UNDERSTANDING UNIFIED VLMS: FROM ATTENTION PROBING TO GUIDANCE TUNING FOR MULTIMODAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified vision-language models (VLMs) perform both text-to-image (T2I) and image-to-text (I2T) generation with one shared Transformer backbone, yet how the backbone processes the condition in either direction remains poorly understood. We study this question from the perspective of conditional guidance: we probe every attention head to show the activation difference between two branches of classifier-free guidance (CFG) with different conditional guidance. On two unified VLMs of autoregressive and masked-diffusion paradigms, the probing reveals a subset of heads that are more sensitive to the change of condition while others are agnostic to it. This characteristic pattern holds in both T2I and I2T generation and is largely stable across data samples, demonstrating the internal mechanism of unified VLMs. Motivated by this, we propose a new inference-time intervention approach for unified multimodal generation, named *HeadTuned-CFG*, which uses the probed condition-change awareness to select and tune the conditional guidance at the attention-head level. Experiments show consistent benefits in both multimodal generation directions of unified VLMs without additional training, and in particular substantial improvements over strong baselines in T2I generation. These results further support the key finding of our probing analysis: the observed pattern of functional specialization is an intrinsic property of unified VLMs.

1 INTRODUCTION

Unified vision-language models (VLMs) are advancing rapidly, aiming to perform multimodal understanding and generation within a single model. Autoregressive models such as Janus (Wu et al., 2024a) and Emu3 (Wang et al., 2024b), as well as discrete diffusion models such as Muddit (Shi et al., 2026) and MMaDA (Yang et al., 2025), have shown strong performance on unified multimodal tasks such as text-to-image (T2I) and image-to-text (I2T) generation, narrowing the gap with task-specific models. Despite substantial differences in their training objectives and decoding strategies, these models commonly rely on a shared Transformer backbone to process information across modalities.

This raises the question of how a single shared backbone supports cross-modal understanding and generation under corresponding prompt guidance. Research has shown that attention heads in LLMs take on identifiable roles, such as induction and retrieval heads (Olsson et al., 2022; Wu et al., 2024b; Wang et al., 2022). Studies of VLMs have followed this line to study the understanding direction, examining how visual information is processed and attended to inside the language model (Neo et al., 2025; Jiang et al., 2025; Kim et al., 2025). However, the interpretability of such mechanisms on unified VLMs still remains under-explored.

In this work, we study this question through the lens of *conditional guidance* at the attention-head level. Specifically, we compare the activations of individual attention heads between the positive and negative branches of classifier-free guidance (CFG) (Ho & Salimans, 2022), which differ only in the conditional guidance. The comparison reveals that a subset of heads responds strongly to the change of condition, while some barely respond, as shown in Figure 1. We further find that this condition-change awareness is largely stable across samples, and that the variation is present in both

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

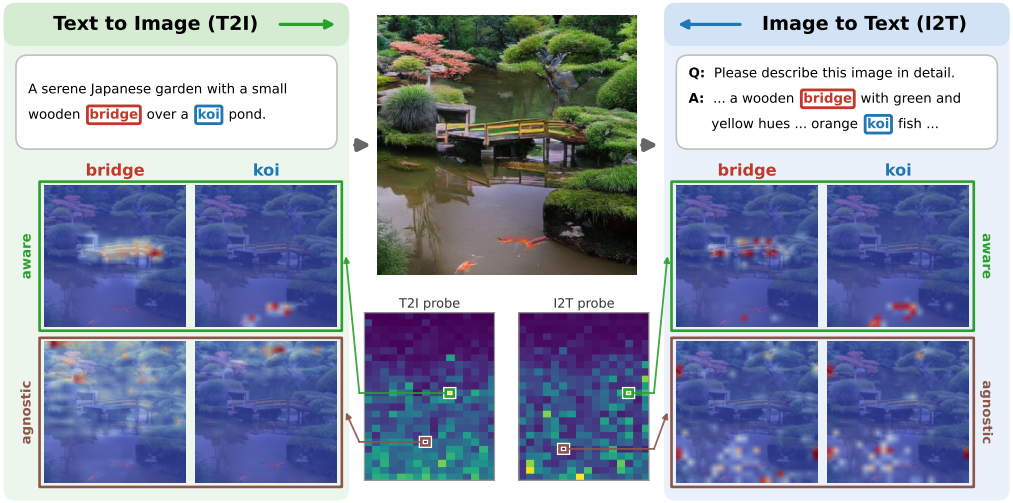

Figure 1: Condition-change aware (green) and agnostic (brown) attention heads of Janus. **Left: T2I generation**; each map shows where the generated image attends to a boxed keyword of the prompt. **Right: I2T generation** with the same image as input; each map shows where a boxed keyword of the generated caption attends in the input image. **Middle: the generated (input) image and head-level condition-change awareness map** of both T2I and I2T probing (layer by head, brighter means higher awareness), with the visualized heads marked. The Muddit counterpart is in Appendix A.

T2I and I2T generation of the two studied VLMs, revealing an intrinsic functional specialization of unified VLMs.

Motivated by these findings, we introduce a head-level guidance tuning method for multimodal generation in unified VLMs (**HeadTuned-CFG**), which explicitly accounts for head-level differences in condition-change awareness when applying CFG. We evaluate HeadTuned-CFG on both multimodal generation directions. For T2I generation, HeadTuned-CFG consistently outperforms vanilla CFG and related guidance methods on standard metrics across two studied VLMs. For I2T generation, it composes directly with CFG-style contrastive decoding such as VCD (Leng et al., 2023) and shows the potential to further improve the robustness of VLMs against hallucination. These results further support our probing analysis as an intrinsic property, facilitating simple yet effective guidance tuning mechanism for unified multimodal generation.

In summary, our contributions are as follows:

- **Understanding via Attention Probing:** We investigate unified VLMs from the perspective of conditional guidance by developing a condition-change probing framework for analyzing their internal mechanism, uncovering a consistent functional specialization pattern of attention heads in both T2I and I2T generation.
- **Guidance Tuning for Multimodal Generation:** Motivated by the observed probing pattern, we propose HeadTuned-CFG, which selects heads based on condition-change awareness and adjusts the guidance strength of the selected heads through a linear scaling of their activations. We also provide a theoretical interpretation of the tuning rule under a local Gaussian model, showing how a KL projection with a prescribed positive-awareness constraint yields its interpolation/extrapolation form.
- **Experiments across Benchmarks and Backbones:** We demonstrate that HeadTuned-CFG consistently benefits both multimodal generation directions of unified VLMs, and in particular substantially outperforms vanilla CFG and related guidance methods in T2I generation.

2 PRELIMINARIES

2.1 UNIFIED VISION-LANGUAGE MODELS

Unified VLMs typically encode multimodal inputs into discrete tokens of a shared vocabulary using a visual tokenizer (Esser et al., 2021) and a text tokenizer, and can be broadly divided by their generation paradigm into autoregressive and masked-diffusion approaches.

Autoregressive VLMs. These models (Team, 2025; Wang et al., 2024b; Wu et al., 2024a) unify generation tasks under the next-token prediction paradigm. A single decoder-only Transformer is trained on interleaved image-text sequences via causal language modeling with the standard autoregressive negative log-likelihood objective:

$$\mathcal{L} = -\sum_i \log p_\theta(x_i \mid x_{<i}), \tag{1}$$

where $x = (x_1, \ldots, x_N)$ is a sequence of discrete image and text tokens and $x_{<i}$ denotes the tokens before position i. At inference time, generation amounts to autoregressively sampling tokens from $p_\theta(x_i \mid x_{<i})$.

Masked-Diffusion VLMs. These models (Shi et al., 2026; Yang et al., 2025) instead formulate multimodal generation as a discrete denoising process. Let x_0 be the clean token sequence. The noisy sequence x_t at step $t \in [0, 1]$ is obtained by replacing each token of x_0 independently with a special `[MASK]` token with probability $\beta(t)$, where the noise schedule $\beta(t)$ increases from $\beta(0) = 0$ to $\beta(1) = 1$. The Transformer is trained to recover the original tokens from their corrupted counterparts via a masked denoising objective:

$$\mathcal{L} = -\sum_t \mathbb{E}_{x_t}[\log p_\theta(x_0 \mid x_t)]\,. \tag{2}$$

At inference time, generation proceeds by iteratively unmasking tokens starting from a fully masked initial sequence.

