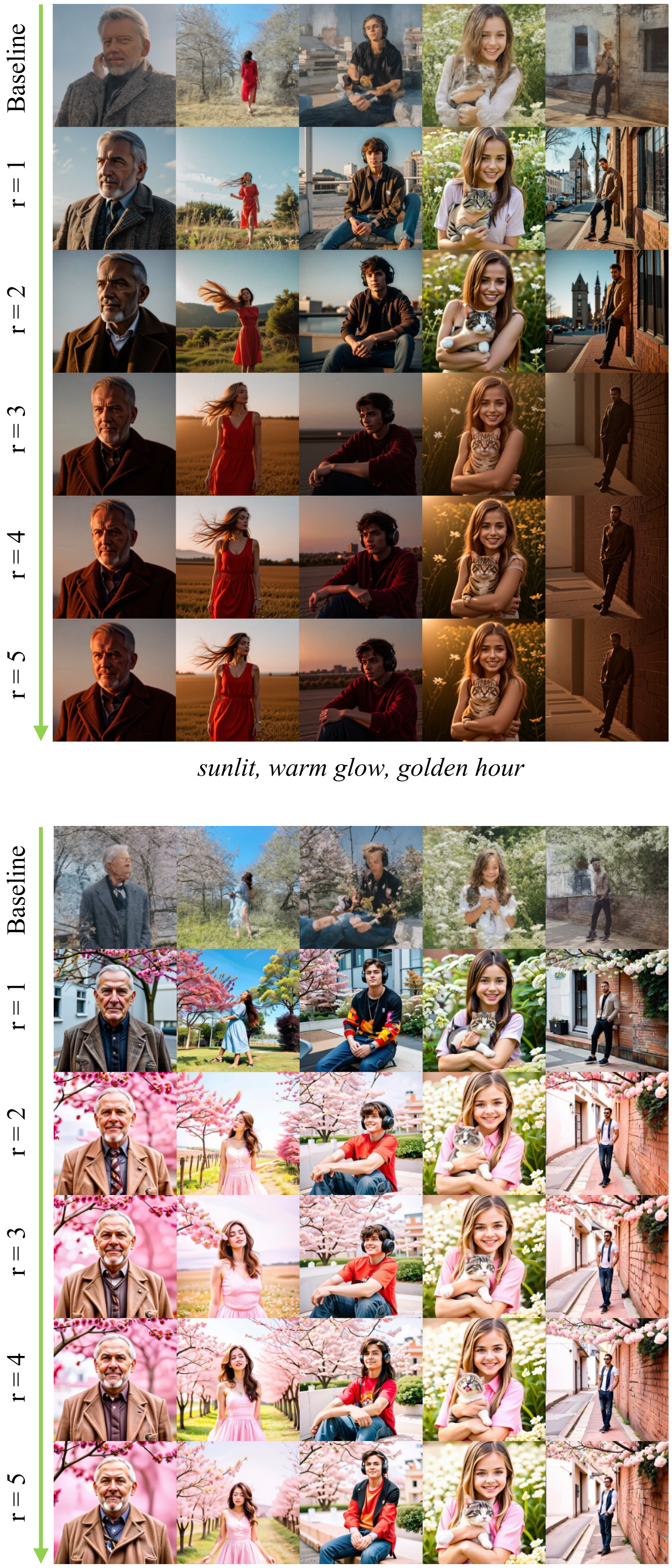
demonstrate that head sets found using a small subset of prompts generalize well to broader and more diverse prompt distributions. At inference time, attention perturbation guidance adds negligible overhead, as it simply modifies a fixed subset of attention maps without altering the model architecture or requiring additional forward passes.



(a) Unconditional generation results

(b) Some heads express style only when using style prompts.

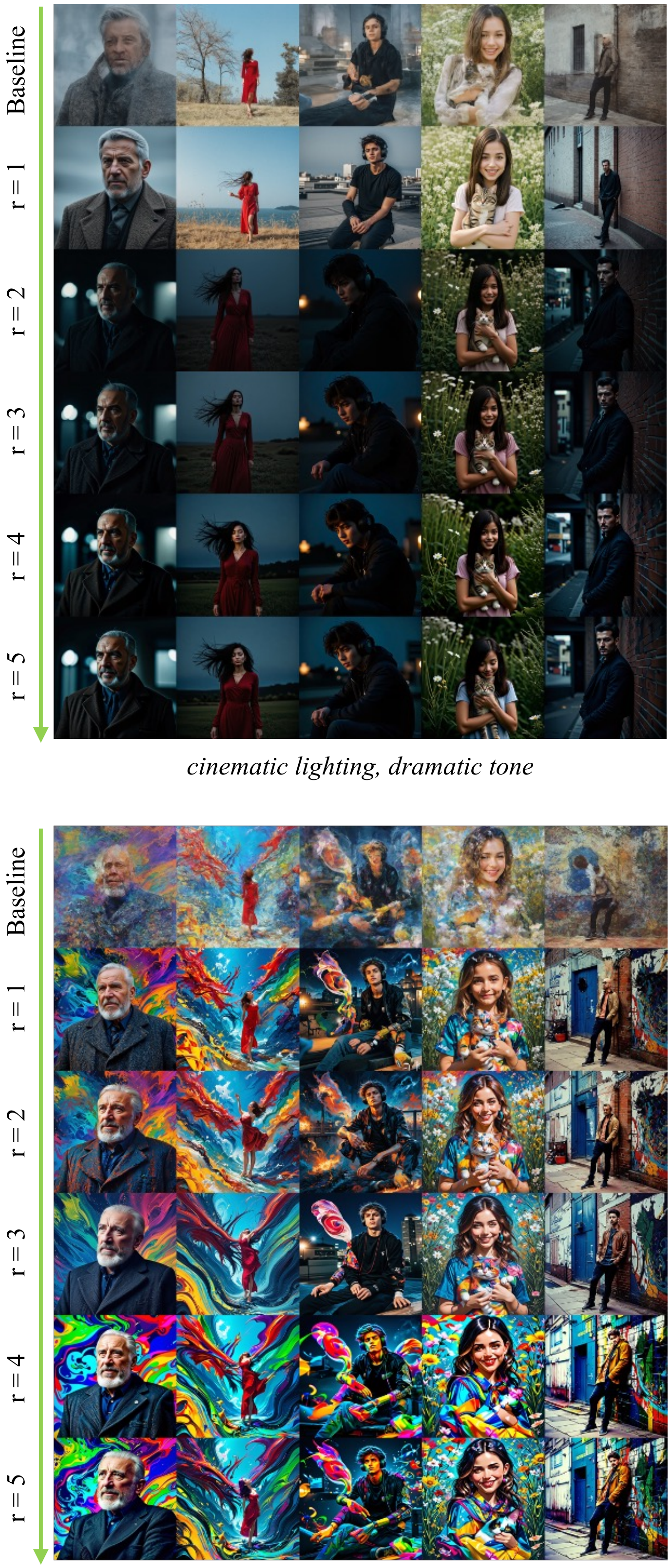
Figure 23: **Investigation whether style enhancement occurs without style prompt when applying HeadHunter (style-oriented quality setting).** $\mathcal{Q}$ indicates prompt and seed pairs. (a) In some cases, applying HeadHunter in even an unconditional setting amplifies the corresponding style to some extent. Style 1: “art nouveau style, elegant, decorative, curvilinear forms, nature-inspired, ornate, detailed”, Style 2: “line art drawing, professional sleek modern minimalist graphic vector graphics", Style 3: “cubist abstraction, fragmented planes, bold geometric angles”, Style 4: “cinematic lighting, dramatic tone”. (b) In the other cases, style enhancement depends on the presence of style prompts. For these heads, the stylization effect weakens without explicit style descriptions, highlighting the importance of including style prompts during the head selection process. Style prompt : "psychedelic style, vibrant colors, swirling patterns, abstract forms, surreal, trippy, colorful reflective voids, perceptual disorientation, sinuous forms, color psychology"



sunlit, warm glow, golden hour

spring bloom, fresh vibrant colors

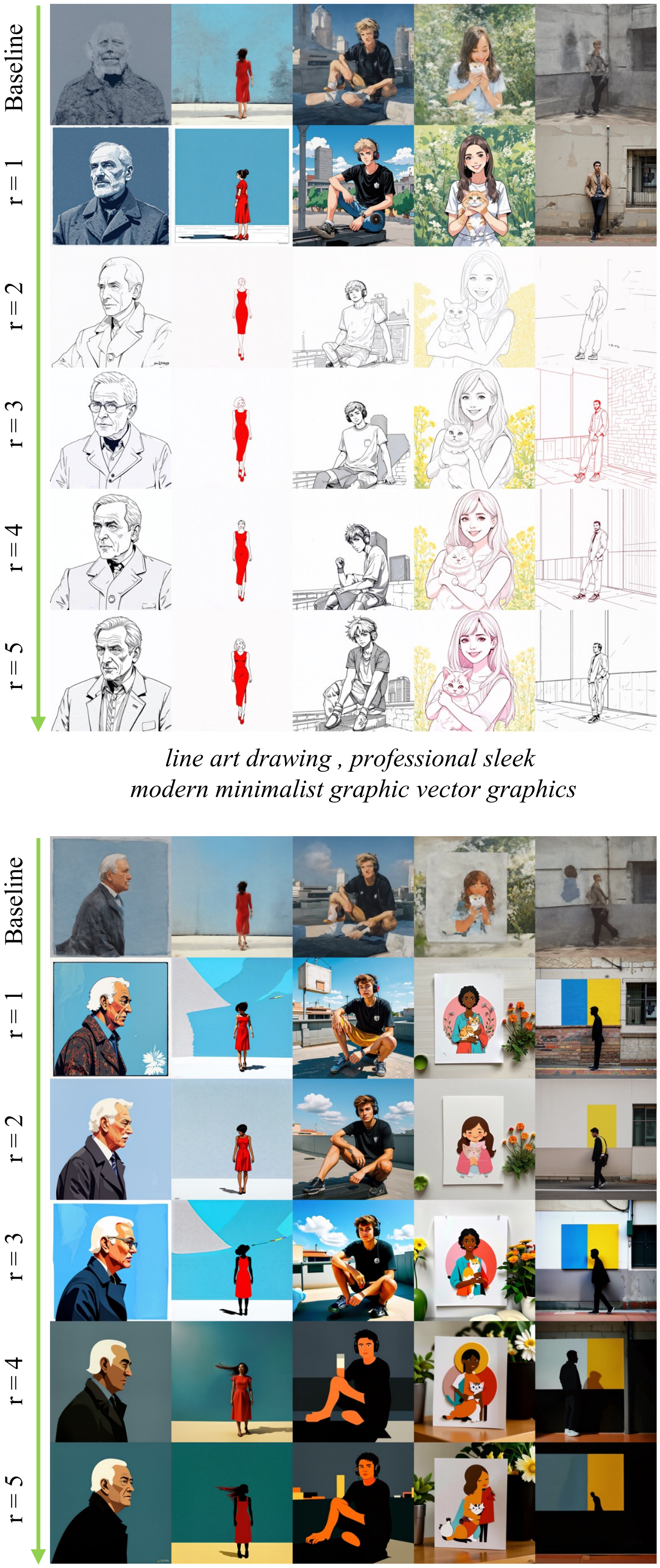
Figure 24: **Qualitative results of HeadHunter for style-oriented image quality improvement on SD3.**



cinematic lighting, dramatic tone

psychedelic style, vibrant colors, swirling patterns, abstract forms, surreal, trippy, colorful

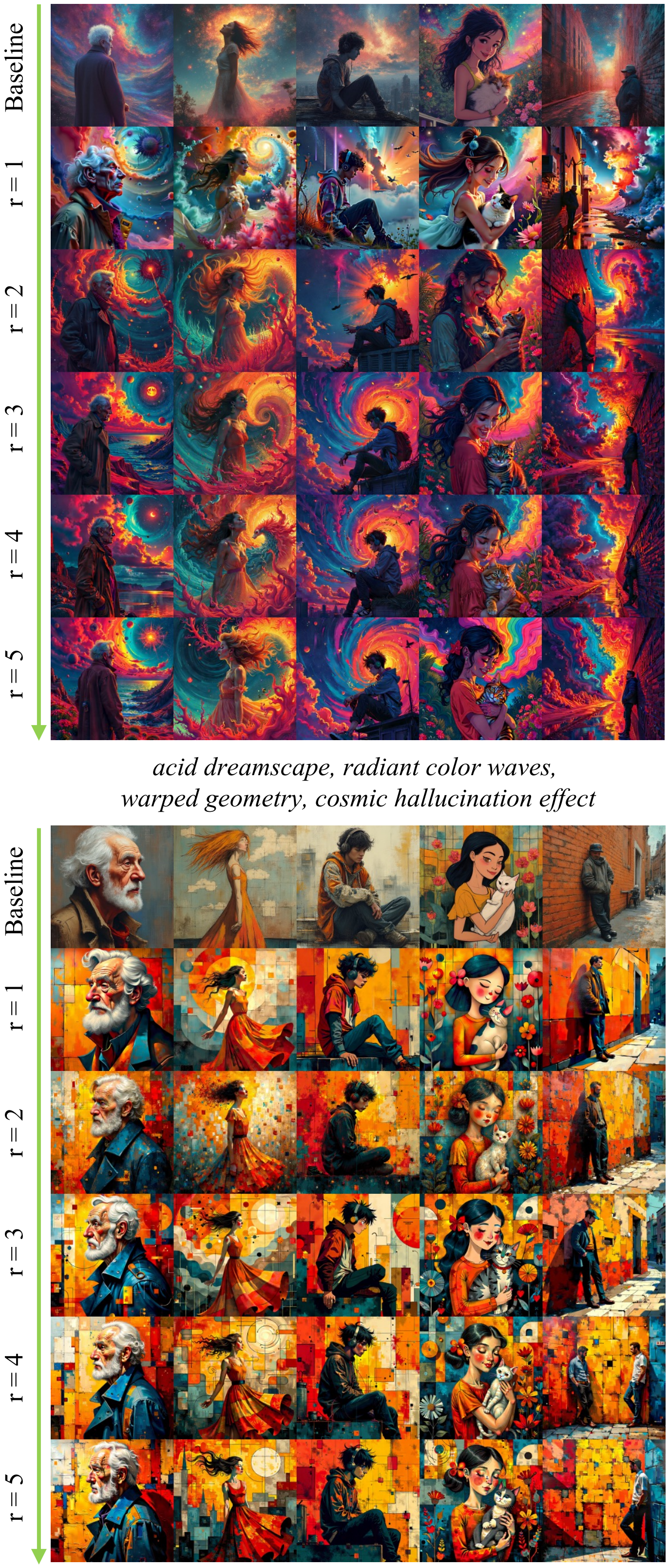
Figure 25: **Qualitative results of HeadHunter for style-oriented image quality improvement on SD3.**