2.2 CLASSIFIER-FREE GUIDANCE

Classifier-free guidance (CFG) (Ho & Salimans, 2022) enhances conditional generation by running two forward passes and extrapolating between their outputs: a *positive* branch conditioned on the condition and a *negative* branch conditioned on a corrupted condition. Denoting their outputs by $\hat{\epsilon}_{\text{pos}}$ and $\hat{\epsilon}_{\text{neg}}$, the guided output is

$$\hat{\epsilon}_{\text{CFG}} = \hat{\epsilon}_{\text{neg}} + w\left(\hat{\epsilon}_{\text{pos}} - \hat{\epsilon}_{\text{neg}}\right), \tag{3}$$

where w is the guidance scale. With $w > 1$, the output is pushed along the direction in which the two branches differ, which promotes the content specific to the condition and suppresses the content that would also be generated without it.

2.3 MULTI-HEAD ATTENTION

The multi-head attention mechanism (Vaswani et al., 2023) splits the query, key, and value matrices into H_l parallel heads, each operating in its own learned projection space. At the l-th layer of the model, the input projections for head $h \in \{1, \ldots, H_l\}$ take the form

$$\mathbf{Q}_{l,h} = \mathbf{Q}_l \mathbf{W}^Q_{l,h}, \qquad \mathbf{K}_{l,h} = \mathbf{K}_l \mathbf{W}^K_{l,h}, \qquad \mathbf{V}_{l,h} = \mathbf{V}_l \mathbf{W}^V_{l,h}, \tag{4}$$

with $\mathbf{Q}_l, \mathbf{K}_l, \mathbf{V}_l \in \mathbb{R}^{N \times d}$ denoting the shared query, key, and value activations, and $\mathbf{W}^Q_{l,h}, \mathbf{W}^K_{l,h}, \mathbf{W}^V_{l,h} \in \mathbb{R}^{d \times \bar{d}}$ denoting the per-head learnable projections of width $\bar{d}$.

Each head then independently produces an attention pattern and aggregates the values:

$$\mathbf{A}_{l,h} = \text{Softmax}\left(\frac{\mathbf{Q}_{l,h}\mathbf{K}^\top_{l,h}}{\sqrt{\bar{d}}}\right), \qquad \mathbf{O}_{l,h} = \mathbf{A}_{l,h}\mathbf{V}_{l,h}. \tag{5}$$

The outputs of all heads $\{\mathbf{O}_{l,h}\}_{h=1}^{H_l}$ are concatenated and mapped back into the model dimension via an output projection matrix $\mathbf{W}^O_l \in \mathbb{R}^{(H_l\bar{d}) \times d}$:

$$\text{MultiHead}(\mathbf{Q}_l, \mathbf{K}_l, \mathbf{V}_l) = \text{Concat}(\mathbf{O}_{l,1}, \ldots, \mathbf{O}_{l,H_l})\,\mathbf{W}^O_l. \tag{6}$$

2.4 Related Work

Classifier-Free Guidance. CFG was proposed for conditional generation in continuous diffusion models (Ho & Salimans, 2022) and is now extended to the T2I generation of unified VLMs (Wu et al., 2024a; Shi et al., 2026). For I2T generation, CFG-style contrastive decoding, such as Visual Contrastive Decoding (VCD) (Leng et al., 2023) and its variants (Favero et al., 2024; Wang et al., 2024c), builds the negative branch from a perturbed image and contrasts it with the original one to suppress object hallucination. Several studies improve CFG by scheduling the guidance scale across generation timesteps (Wang et al., 2024a; Kynkäänniemi et al., 2024). However, these approaches treat the model as a black box and apply a global scale, rarely examining the underlying architectural mechanisms, such as the roles of individual attention heads.

Another line of work intervenes inside the architecture: Perturbed-Attention Guidance (PAG) (Ahn et al., 2025a) and related methods (Hong et al., 2023; Hong, 2024; Ahn et al., 2025b) take a forward branch perturbed at intermediate attention as the reference and apply the same linear extrapolation as CFG, which extends guidance to tasks without explicit conditioning, such as inverse problems and inpainting. They do not outperform CFG on common conditional generation, and are therefore mainly used as a substitute for CFG where it is inapplicable or stacked on CFG to improve it. Moreover, designed for the diffusion paradigm, they cannot be equivalently adapted to autoregressive models.

Unified-VLM Interpretability. Attention heads in LLMs take on identifiable roles: induction heads implement in-context copying (Olsson et al., 2022), retrieval heads fetch information from long contexts (Wu et al., 2024b), and a small circuit of heads implements a specific task (Wang et al., 2022). Interpretability studies of VLMs have followed this line in the understanding direction: Neo et al. (2025) trace how visual tokens are processed across the layers of the language model, Jiang et al. (2025) relate object hallucination to the attention paid to image tokens in the middle layers, and Kim et al. (2025) attribute the image-to-text information flow to a subset of heads. However, the interpretability of such mechanisms on unified VLMs remains poorly explored.

3 Understanding Condition-change Awareness of Attention Heads: A Probing-based Analysis

To understand how individual attention heads behave differently under CFG, we probe the response of attention heads to the change of the condition on two unified VLMs: Muddit (Shi et al., 2026), a masked discrete diffusion model with an MMDiT (Esser et al., 2024) backbone, and Janus-1.3B (Wu et al., 2024a), an autoregressive model with a LLaMA-style (Touvron et al., 2023) decoder.

Algorithm 1 Head-level condition-change awareness probing.

Input: prompt set $\mathcal{P}$; generation steps $\mathcal{T}$
Output: condition-change awareness scores $\delta^{(l,h)}(p,t)$ for every head (l,h)
for $p \in \mathcal{P}$ **do**
    run one CFG generation, hooking every head's activation on both the positive and negative branches
    for $t \in \mathcal{T}$ **do**
        $\delta^{(l,h)}(p,t) \leftarrow \|\mathbf{h}_{\text{pos}}^{(l,h),t} - \mathbf{h}_{\text{neg}}^{(l,h),t}\|_2 \quad \forall (l,h)$
    end for
end for
return $\{\delta^{(l,h)}(p,t)\}$ ▷ diffusion: kept per step; AR: averaged over t

Let $\mathcal{P}$ denote the set of probing samples, i.e., prompts for T2I and images for I2T, and $\mathcal{T}$ the forward passes of one generation, i.e., the denoising steps for diffusion models and the generation steps for AR models. At each step, the model performs a standard CFG forward pass, and we hook $\mathbf{h}^{(l,h)}$, an intermediate activation of each attention head (l,h), i.e., the h-th head of the l-th layer, on both branches; the exact activation and its positions depend on the model architecture and are given in Appendix B. We then compute $\delta^{(l,h)} = \|\mathbf{h}_{\text{pos}}^{(l,h)} - \mathbf{h}_{\text{neg}}^{(l,h)}\|_2$ as its *condition-change awareness*, the response of the head to the change of the condition between the two branches. Algorithm 1 gives the probing procedure. We apply this probing to the T2I and I2T generation of both models with

$|\mathcal{P}| = 500$ samples per task (settings in Appendix B), and average $\delta^{(l,h)}$ over $\mathcal{P}$, and over steps for diffusion models, to obtain the layer-by-head awareness map of each model and task in Figure 2.

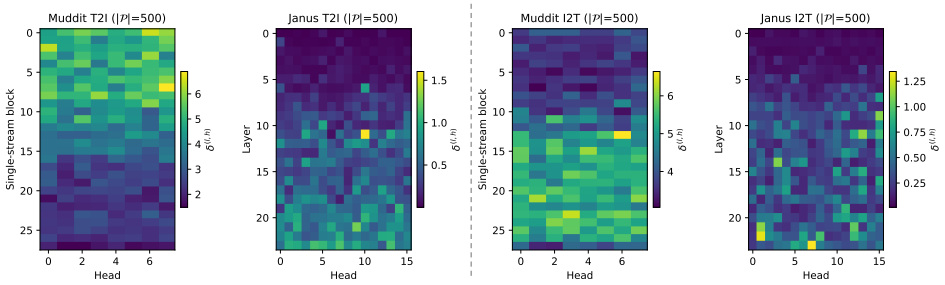

Figure 2: Layer-by-head condition-change awareness $\delta^{(l,h)}$ of Muddit and Janus, averaged over $|\mathcal{P}| = 500$ probing samples (and over steps for Muddit). Left of the dashed line: T2I generation; right: I2T generation.

3.1 Text-to-Image Generation

We first probe T2I generation on 500 GenEval (Ghosh et al., 2023) prompts for both models (Appendix B.1).

Different Heads Exhibit Heterogeneous Condition-change Awareness. We quantify the condition-change awareness of each head per prompt by $\delta^{(l,h)}$ averaged over generation steps. As shown in Figure 2 (left), both Muddit and Janus exhibit substantial variation across heads: some heads respond strongly to the condition change, whereas others barely respond.