line art drawing , professional sleek modern minimalist graphic vector graphics

flat papercut style, silhouette, clean cuts, paper, sharp edges, minimalist, color block

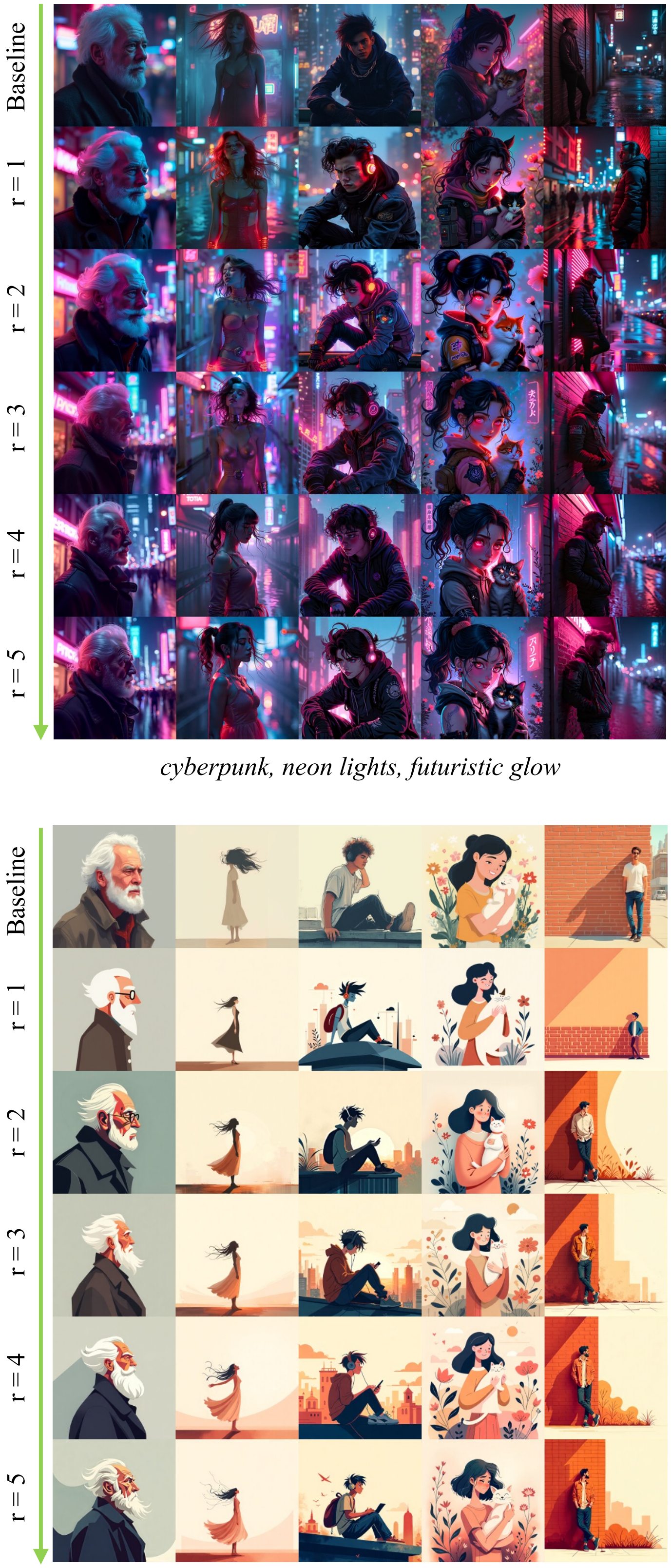
Figure 26: **Qualitative results of HeadHunter for style-oriented image quality improvement on SD3.**



acid dreamscape, radiant color waves,
warped geometry, cosmic hallucination effect

cubist artwork , geometric shapes
abstract innovative revolutionary

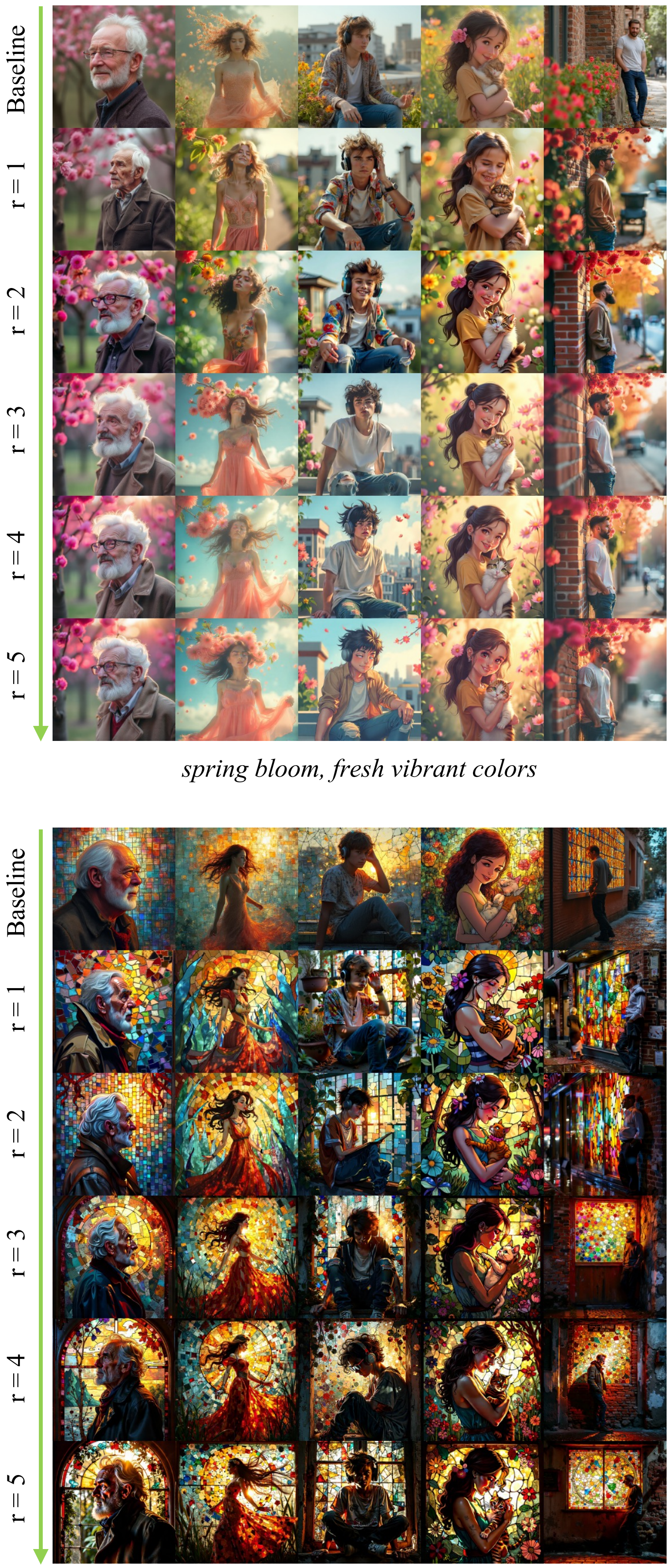
Figure 27: **Qualitative results of HeadHunter for style-oriented image quality improvement on FLUX.1-Dev.**