Head-level Condition-change Awareness Is Stable Across Prompts. We next ask whether these head-level differences reflect persistent head behavior or are specific to individual prompts. We compute the mean pairwise Spearman rank correlation between per-prompt head rankings over $|\mathcal{P}| = 500$ prompts. For Janus T2I, each prompt yields a single ranking averaged over decode positions, with a correlation of 0.970 ± 0.014. For Muddit T2I, rankings are computed independently at each denoising step; the correlation averages 0.975 across steps and remains above 0.97 except for the final few steps (Figure 6, left, in Appendix C). These results indicate that a head's relative awareness of condition changes is largely intrinsic rather than prompt-specific, providing a stable signal for characterizing head behavior.

Condition-change Aware Heads Attend to Semantically Relevant Regions. Figure 1 reveals a clear functional distinction between condition-change aware and condition-change agnostic heads. We select heads from the high end of the probed $\delta^{(l,h)}$ as condition-change aware and from the low end as condition-change agnostic, and visualize the attention of each generated image token to words in the prompt. Aware heads concentrate their attention on the region of the queried keyword, whereas agnostic heads attend diffusely with little semantic correspondence.

3.2 Image-to-Text Generation

For I2T generation, CFG is not a standard component of decoding. However, methods such as Visual Contrastive Decoding (VCD) (Leng et al., 2023) use a specially constructed perturbed image as the negative branch of CFG and improve the robustness of VLMs against hallucination. Our probing analysis follows this construction, taking the perturbed image as the negative condition. This is conducted on the captioning task over 500 MS-COCO (Lin et al., 2015) images (Appendix B.2).

Results. I2T generation exhibits the same heterogeneity as T2I generation: on both models a subset of heads responds strongly to the change of the image while the rest barely respond (Figure 2, right). Similarly, Figure 1 (right) shows the distinct pattern on captioning, where the aware head attends to the object relevant to the decoded keyword. The ranking of Janus is stable across samples, with a mean pairwise Spearman rank correlation of 0.962 (Table 5); However, the ranking of Muddit degrades steadily over the 32 denoising steps, staying above 0.90 only in the first 9 steps and falling to 0.67 at the last (Figure 6, right, in Appendix C).

4 METHOD: TUNING CFG AT HEAD-LEVEL

This section presents HeadTuned-CFG, which tunes CFG at the head level according to the probed condition-change awareness in two steps: selecting heads by their awareness and adjusting the guidance strength of the selected heads.

Motivation. Section 3 reveals functional specialization among attention heads under CFG. We hypothesize that the response of a condition-change agnostic head to the change of condition carries little condition-specific information and acts as noise, which CFG nevertheless amplifies together with the informative responses, whereas the response of a condition-change aware head points in the direction of the condition itself. This suggests adjusting the guidance per head rather than on the final output alone: interpolating the positive-branch activation of agnostic heads toward the negative branch suppresses the amplified noise, and extrapolating that of aware heads away from the negative branch reinforces the condition-specific direction and strengthens condition adherence. To verify this conjecture, we provide a theoretical justification in a simplified setting under a local Gaussian model in Theorem 1, with the proof deferred to Appendix D.

Theorem 1. *For one head, suppress (l, h) and let $p_b = \mathcal{N}(\boldsymbol{\mu}_b, \Sigma)$ with $\Sigma \succ 0$ for $b \in \{\text{pos}, \text{neg}\}$. Define LDA direction $\mathbf{w} = \Sigma^{-1}(\boldsymbol{\mu}_{\text{pos}} - \boldsymbol{\mu}_{\text{neg}})$ with $\mu_{\text{pos}} \neq \mu_{\text{neg}}$, and induced positive-awareness score $T(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$. The KL projection*

$$\min_q \mathrm{KL}(q \| p_{\text{neg}}) \quad \textit{s.t.} \quad \mathbb{E}_q[T(\mathbf{x})] \geq \tau \tag{7}$$

has solution $q^* = \mathcal{N}(\mathbf{h}_g, \Sigma)$, whose maximum a posteriori (MAP) estimate and tuning strength are

$$\mathbf{h}_g = \boldsymbol{\mu}_{\text{neg}} + s^*(\boldsymbol{\mu}_{\text{pos}} - \boldsymbol{\mu}_{\text{neg}}), \qquad s^* = \left[\frac{\tau - \mathbf{w}^\top \boldsymbol{\mu}_{\text{neg}}}{\mathbf{w}^\top \Sigma \mathbf{w}} \right]_+ . \tag{8}$$

Moreover, $s^* > 1$ (resp., $s^* < 1$) if and only if $\tau > \mathbb{E}_{p_{\text{pos}}}[T(\mathbf{x})]$ (resp., $\tau < \mathbb{E}_{p_{\text{pos}}}[T(\mathbf{x})]$).

Theorem 1 interprets head-wise tuning as a KL projection: we seek the distribution closest to the negative branch that attains a prescribed positive-awareness level. Under the local Gaussian model, its most probable activation is obtained by *interpolating or extrapolating* between the two branch means. Replacing these means with observed activations recovers our tuning rule. Targets above or below the unchanged positive branch's awareness level yield extrapolation ($s > 1$) or interpolation ($s < 1$), respectively. This motivates the two modes in our HeadTuned-CFG algorithm (see Algorithm 2): (a) extrapolating the top-k heads ranked by condition-change awareness to strengthen conditioning, and (b) interpolating the bottom-k heads toward the negative branch to suppress their uninformative response to the condition. Each mode uses a shared tuning strength s across the selected heads.

Algorithm. Given the condition-change awareness scores $\delta^{(l,h)}$, HeadTuned-CFG proceeds in two steps, *select* and *tune* (Algorithm 2), controlled by two hyperparameters: the fraction k of heads to tune and the tuning strength s.

Algorithm 2 HeadTuned-CFG (*select* and *tune*).

Input: scores $\delta^{(l,h)}$; mode (interpolation: bottom-k heads, $s < 1$; extrapolation: top-k heads, $s > 1$); Φ_s; w; steps $\mathcal{T}$, tuning steps $\mathcal{T}' \subseteq \mathcal{T}$
Output: generated sample
for $t \in \mathcal{T}$ **do**
  if $t \in \mathcal{T}'$ **then**
    $\mathcal{S}_k \leftarrow$ the k heads of the mode by $\delta^{(l,h)}$ ▷ per step for diffusion; fixed for AR
    run the positive and negative branches, replacing in the forward pass
    $\mathbf{h}_{\text{pos}}^{(l,h)} \leftarrow \Phi_s\left(\mathbf{h}_{\text{pos}}^{(l,h)}, \mathbf{h}_{\text{neg}}^{(l,h)}\right) \quad \forall (l,h) \in \mathcal{S}_k$
  else
    run the positive and negative branches
  end if
  combine the two branches with guidance scale w ▷ Eq. 3
end for
return sample

Select: heads are ranked by $\delta^{(l,h)}$, and the bottom-k fraction (interpolation mode) or the top-k fraction (extrapolation mode) forms the tuning set $\mathcal{S}_k$; for masked-diffusion models the set is selected at each denoising step from the per-step scores, and for AR models once from the averaged scores.

Tune: at a subset of steps $\mathcal{T}' \subseteq \mathcal{T}$, a tuning operator Φ_s is applied to the selected heads in $\mathcal{S}_k$ and modifies their positive-branch activation, displacing it along the direction between the positive and negative branches: $s = 1$ leaves the head unchanged and recovers vanilla CFG, $s < 1$ interpolates it toward the negative branch, and $s > 1$ extrapolates it away from the negative branch. For simplicity, the same s is applied uniformly to all selected heads. The interpolation mode thus tunes the bottom-k heads with $s < 1$ to suppress the noise amplified by CFG, whereas the extrapolation mode tunes the top-k heads with $s > 1$ to strengthen condition adherence. The concrete form of Φ_s and the choice of $\mathcal{T}'$ depend on the architecture and are given in Appendix E. Figure 3 illustrates the procedure on Muddit.

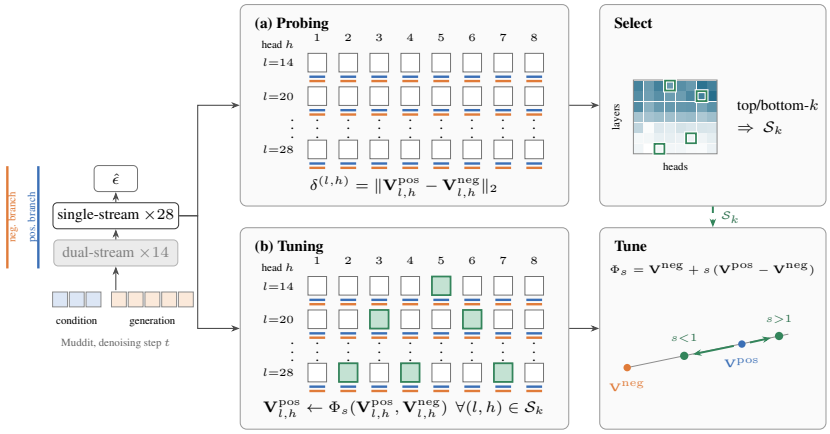

Figure 3: Proposed HeadTuned-CFG method on Muddit. Illustration of head-level probing, selection, and tuning process of the single-stream attention heads.