cyberpunk, neon lights, futuristic glow

flat vector design, clean geometry,
negative space, minimalist

Figure 28: **Qualitative results of HeadHunter for style-oriented image quality improvement on FLUX.1-Dev.**



spring bloom, fresh vibrant colors

stained glass style, colorful glass fragments, lead came details, backlit, translucent, mosaic

Figure 29: **Qualitative results of HeadHunter for style-oriented image quality improvement on FLUX.1-Dev.**

E Additional results and analysis of head-level perturbation guidance

In this section, we provide an extended analysis of head-level perturbation guidance with a focus on interpretable visual concepts. Specifically, Sec. E.1 presents more individual head-level guidance results in Stable Diffusion 3 and FLUX.1-Dev. In Sec. E.2, we showcase examples of head-level guidance that exhibit clearly interpretable semantic effects. Finally, Sec. E.3 explores the compositional behavior of head-level guidance and how combining them yields richer visual outcomes.

E.1 Additional results of individual head-level guidance

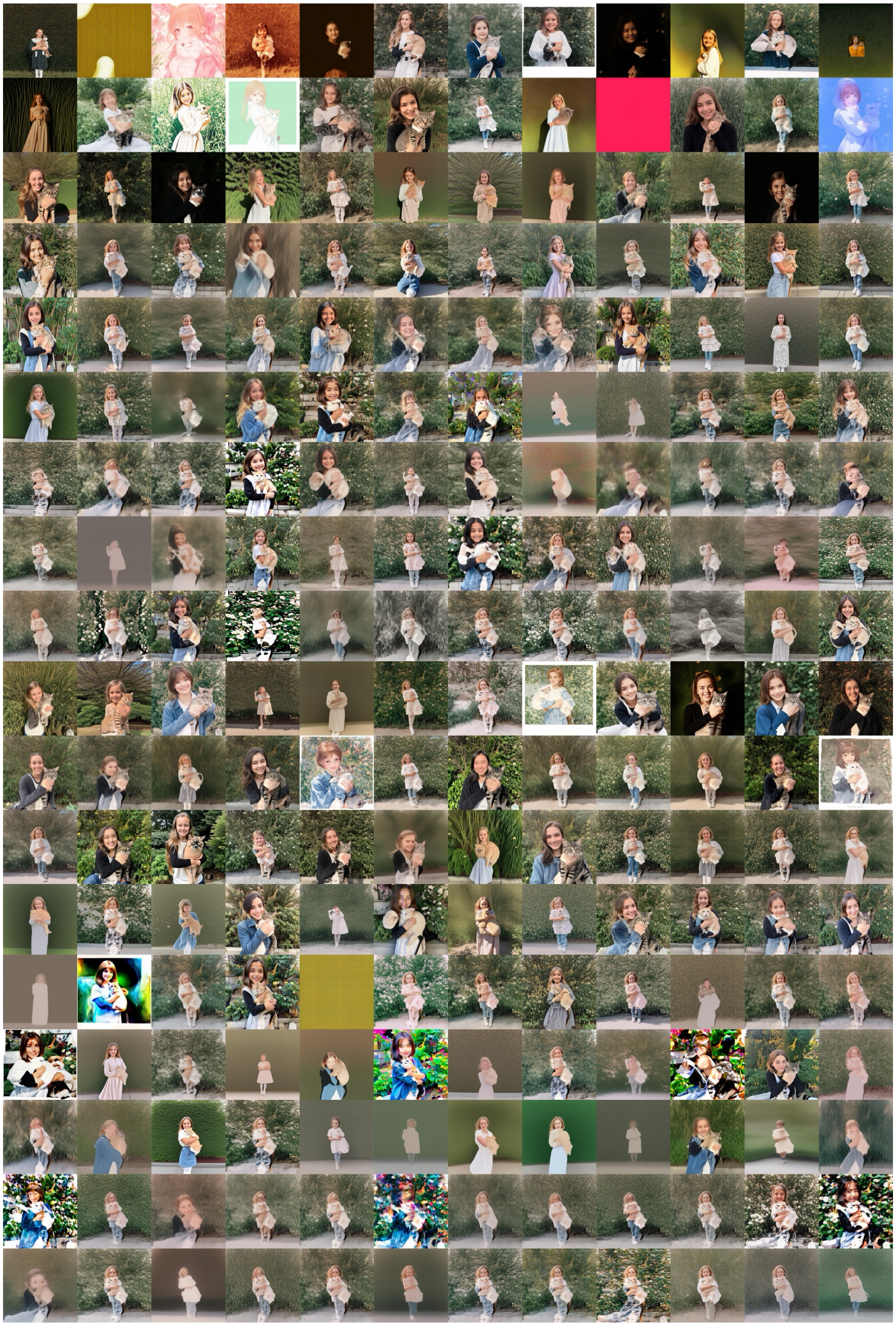


Figure 30: **Individual head-level guidance results on SD3.** Each cell corresponds to the output guided by perturbing a single head. Some heads induce notable effects, such as changes in lighting, structure, or color.

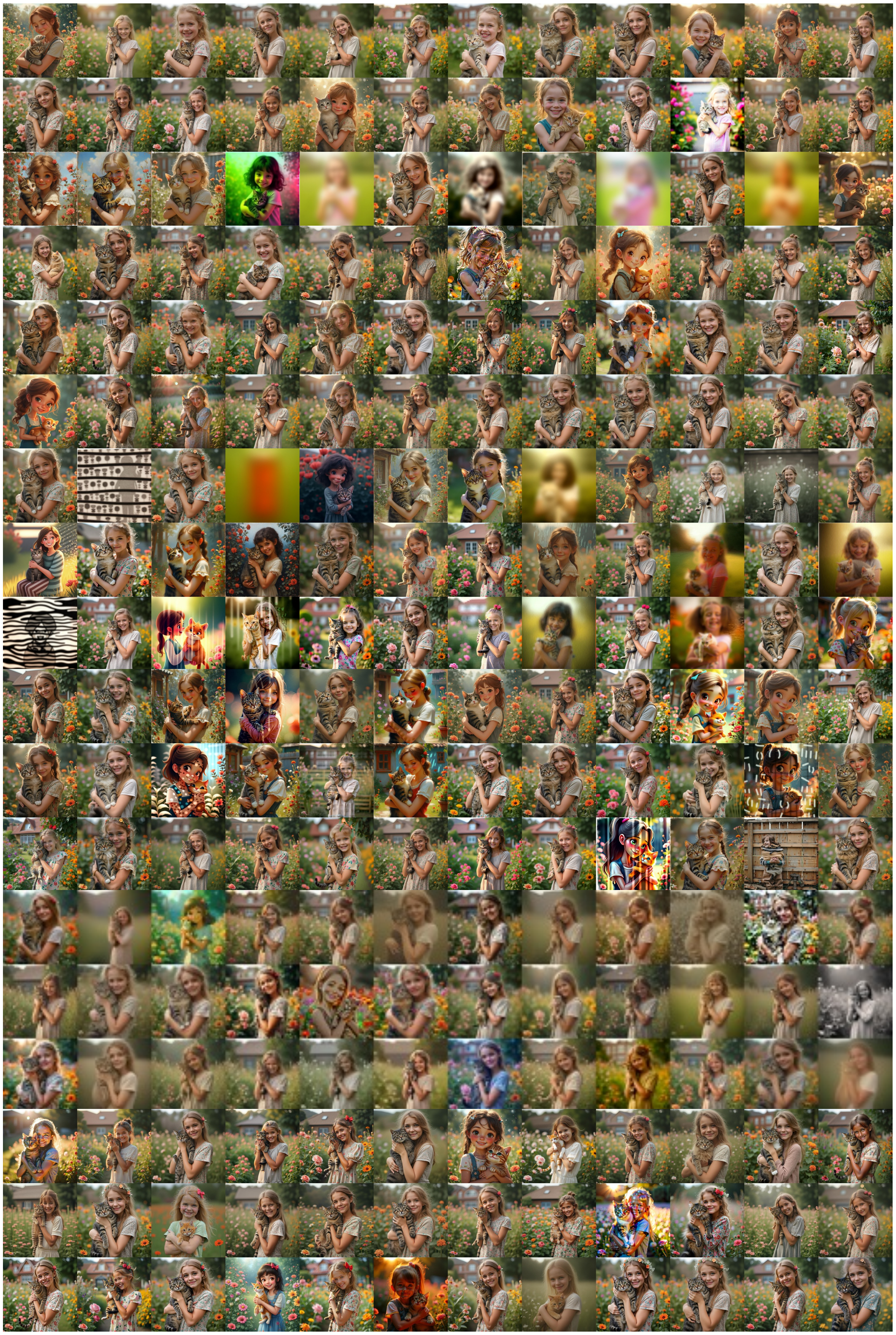
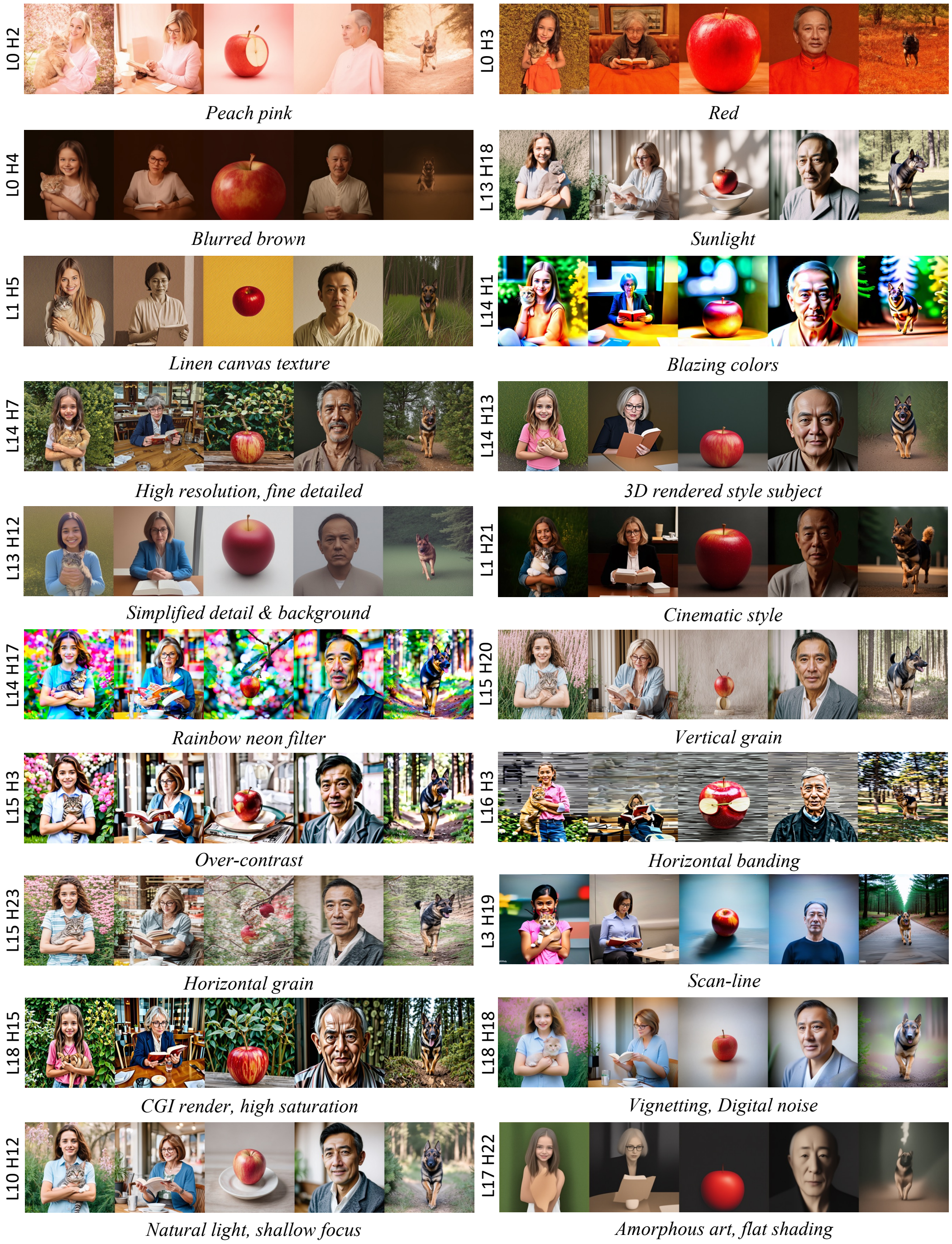


Figure 31: **Individual head-level guidance results on FLUX.1-Dev.** Each cell corresponds to the output guided by perturbing a single head. Some heads induce notable effects, such as changes in lighting, structure, or color.