5 Experiments

5.1 Text-to-Image Generation

Setup. We evaluate Muddit and Janus on MJHQ-30K (Li et al., 2024), which contains 30,000 prompts in 10 categories of 3,000 prompts each. FID (Heusel et al., 2018) is the primary metric. Generation settings (resolution, denoising steps, and the negative branch) follow the probing setup in Appendix B.1, and the head ranking is the T2I probe of Section 3.1, obtained on GenEval prompts disjoint from MJHQ. We use the interpolation mode (bottom-k heads, $s < 1$) here, and grid-search (w, k, s) for the lowest FID on the *animals* category (MJHQ-animals; Appendix H.1), then apply the selected setting unchanged to all 10 categories (MJHQ-full) and compare it with vanilla CFG at its best w. PAG (Ahn et al., 2025a), a guidance method that is likewise designed inside the architecture and can be stacked on CFG, is included as an additional baseline on Muddit, alone and on top of CFG, with the hyperparameters recommended in its paper. Tables 1 and 2 report the FID per category and on MJHQ-full; Appendix F describes and reports the full set of other scores.

Results. On Muddit, HeadTuned-CFG lowers the FID on MJHQ-full from 20.90 to 16.29 and the FID of every category (Table 1); on Janus it lowers the FID on MJHQ-full from 10.51 to 9.92 and the FID of 9 of the 10 categories (Table 2). PAG alone improves over a single forward pass on Muddit but falls short of CFG, and stacking it on CFG degrades CFG (Table 1). The full metrics in Appendix F show that the lower FID is not obtained at the expense of the other scores. In short, HeadTuned-CFG outperforms the related guidance methods on the T2I generation of two models of different paradigms and architectures.

Table 1: Muddit T2I: FID per MJHQ category ($n = 3{,}000$ each) and on MJHQ-full ($n = 30{,}000$); CFG ($w = 3$), PAG ($\lambda = 2.5$, single-stream blocks 9–10), PAG+CFG ($\lambda = 1.5$ on CFG with $w = 3$), and HeadTuned-CFG ($w = 9$, $k = 50\%$, $s = 0.4$).

	animals	art	fashion	food	indoor	landscape	logo	people	plants	vehicles	full
CFG	33.91	42.82	38.79	39.53	43.82	40.93	40.27	43.85	49.93	36.89	20.90
PAG	52.31	57.31	53.30	64.44	61.24	65.27	51.48	60.09	76.86	54.04	34.37
PAG+CFG	38.21	44.40	40.03	42.27	43.74	45.92	39.46	45.28	56.44	38.59	22.39
HeadTuned-CFG	**30.41**	**36.70**	**34.02**	**33.60**	**36.17**	**34.56**	**36.65**	**36.88**	**43.01**	**30.93**	**16.29**

Table 2: Janus T2I: FID per MJHQ category ($n = 3{,}000$ each) and on MJHQ-full ($n = 30{,}000$). CFG ($w = 4$) versus HeadTuned-CFG ($w = 4$, $k = 15\%$, $s = 0.75$); PAG is not applicable to Janus.

	animals	art	fashion	food	indoor	landscape	logo	people	plants	vehicles	full
CFG	27.32	31.33	30.10	**31.05**	28.06	27.53	37.90	33.97	30.20	23.66	10.51
HeadTuned-CFG	**25.95**	**29.93**	**28.83**	31.37	**27.03**	**24.81**	**37.60**	**32.02**	**29.37**	**23.16**	**9.92**

5.2 Ablation Study on T2I Head Selection

To test whether the gain comes from the probed ranking, on Muddit we keep all parameters $w = 9$, $k = 50\%$, $s = 0.4$, while only replacing the probe-based selection by a late-layer prior (all heads of the last 14 single-stream blocks, where most of the probed bottom-50% heads lie), a random set of the same size, and the reverse selection that tunes the top-50% heads, i.e., the condition-change aware heads (Figure 4). On MJHQ-animals, the FID ranks probe-based (30.41) $<$ late-layer prior (30.71) $<$ random (39.31) $<$ reverse (41.27), against 33.91 for vanilla CFG: the probe-based selection is better than the layer prior, whereas random and reverse selection are both worse than no tuning, which confirms that the probed condition-change agnostic heads are the right heads to tune. The same ablation on Janus gives the same ordering (Appendix G).

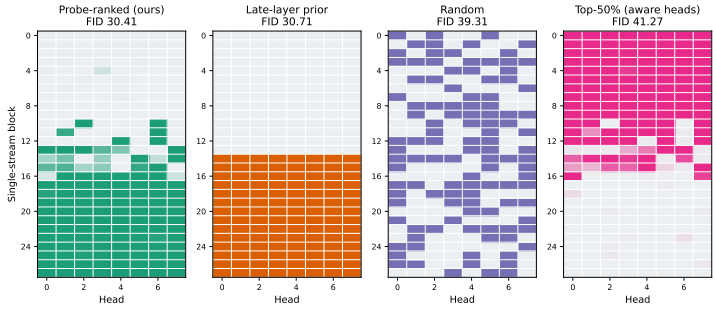

Figure 4: Heads tuned by the four selection rules on Muddit (single-stream block by head), with FID on MJHQ-animals at $w = 9$, $k = 50\%$, $s = 0.4$. The probed and the top-50% selections change per denoising step; their panels show the fraction of the 64 steps in which a head is tuned.

5.3 Image-to-Text Generation

Setup. We evaluate object hallucination of Janus on POPE (Li et al., 2023) and on the hallucination subset of MME (Fu et al., 2025). POPE asks yes/no questions about the presence of objects in MS-COCO images on three splits (*random*, *popular*, and *adversarial*; 3,000 questions each) and reports accuracy and F1; the hallucination subset of MME covers *existence*, *count*, *position*, and *color*, scored as in VCD (Leng et al., 2023) out of 200 per subtask. We follow the protocol and the decoding settings of VCD (Appendix H.2); the head ranking is the I2T probe of Section 3.2, obtained on the captioning task. *Regular* denotes decoding without guidance, *VCD* the decoding above, and *HeadTuned-VCD* adds head tuning to VCD in the extrapolation mode (top-k heads, $s > 1$) to further strengthen the output of the condition-change aware heads. As for T2I, we sweep w for VCD and grid-search (w, k, s) for HeadTuned-VCD and report the best setting of each (Appendix H.2); on POPE the search uses the random split, and the selected setting is applied unchanged to the other two splits.

Table 3: Janus on POPE, accuracy and F1 (%) per split ($n = 3{,}000$ each): Regular decoding, VCD ($w = 8$, $\beta = 0.2$, $T = 500$), and HeadTuned-VCD ($w = 8$, $k = 20$ heads, $s = 1.3$).

	random		popular		adversarial		overall	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Regular	84.17	81.60	82.97	80.47	81.33	78.99	82.82	80.35
VCD	87.77	86.33	85.97	84.63	83.93	82.79	85.89	84.58
HeadTuned-VCD	**88.80**	**87.76**	**86.43**	**85.54**	**84.13**	**83.50**	**86.45**	**85.60**

Table 4: Janus on the hallucination subset of MME, score (accuracy + image-level accuracy, out of 200) per subtask and in total: Regular decoding, VCD ($w = 10$, $\beta = 0.2$, $T = 500$), and HeadTuned-VCD ($w = 8$, $k = 15$ heads, $s = 1.1$).

	existence	count	position	color	total
Regular	138.33	56.67	90.00	93.33	378.33
VCD	**170.00**	81.67	**91.67**	123.33	466.67
HeadTuned-VCD	**170.00**	**83.33**	**91.67**	**130.00**	**475.00**

Results. On POPE, HeadTuned-VCD raises the accuracy of the random split from 87.77 to 88.80 and the F1 from 86.33 to 87.76 over VCD, and improves on every split in both metrics when applied unchanged to the other splits (Table 3). On the hallucination subset of MME, VCD raises the total score from 378.33 to 466.67, with the largest gains on existence and color; HeadTuned-VCD further raises it to 475.00, matching VCD on existence and position and improving count and color (Table 4).

Limitation. We also applied HeadTuned-VCD to Muddit on the same benchmarks, but it did not improve over VCD. Several factors contribute: Muddit itself performs poorly on POPE and MME, and VCD brings only a marginal gain on it. Moreover, as discussed in Section 3.2, its condition-change awareness ranking in I2T generation is unstable across samples in the late denoising steps, which makes the selected heads unreliable across samples. We discuss this limitation in I2T generation here, as it may motivate further study in the future.