E.2 Additional results on interpretable head-level guidance



Peach pink

Red

Blurred brown

Sunlight

Linen canvas texture

Blazing colors

High resolution, fine detailed

3D rendered style subject

Simplified detail & background

Cinematic style

Rainbow neon filter

Vertical grain

Over-contrast

Horizontal banding

Horizontal grain

Scan-line

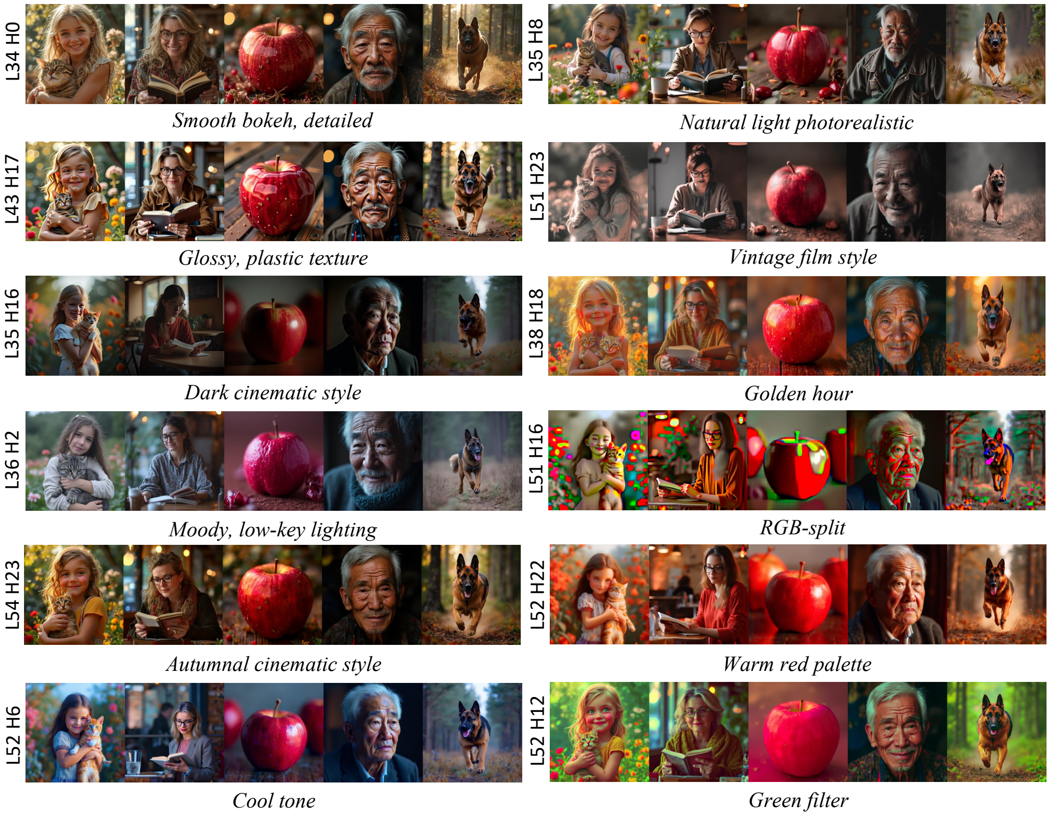
CGI render, high saturation

Vignetting, Digital noise

Natural light, shallow focus

Amorphous art, flat shading

Figure 32: **Interpretable heads in SD3.** Some heads exhibit consistent and clear visual effects when used for guidance. For instance, heads that add glow, alter geometry, or introduce specific lighting styles. These interpretable heads support the view that attention heads encode specific semantic concepts.



Smooth bokeh, detailed

Natural light photorealistic

Glossy, plastic texture

Vintage film style

Dark cinematic style

Golden hour

Moody, low-key lighting

RGB-split

Autumnal cinematic style

Warm red palette

Cool tone

Green filter

Figure 33: **Interpretable heads in FLUX.1-Dev.** Similar to SD3, FLUX also contains heads with distinctive effects, suggesting the generality of head-level interpretability across architectures.

E.3 Additional results on the combinational effect of head-level guidance

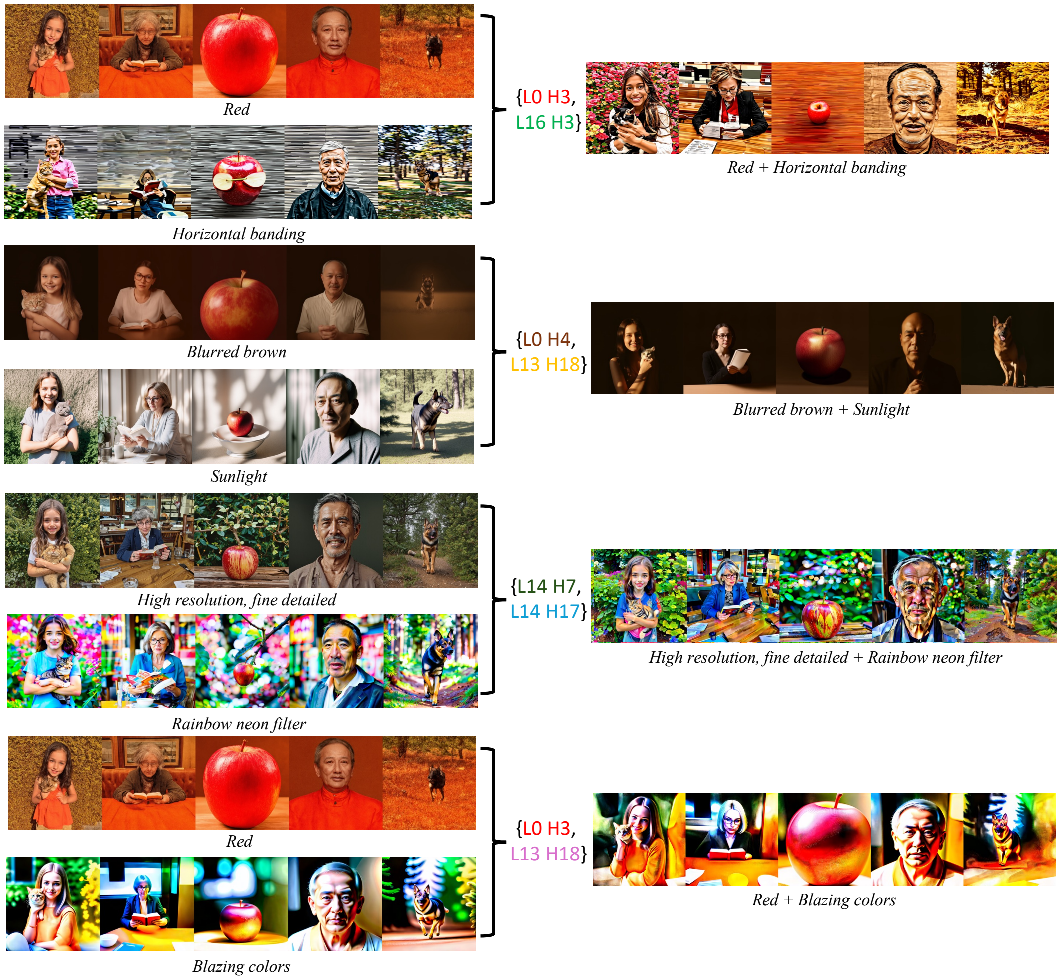


Figure 34: **Compositional effects of head-level guidance in SD3.** Combining heads leads to enriched outputs by *blending* their individual effects. This shows that attention heads can be composed to control more complex or stylized generations.