6 Conclusion and Future Work

In this work, we investigate the shared backbone of unified VLMs through the lens of conditional guidance. Probing the intermediate activations of attention heads across the two branches of CFG reveals that the heads are functionally specialized: a subset of them is aware of the change of condition, while some are agnostic to it. This specialization appears consistently in both T2I and I2T generation and is largely stable across samples, suggesting that it is an intrinsic property of unified VLMs. Building on this specialization, HeadTuned-CFG uses the probed condition-change awareness to select heads and tune the guidance at the head level, and brings consistent benefits to both generation directions of two unified VLMs of different paradigms, including substantial improvements over vanilla CFG and related guidance methods in T2I generation. These results in turn support the probing analysis, and establish conditional guidance as a new perspective for understanding unified VLMs.

Two directions follow from this work. First, HeadTuned-CFG currently selects its hyperparameters (w, k, s) by grid search on a validation subset; a more efficient tuning, e.g., setting the head-wise strength from the probed condition-change awareness itself as suggested by the analysis in Appendix D, would make the method less costly to apply to new models and tasks. Second, the head-level specialization revealed by our probing is a property of the pretrained backbone, which raises an intriguing question of how it emerges. Understanding this could inform the architecture design and the training paradigm of unified VLMs, for instance by encouraging or exploiting such specialization when both T2I and I2T generation are trained jointly.

AI USE STATEMENT

In this work, we used generative AI tools for polishing writing, finding related work, research ideation, and experiment execution. We have not used generative AI tools for generating synthetic datasets for this work. We have reviewed all AI-assisted work. We checked LLM-generated research ideas for potential plagiarism through a manual literature survey, and LLM-generated code was verified and tested for correctness by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We describe all the experiment settings and implementation details in the paper and the supplementary materials. For the theoretical result, the assumptions and a complete proof are given in Appendix D. We will publish the code after acceptance.

REFERENCES

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, 2025a. URL https://arxiv.org/abs/2403.17377.

Donghoon Ahn, Jiwon Kang, Sanghyun Lee, Minjae Kim, Jaewon Min, Wooseok Jang, Sangwu Lee, Sayak Paul, Susung Hong, and Seungryong Kim. Where and how to perturb: On the design of perturbation guidance in diffusion and flow models, 2025b. URL https://arxiv.org/abs/2506.10978.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2021. URL https://arxiv.org/abs/2012.09841.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding, 2024. URL https://arxiv.org/abs/2403.14003.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2025. URL https://arxiv.org/abs/2306.13394.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment, 2023. URL https://arxiv.org/abs/2310.11513.

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, 2018. URL https://arxiv.org/abs/1706.08500.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention, 2024. URL https://arxiv.org/abs/2408.00760.

Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance, 2023. URL https://arxiv.org/abs/2210.00939.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens, 2025. URL https://arxiv.org/abs/2411.16724.

Jinyeong Kim, Seil Kang, Jiwoo Park, Junhyeok Kim, and Seong Jae Hwang. Interpreting attention heads for image-to-text information flow in large vision-language models, 2025. URL https://arxiv.org/abs/2509.17588.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models, 2019. URL https://arxiv.org/abs/1904.06991.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models, 2024. URL https://arxiv.org/abs/2404.07724.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding, 2023. URL https://arxiv.org/abs/2311.16922.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation, 2024. URL https://arxiv.org/abs/2402.17245.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023. URL https://arxiv.org/abs/2305.10355.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollr. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models, 2020. URL https://arxiv.org/abs/2002.09797.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models, 2025. URL https://arxiv.org/abs/2410.07149.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads, 2022. URL https://arxiv.org/abs/2209.11895.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016. URL https://arxiv.org/abs/1606.03498.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL https://arxiv.org/abs/2210.08402.

Qingyu Shi, Jinbin Bai, Zhuoran Zhao, Wenhao Chai, Kaidong Yu, Jianzong Wu, Yunhai Tong, Xiangtai Li, Xuelong Li, and Shuicheng Yan. Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model, 2026. URL https://arxiv.org/abs/2505.23606.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2025. URL https://arxiv.org/abs/2405.09818.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small, 2022. URL https://arxiv.org/abs/2211.00593.

Xi Wang, Nicolas Dufour, Nefeli Andreou, Marie-Paule Cani, Victoria Fernandez Abrevaya, David Picard, and Vicky Kalogeiton. Analysis of classifier-free guidance weight schedulers, 2024a. URL https://arxiv.org/abs/2404.13040.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need, 2024b. URL https://arxiv.org/abs/2409.18869.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding, 2024c. URL https://arxiv.org/abs/2403.18715.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation, 2024a. URL https://arxiv.org/abs/2410.13848.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality, 2024b. URL https://arxiv.org/abs/2404.15574.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis, 2023. URL https://arxiv.org/abs/2306.09341.

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, 2023. URL https://arxiv.org/abs/2304.05977.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models, 2025. URL https://arxiv.org/abs/2505.15809.

A Visualization of Condition-change Aware and Agnostic Heads on Muddit

Figure 5 repeats Figure 1 on Muddit with the same prompt and the heads selected by the Muddit probes. The T2I maps show where the CFG difference of the head aligns with the keyword in value space, and the I2T maps show the attention of the decoded keyword to the image, as for Janus.

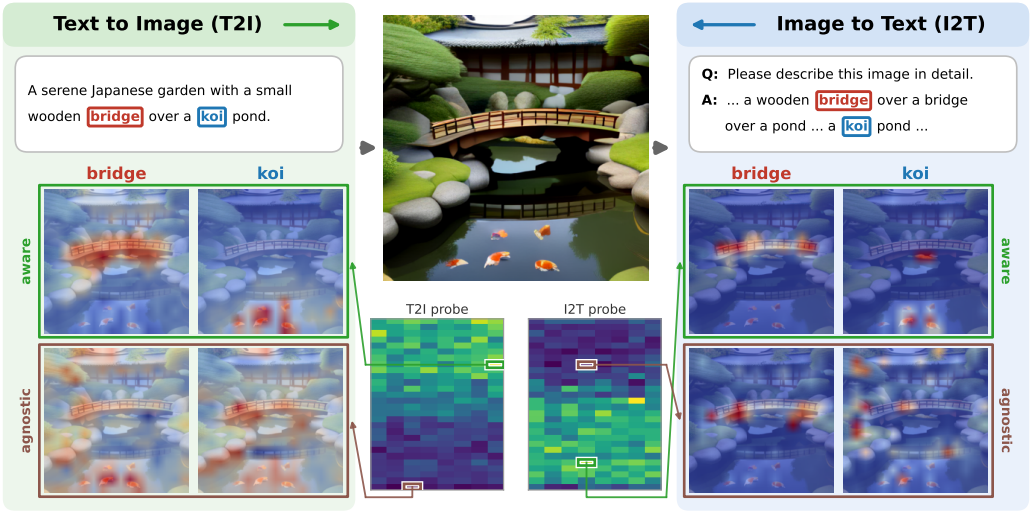

Figure 5: Condition-change aware (green) and agnostic (brown) attention heads of Muddit, in the layout of Figure 1: the T2I maps show the alignment of each head's CFG difference with the keyword in value space, the I2T maps the attention of the decoded keyword to the image, and the middle column the generated image and the Muddit probes with the visualized heads marked.

B Implementation Details of Probing

Probing procedure. Probing runs standard CFG generation with the positive and negative branches in one batch and hooks the activations of every head without modifying them. At a hooked head (l, h) and step t, the score $\delta^{(l,h)}(p, t)$ in Algorithm 1 is the ℓ_2 norm of $\mathbf{h}_{\text{pos}}^{(l,h),t} - \mathbf{h}_{\text{neg}}^{(l,h),t}$ over the head dimension d, computed at the generation positions and averaged over them. In Muddit the hook reads the per-head value vectors $\mathbf{V}_{l,h}$, and in Janus the per-head attention outputs $\mathbf{O}_{l,h}$ (Section 2.3). The prompt set $\mathcal{P}$ contains $|\mathcal{P}| = 500$ samples, and the ranking used for selection averages $\delta^{(l,h)}$ over $\mathcal{P}$. The construction of the two branches, the probed layers and positions, and the generation parameters are given per task and model below.

Muddit. Muddit uses an MMDiT backbone with 14 dual-stream and 28 single-stream blocks of 8 heads each; every denoising step processes the full sequence in one forward pass. The sequence concatenates the tokens of the two modalities; we call the positions of the conditioning modality the *condition positions* and those of the generated modality the *generation positions*. The 28 single-stream blocks are probed, and the ranking is over these 224 heads. Scores are kept per denoising step, giving $\delta^{(l,h)}(p, t)$ of shape $|\mathcal{T}| \times 28 \times 8$ per sample. Generation runs at 512×512 resolution.