F Head-level perturbation guidance with different perturbation methods

In the main paper, we show the head-level analysis mostly on identity-matrix replacement perturbation. In Fig. 37. We show the results with another perturbation and the analysis. Note that one can freely choose any other perturbation methods, and we show some cases as examples.

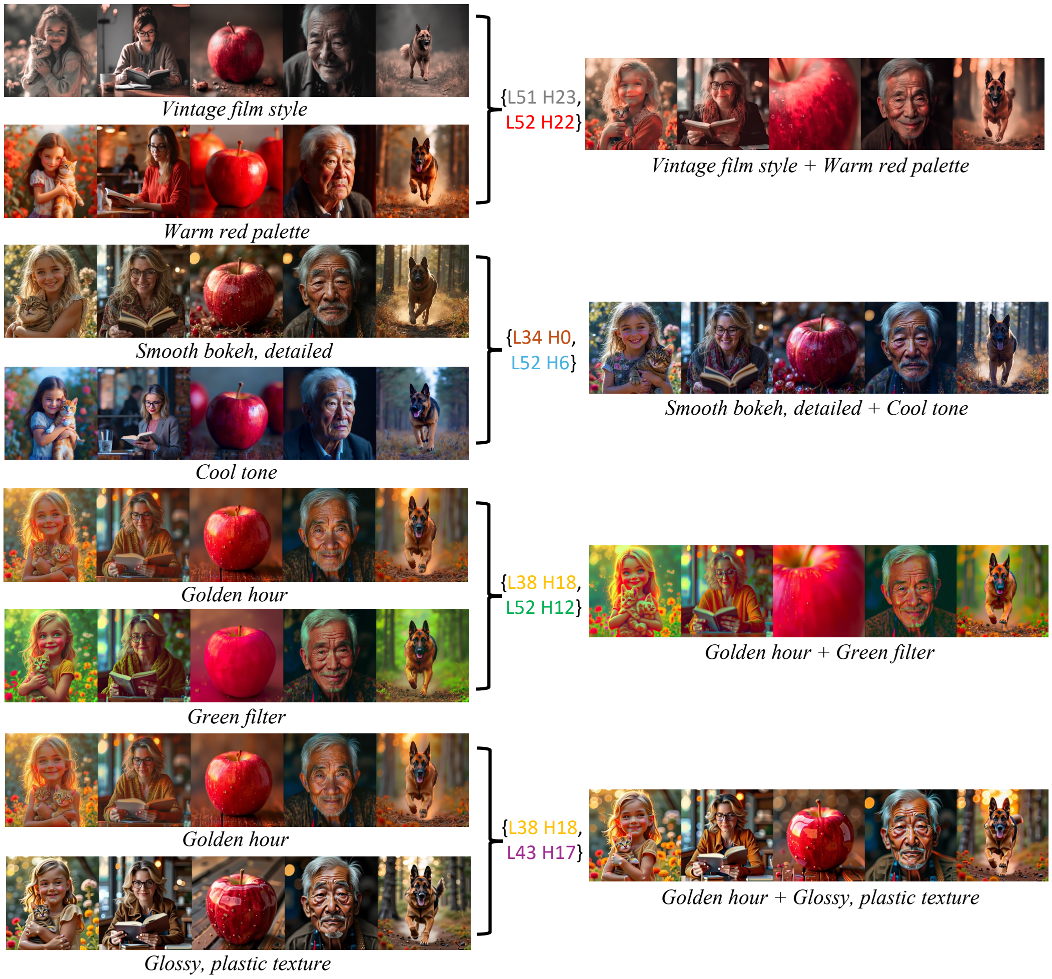
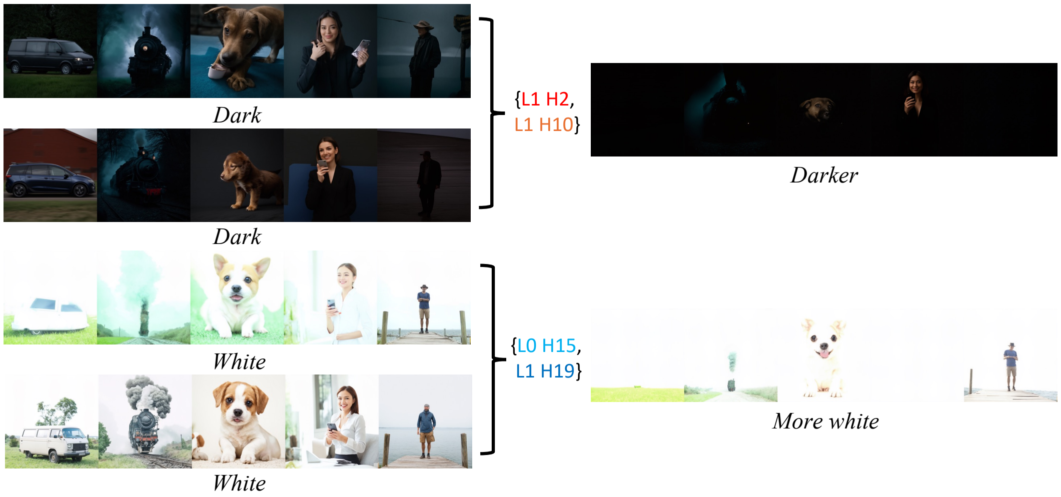