Janus. Janus-1.3B uses a LLaMA-style decoder with 24 layers of 16 heads; the condition is encoded in a prefill pass, after which the output is decoded autoregressively with a KV cache. The condition positions are the prefill tokens and the generation positions the decoded tokens. All 24 layers are probed at the decode steps; with the KV cache, each decode step computes attention outputs only for the newly generated position, which is therefore the probed position. Scores are averaged over the decode steps, giving $\delta^{(l,h)}(p)$ of shape 24×16 per sample. Image generation runs at 384×384 resolution, i.e., 576 image tokens decoded in 576 steps.

B.1 Probing on T2I Generation

$\mathcal{P}$ contains 500 GenEval prompts for both models. In Muddit, the negative branch replaces the prompt with a negative prompt, and generation uses $|\mathcal{T}| = 64$ and $w = 9$. In Janus, the negative branch replaces the text tokens with padding tokens, and generation uses $w = 5$.

B.2 Probing on I2T Generation

$\mathcal{P}$ contains 500 images of MS-COCO, and the prompt asks for a detailed description of the image. The negative branch replaces the image with a noised copy as in VCD ($T = 500$), and generation uses $w = 2$ with greedy decoding.

C Cross-sample Stability of the Condition-change Awareness Ranking

Figure 6 plots the mean pairwise Spearman rank correlation of the per-sample head rankings at each denoising step for Muddit ($|\mathcal{P}| = 500$ samples per task); the I2T ranking is probed as in Appendix B.2 with $|\mathcal{T}| = 32$ and $w = 1.25$. Table 5 reports the same statistic for Janus, whose samples each yield one ranking averaged over decode positions. Both rankings of Janus and the T2I ranking of Muddit are stable across samples, whereas the I2T ranking of Muddit degrades over the denoising steps, from 0.97 at the first step to 0.67 at the last.

Table 5: Mean pairwise Spearman rank correlation of the condition-change awareness ranking of Janus across $|\mathcal{P}| = 500$ probing samples.

	T2I	I2T
Spearman ρ	0.970 ± 0.014	0.962 ± 0.023

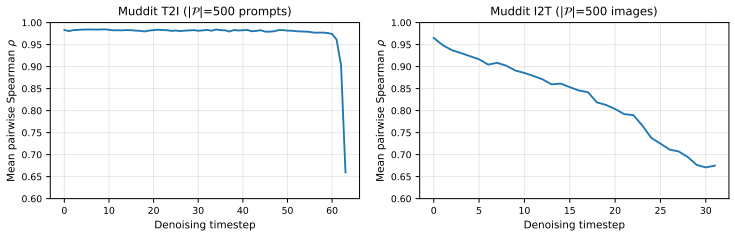

Figure 6: Cross-sample stability of the condition-change awareness ranking in Muddit, per denoising step. Left: T2I generation. Right: I2T generation.

D Theoretical Justification of HeadTuned-CFG

This section proves Theorem 1 and gives a theoretical justification of the head-wise update in HeadTuned-CFG. We show that the tuned head activation is the MAP estimate of the distribution obtained by a KL projection from the negative branch under a positive-awareness constraint. This view also explains the two modes of Section 4: the probed ranking separates condition-change aware heads from condition-change agnostic heads, and the two groups receive different positive-awareness targets, so that the former amplify condition-specific information ($s > 1$) while the latter retain only a fraction of their uninformative response and stay close to the negative branch ($s < 1$).

Consider one probed head (l, h) with positive- and negative-branch activations $\mathbf{h}_{\text{pos}}^{(l,h)}$ and $\mathbf{h}_{\text{neg}}^{(l,h)}$. HeadTuned-CFG modifies the positive branch by the tuning operator

$$\Phi_s\left(\mathbf{h}_{\text{pos}}^{(l,h)}, \mathbf{h}_{\text{neg}}^{(l,h)}\right) = \mathbf{h}_{\text{neg}}^{(l,h)} + s\left(\mathbf{h}_{\text{pos}}^{(l,h)} - \mathbf{h}_{\text{neg}}^{(l,h)}\right).$$

Head-wise CFG model. For head (l, h), suppose the activations of the two branches locally follow two Gaussian distributions with a shared covariance,

$$\mathbf{x}_{\text{pos}}^{(l,h)} \sim p_{\text{pos}}^{(l,h)} := \mathcal{N}\Big(\boldsymbol{\mu}_{\text{pos}}^{(l,h)}, \Sigma_{l,h}\Big), \qquad \mathbf{x}_{\text{neg}}^{(l,h)} \sim p_{\text{neg}}^{(l,h)} := \mathcal{N}\Big(\boldsymbol{\mu}_{\text{neg}}^{(l,h)}, \Sigma_{l,h}\Big).$$

Linear discriminant analysis (LDA) gives the discriminant direction

$$\mathbf{w}_{l,h} = \Sigma_{l,h}^{-1}\left(\boldsymbol{\mu}_{\text{pos}}^{(l,h)} - \boldsymbol{\mu}_{\text{neg}}^{(l,h)}\right),$$

and we call $T_{l,h}(\mathbf{x}) = \mathbf{w}_{l,h}^{\top}\mathbf{x}$ the *positive-awareness score* of an activation $\mathbf{x}$ at this head.

We then seek the distribution $q_{l,h}$ closest to the negative-branch distribution $p_{\text{neg}}^{(l,h)}$ that attains a target positive-awareness level $\tau_{l,h}$:

$$\min_{q_{l,h}} \mathrm{KL}\Big(q_{l,h} \,\|\, p_{\text{neg}}^{(l,h)}\Big) \quad \text{s.t.} \quad \mathbb{E}_{q_{l,h}}[T_{l,h}(\mathbf{x})] \geq \tau_{l,h}. \tag{9}$$

The Lagrangian is

$$\mathcal{L}(q_{l,h}, s_{l,h}) = \mathrm{KL}\Big(q_{l,h} \,\|\, p_{\text{neg}}^{(l,h)}\Big) + s_{l,h}\left(\tau_{l,h} - \mathbb{E}_{q_{l,h}}[T_{l,h}(\mathbf{x})]\right),$$

with multiplier $s_{l,h} \geq 0$. For fixed $s_{l,h}$, minimizing over $q_{l,h}$ gives the exponential tilt

$$q_{l,h,s_{l,h}}^{*}(\mathbf{x}) \propto p_{\text{neg}}^{(l,h)}(\mathbf{x}) \exp(s_{l,h} T_{l,h}(\mathbf{x})).$$

Substituting it back yields the dual objective

$$g_{l,h}(s_{l,h}) = s_{l,h}\tau_{l,h} - \log Z_{l,h}(s_{l,h}),$$

where

$$Z_{l,h}(s_{l,h}) = \mathbb{E}_{p_{\text{neg}}^{(l,h)}}[\exp(s_{l,h} T_{l,h}(\mathbf{x}))] = \exp\left(s_{l,h}\, \mathbf{w}_{l,h}^{\top}\boldsymbol{\mu}_{\text{neg}}^{(l,h)} + \frac{s_{l,h}^2}{2}\, \mathbf{w}_{l,h}^{\top}\Sigma_{l,h}\mathbf{w}_{l,h}\right). \tag{10}$$

The last equality holds because, under $p_{\text{neg}}^{(l,h)}$, $T_{l,h}(\mathbf{x}) = \mathbf{w}_{l,h}^{\top}\mathbf{x}$ is a one-dimensional Gaussian with moment-generating function $\mathbb{E}[\exp(sT)] = \exp\big(s\,\mathbb{E}[T] + s^2\mathrm{Var}(T)/2\big)$. The optimal multiplier solves the one-dimensional dual problem $s_{l,h}^{*} = \arg\max_{s_{l,h}\geq 0} g_{l,h}(s_{l,h})$; with Eq. 10,

$$g_{l,h}(s_{l,h}) = s_{l,h}\left(\tau_{l,h} - \mathbf{w}_{l,h}^{\top}\boldsymbol{\mu}_{\text{neg}}^{(l,h)}\right) - \frac{s_{l,h}^2}{2}\, \mathbf{w}_{l,h}^{\top}\Sigma_{l,h}\mathbf{w}_{l,h},$$

and maximizing this concave quadratic gives

$$s_{l,h}^{*} = \left[\frac{\tau_{l,h} - \mathbf{w}_{l,h}^{\top}\boldsymbol{\mu}_{\text{neg}}^{(l,h)}}{\mathbf{w}_{l,h}^{\top}\Sigma_{l,h}\mathbf{w}_{l,h}}\right]_{+} = \left[\frac{\tau_{l,h} - \left(\boldsymbol{\mu}_{\text{pos}}^{(l,h)} - \boldsymbol{\mu}_{\text{neg}}^{(l,h)}\right)^{\top}\Sigma_{l,h}^{-1}\boldsymbol{\mu}_{\text{neg}}^{(l,h)}}{\left(\boldsymbol{\mu}_{\text{pos}}^{(l,h)} - \boldsymbol{\mu}_{\text{neg}}^{(l,h)}\right)^{\top}\Sigma_{l,h}^{-1}\left(\boldsymbol{\mu}_{\text{pos}}^{(l,h)} - \boldsymbol{\mu}_{\text{neg}}^{(l,h)}\right)}\right]_{+}. \tag{11}$$

Therefore the solution of the primal problem in Eq. 9 is

$$q_{l,h,s_{l,h}^{*}}^{*} = \mathcal{N}\Big(\boldsymbol{\mu}_{\text{neg}}^{(l,h)} + s_{l,h}^{*}\,\Sigma_{l,h}\mathbf{w}_{l,h},\ \Sigma_{l,h}\Big),$$

whose MAP estimate is

$$\mathbf{h}_{g}^{(l,h)} = \boldsymbol{\mu}_{\text{neg}}^{(l,h)} + s_{l,h}^{*}\,\Sigma_{l,h}\mathbf{w}_{l,h} = \boldsymbol{\mu}_{\text{neg}}^{(l,h)} + s_{l,h}^{*}\left(\boldsymbol{\mu}_{\text{pos}}^{(l,h)} - \boldsymbol{\mu}_{\text{neg}}^{(l,h)}\right).$$

Estimating the class means by the observed head activations, $\boldsymbol{\mu}_{\text{pos}}^{(l,h)} \approx \mathbf{h}_{\text{pos}}^{(l,h)}$ and $\boldsymbol{\mu}_{\text{neg}}^{(l,h)} \approx \mathbf{h}_{\text{neg}}^{(l,h)}$, yields

$$\mathbf{h}_{g}^{(l,h)} = \mathbf{h}_{\text{neg}}^{(l,h)} + s_{l,h}^{*}\left(\mathbf{h}_{\text{pos}}^{(l,h)} - \mathbf{h}_{\text{neg}}^{(l,h)}\right),$$

which is exactly the operator Φ_s with the head-wise strength $s = s_{l,h}^{*}$ of Eq. 11. In practice HeadTuned-CFG applies one shared s to all selected heads (Section 4).

Extrapolation and interpolation. The vanilla positive branch is Φ_s with $s = 1$, i.e., $q^*_{l,h,1} = p^{(l,h)}_{\text{pos}}$, whose expected positive-awareness score is

$$\mathbb{E}_{p^{(l,h)}_{\text{pos}}}[T_{l,h}(\mathbf{x})] = \mathbf{w}^\top_{l,h}\boldsymbol{\mu}^{(l,h)}_{\text{neg}} + \mathbf{w}^\top_{l,h}\Sigma_{l,h}\mathbf{w}_{l,h}.$$

By Eq. 11, extrapolation ($s^*_{l,h} > 1$) is equivalent to $\tau_{l,h} > \mathbb{E}_{p^{(l,h)}_{\text{pos}}}[T_{l,h}(\mathbf{x})]$, and interpolation ($s^*_{l,h} < 1$) to $\tau_{l,h} < \mathbb{E}_{p^{(l,h)}_{\text{pos}}}[T_{l,h}(\mathbf{x})]$: the target lies above or below the positive-awareness level of the unchanged positive branch.

HeadTuned-CFG ranks heads by the condition-change awareness score $\delta^{(l,h)} = \|\mathbf{h}^{(l,h)}_{\text{pos}} - \mathbf{h}^{(l,h)}_{\text{neg}}\|_2$ to separate condition-change aware heads from condition-change agnostic heads, and the ranking guides the assignment of the head-wise targets $\tau_{l,h}$. Heads with large $\delta^{(l,h)}$ are condition-change aware; assigning them targets above their unchanged positive-branch level gives $s^*_{l,h} > 1$, the top-k extrapolation mode, which amplifies condition-specific information. Heads with small $\delta^{(l,h)}$ are condition-change agnostic; assigning them lower targets gives $s^*_{l,h} < 1$, the bottom-k interpolation mode, which suppresses their uninformative response and keeps them close to the negative branch.

E Implementation Details of HeadTuned-CFG

Tuning operator. For a head $(l, h) \in \mathcal{S}_k$, Φ_s acts on the per-head slice of an activation tensor and is the linear map

$$\Phi_s\left(\mathbf{h}^{(l,h)}_{\text{pos}}, \mathbf{h}^{(l,h)}_{\text{neg}}\right) = \mathbf{h}^{(l,h)}_{\text{neg}} + s\left(\mathbf{h}^{(l,h)}_{\text{pos}} - \mathbf{h}^{(l,h)}_{\text{neg}}\right). \tag{12}$$

The two branches are run in one batch, so $\mathbf{h}^{(l,h)}_{\text{neg}}$ is available at the same layer. Φ_s is implemented as a forward hook that replaces $\mathbf{h}^{(l,h)}_{\text{pos}}$ in place; the negative branch and all downstream computation, including the CFG combination in Eq. 3, are unchanged. Φ_s acts on the activation probed in Appendix B, $\mathbf{V}_{l,h}$ for Muddit and $\mathbf{O}_{l,h}$ for Janus; the tuned layers, the token positions it covers, and the tuning steps $\mathcal{T}'$ are chosen per architecture as follows.

Muddit. The 28 single-stream blocks are tuned, restricted to the generation positions; the condition positions and the dual-stream blocks are left untouched. Since every denoising step recomputes the full sequence, the tuning steps are $\mathcal{T}' = \mathcal{T}$, and $\mathcal{S}_k$ is re-selected at each step from the per-step scores.

Janus. All 24 layers are tuned. The tuning steps are $\mathcal{T}' = \{\text{prefill}\}$: Φ_s is applied only in the prefill pass and hence only at the condition positions. The tuned outputs change the hidden states of the condition positions, whose keys and values in the following layers are stored in the KV cache and attended to by all decode steps, which run unmodified.

F Full Metrics on MJHQ-full

This appendix reports the scores omitted from Tables 1 and 2 on MJHQ-full ($n = 30{,}000$), computed against all 30,000 real images: CLIP score (Radford et al., 2021) (text–image alignment), ImageReward (Xu et al., 2023) and HPSv2 (Wu et al., 2023) (learned human preference), precision/recall (Kynkäänniemi et al., 2019) and density/coverage (Naeem et al., 2020) (fidelity and diversity of the generated distribution), Inception Score (Salimans et al., 2016), and the aesthetic score of the LAION aesthetic predictor (Schuhmann et al., 2022).

F.1 Muddit

Table 6 lists the scores of vanilla CFG, PAG, PAG+CFG, and HeadTuned-CFG on Muddit; HeadTuned-CFG is the best on every score except the aesthetic score, where it is within 0.07 of the best.

F.2 Janus

Table 7 lists the scores of vanilla CFG and HeadTuned-CFG on Janus: HeadTuned-CFG lowers the FID while the other scores remain comparable. To show that the lower FID is not obtained at the

Table 6: Muddit on MJHQ-full, overall ($n = 30{,}000$): CFG ($w = 3$), PAG ($\lambda = 2.5$ on single-stream blocks 9–10), PAG+CFG ($\lambda = 1.5$ on top of CFG with $w = 3$), and HeadTuned-CFG ($w = 9$, $k = 50\%$, $s = 0.4$). IR is ImageReward; HPS is human preference score; Aes. is aesthetic score; P/R is precision/recall; D/C is density/coverage.

	FID ↓	CLIP ↑	IR ↑	HPS ↑	P/R ↑	D/C ↑	IS ↑	Aes. ↑
CFG	20.90	0.320	0.63	0.257	0.733/0.464	0.856/0.733	21.83	5.85
PAG	34.37	0.294	−0.26	0.217	0.597/0.382	0.577/0.508	16.45	5.66
PAG+CFG	22.39	0.318	0.64	0.264	0.720/0.431	0.851/0.712	20.77	**5.91**
HeadTuned-CFG	**16.29**	**0.324**	**0.82**	**0.271**	**0.762/0.529**	**0.969/0.808**	**22.77**	5.84

expense of the other scores, Figure 7 sweeps $w \in \{3, 4, 7.5\}$ on MJHQ-full for vanilla CFG and for HeadTuned-CFG with the pattern of Table 2 ($k = 15\%$, $s = 0.75$) fixed; the CLIP, ImageReward, HPS, and aesthetic panels also include HeadTuned-CFG at $w = 7.5$, $s = 0.8$. In every panel, the HeadTuned-CFG curve is Pareto-optimal over the vanilla CFG curve.

Table 7: Janus on MJHQ-full, overall ($n = 30{,}000$): CFG ($w = 4$) versus HeadTuned-CFG ($w = 4$, $k = 15\%$, $s = 0.75$); abbreviations as in Table 6.