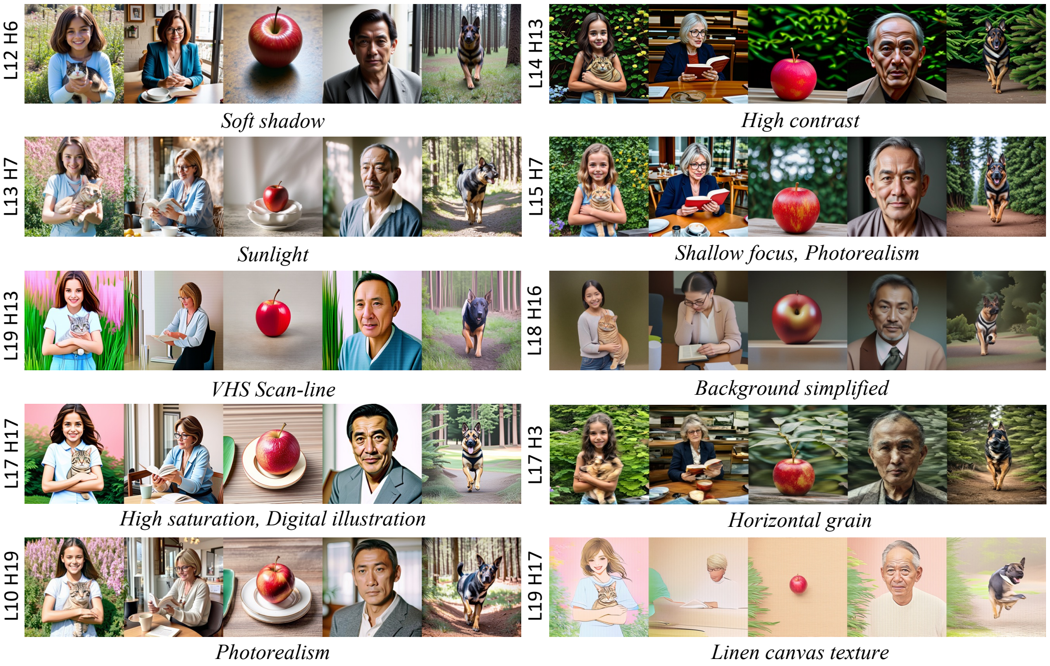
Figure 35: **Compositional effects of head-level guidance in FLUX.1-Dev.** Similar head-level guidance effects can be composed in FLUX.1-Dev to amplify or adjust stylistic elements.



(a) Individual head perturbation

(b) Multiple heads perturbation

Figure 36: **Concept amplification by combining stylistic heads.** Composing “dark heads” makes the generation darker, while “white heads” make it brighter. This demonstrates fine-grained control over stylistic dimensions via head-level selection.



(a) Uniform Guidance

(b) Smoothed Energy Guidance

Figure 37: **Head-Level perturbation guidance with different perturbations.** We show interpretable heads with different perturbation methods.

G Experimental details

Unless otherwise specified, all images are generated using `Stable Diffusion 3-Medium` with 20 Euler sampling steps. The default guidance scale is set to $w = 5.0$.

Fig. 1 Motivating examples for head-level guidance. Perturbations are applied individually to heads from layers 1, 3, and 14. Prompts are sampled from the MS-COCO validation set.

Fig. 2 Generated images from head- and layer-level perturbation guidance. In (a), heads with PickScore above a fixed threshold are selected (indicated by red boxes). Prompts include ‘‘`a medieval castle in the forest`’’ and ‘‘`a futuristic cityscape`’’.

Fig. 3 Effect of head-level guidance on concept amplification. Each head is perturbed independently. Prompts used include: `"smiling girl holding a cat, in a flower garden"`, `"middle-aged woman with glasses reading a book in a café"`, `"A macro shot of a red apple"`, `"a close up portrait of an elderly Japanese man, soft lighting, 8k"`, and `"A German shepherd bounding through a pine forest"`. Guidance scale is set to $w = 5.0$.

Fig. 4 Results of HeadHunter for general image quality improvement. HeadHunter is applied with $R = 1$ and $k = 24$. Prompts used include ‘‘`smiling girl holding a cat, in a flower garden`’’, ‘‘`businessman with sleek, glass-shiny black hair neatly parted, waiting for a morning train on a city platform`’’, and ‘‘`towering lighthouse casting a rotating beam across storm-tossed waves at twilight`’’. Metrics such as PickScore and AES are used for evaluation.

Fig. 9 Linear interpolation between attention map A and identity matrix I (SoftPAG). Images are generated with perturbation applied to layer 9. Prompts are generated from ChatGPT.

Fig. 14 Linear interpolation between attention map A and uniform matrix U (UG). Perturbation is applied to layer 9. Prompts are the same as those used in Fig. 9.

H Experimental resources

All experiments, including testing perturbation methods, analyzing head-level guidance, running HeadHunter, and conducting ablation studies, are performed using a mix of 8 NVIDIA H100 GPUs, 6 NVIDIA RTX 3090 GPUs, and 2 NVIDIA A6000 GPUs.

I Limitations & Broader Impacts

While our method is effective, there remains room for improving its efficiency. Running HeadHunter iteratively can be computationally intensive, especially for large-scale models with many heads. That said, the selected heads tend to generalize well across prompts and latents, making reuse practical. Exploring faster head and parameter selection strategies offers an exciting direction for future work.

Our work improves quality and stylistic control in diffusion models through guidance. This can benefit creative applications such as digital art, design. However, enhanced quality may also increase the risk of misuse, including the creation of deceptive or harmful content like deepfakes. While our work does not involve identity synthesis or model release, we acknowledge this risk and recommend responsible deployment practices such as usage restrictions and content disclosure.

Table 3: **Prompt and seed list used for general quality improvement via HeadHunter.**

Prompt	Seed
a close up of an old Japanese man, soft lighting, 8k	0
A black man with long braids, wearing white robes and heavy metal jewelry, pointing at the viewer against a dark background in the style of cinematic photography for a fashion shoot.	0
A photo of an Asian girl with her hand on her nose, with red nail polish and a bandaid, with black hair, a close up portrait, in the style of Rinko Kawauchi.	0
Japanese woman with short black hair, bangs hairstyle, hand covering eye, ring rings on, cool pose, outdoor sunlight background, natural lighting, portrait photography, in the style of Leica Q2 camera	0
A group of models of diverse ethnicities wearing various Nike sneakers, all dressed in white and colorful with patterns and designs inspired by the pop art movement, posed together for an editorial photoshoot against a backdrop featuring abstract collage-like elements in shades of red and blue. The scene was vibrant and dynamic, capturing their energetic expressions and stylish outfits in the style of the pop art movement.	0
photo of a black special forces soldier doing an olympic ski jump, holding two katanas in his hands, white background, 35mm film stills in the style of Cai Guo-Qiang and James	0
A tall German man sitting in a bench, reading a book, medium shot	0
A breakdancer doing a backflip	0
A photograph of an old oak tree in the middle, surrounded by dense woodland foliage, with sunlight filtering through the leaves and casting dappled light on the ground. The image was shot using a Hasselblad camera, Kodak Gold 200 film, and Portra 800 film stock.	0
a close-up of a woman's hand with delicate fingers, soft lighting, photorealistic	0
a photo of beautiful girl	0-9

Table 4: **Content prompts and corresponding seeds used for style-oriented quality improvement via HeadHunter.**

Prompt	Seed
portrait of an elderly man with white hair, wearing a wool coat, looking into the distance	0
young woman in a red dress, standing in the wind, eyes closed	1
teenage boy with messy hair, wearing headphones, sitting on a rooftop	2
smiling girl holding a cat, in a flower garden	3
a man leaning against a brick wall, hands in pockets, calmly observing the street	4