	FID ↓	CLIP ↑	IR ↑	HPS ↑	P/R ↑	D/C ↑	IS ↑	Aes. ↑
CFG	10.51	**0.321**	**0.59**	**0.251**	0.660/**0.666**	0.660/0.797	**24.62**	**5.53**
HeadTuned-CFG	**9.92**	0.318	0.56	0.249	**0.682**/0.660	**0.702/0.811**	24.51	5.52

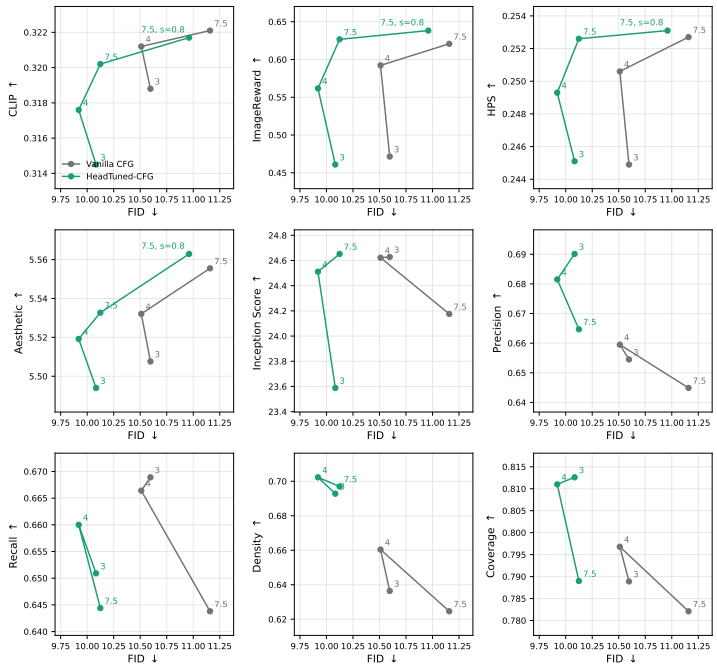

Figure 7: Janus on MJHQ-full ($n = 30{,}000$): FID against CLIP, ImageReward, HPS, aesthetic score, Inception Score, precision, recall, density, and coverage over the guidance scale w (labels), for vanilla CFG and HeadTuned-CFG ($k = 15\%$, $s = 0.75$ unless noted).

G Ablation on Head Selection

Figure 4 in Section 5.2 shows the head sets of the Muddit ablation; here we report the same ablation on Janus. On Janus we keep $w = 4$, $k = 15\%$, $s = 0.75$ and replace the probe-based selection by

an early-layer prior (the first layers, where most of the probed bottom-15% heads lie), a random set of the same size, and the reverse selection that tunes the top-15% heads (Figure 8). As on Muddit, the FID ranks probe-based (25.95) < early-layer prior (26.55) < random (27.12) < reverse (27.32), against 27.32 for vanilla CFG: the reverse selection is no better than no tuning.

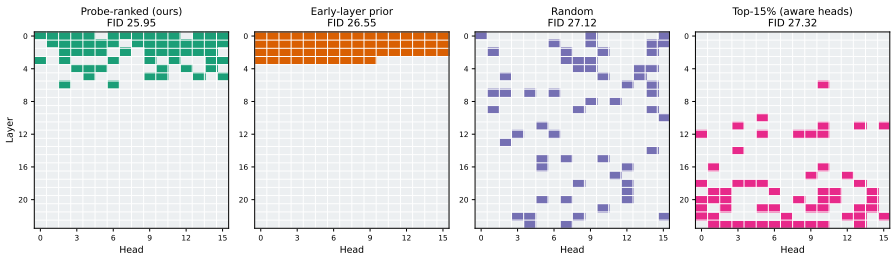

Figure 8: Heads tuned by the four selection rules on Janus (layer by head; colored = tuned), with FID on MJHQ-animals at $w = 4$, $k = 15\%$, $s = 0.75$.

H Grid Search of the Best Head Pattern under Fixed w

H.1 T2I Generation

Resolution and generation steps follow Appendix B.1; per-category scores are computed against the real MJHQ images of the same category.

H.1.1 Muddit

Table 8 sweeps vanilla CFG for T2I generation on MJHQ-animals; Tables 9 and 10 give the grid search of (k, s) at $w = 9$ and $w = 3$.

Table 8: Vanilla CFG on Muddit, MJHQ-animals ($n = 3{,}000$), over the guidance scale w.

w	1.0	1.5	2.0	2.5	3.0	5.0	7.0	9.0
FID	77.96	41.95	36.17	34.51	**33.91**	34.05	35.35	37.40

Table 9: HeadTuned-CFG search on Muddit, MJHQ-animals, $w = 9$: FID over the head fraction k (rows) and strength s (columns). Vanilla CFG at $w = 9$: FID 37.40.

$k \backslash s$	0.2	0.3	0.4	0.5	0.6
40%	31.60	31.33	31.91	–	–
45%	–	31.11	30.83	31.51	–
50%	–	30.61	**30.41**	31.21	–
55%	–	–	31.63	31.01	31.86
60%	–	–	31.77	31.76	32.25

Table 10: HeadTuned-CFG search on Muddit, MJHQ-animals, $w = 3$: FID over the head fraction k (rows) and strength s (columns). Vanilla CFG at $w = 3$: FID 33.91.

$k \backslash s$	0.6	0.7	0.8	0.9
15%	–	–	33.87	**33.61**
25%	–	34.09	33.70	33.80
35%	34.57	33.83	33.98	34.00

H.1.2 Janus

Table 11 sweeps vanilla CFG for T2I generation on MJHQ-animals; Tables 12 and 13 give the grid search of (k, s) at $w = 4$ and $w = 7.5$.

Table 11: Vanilla CFG on Janus, MJHQ-animals ($n = 3{,}000$), over the guidance scale w.

w	3.0	3.5	4.0	4.5	5.0	7.5
FID	27.45	28.07	**27.32**	27.57	27.92	28.87

Table 12: HeadTuned-CFG search on Janus, MJHQ-animals, $w = 4$: FID over the head fraction k (rows) and strength s (columns). Vanilla CFG at $w = 4$: FID 27.32.

$k\backslash s$	0.60	0.65	0.70	0.75	0.80
10%	26.51	26.57	26.34	27.09	–
15%	–	26.65	26.04	**25.95**	26.92
20%	–	28.12	27.08	26.32	27.19
25%	–	29.74	27.23	26.40	27.22
30%	–	–	28.88	26.95	27.35

Table 13: HeadTuned-CFG search on Janus, MJHQ-animals, $w = 7.5$: FID over the head fraction k (rows) and strength s (columns). Vanilla CFG at $w = 7.5$: FID 28.87.

$k\backslash s$	0.70	0.75	0.80
10%	27.11	26.76	28.01
15%	26.95	**26.48**	27.97
20%	27.13	26.53	28.21

H.2 I2T Generation

Protocol. Following VCD, the noised copy of the image is obtained with $T = 500$ forward-diffusion steps, and decoding uses the adaptive plausibility constraint with $\beta = 0.2$; each MME subtask has 30 images with two yes/no questions per image and is scored by accuracy plus image-level accuracy (both questions of an image correct), i.e., out of 200. Following the official implementation, each POPE question is appended with "Please answer this question with one word." (MME questions already end with a yes/no instruction), and answers are sampled with temperature 1 and a per-question seed shared by all decoding variants. Images are encoded at 384×384 resolution (576 image tokens), as in the I2T probe. For HeadTuned-VCD, k is given as a number of heads. The guidance scale of VCD is swept on each benchmark and fixed at its best value, $w = 8$ on POPE and $w = 10$ on MME; on POPE the head pattern (k, s) is searched on top of VCD at $w = 8$ on the random split, and on MME (w, k, s) is searched jointly over $w \in \{6, 8, 10\}$.

POPE. Table 14 sweeps the guidance scale w of VCD on the POPE random split for Janus; VCD saturates from $w = 8$ on, which we take as the baseline. Table 15 gives the search of the head pattern on top of VCD at $w = 8$.

Table 14: VCD on Janus, POPE random split ($n = 3{,}000$): accuracy and F1 (%) over the guidance scale $w = 1 + \alpha$ ($\beta = 0.2, T = 500$). Regular decoding: accuracy 84.17, F1 81.60.

w	2	4	6	8	10	12
Acc	87.10	87.60	87.70	**87.77**	**87.77**	**87.77**
F1	85.47	86.12	86.25	**86.33**	**86.33**	**86.33**

Table 15: HeadTuned-VCD search on Janus, POPE random split, on top of VCD at $w = 8$ (accuracy 87.77, F1 86.33): accuracy and F1 (%) over the number of tuned heads k (rows) and strength s (columns).

	Acc.				F1			
$k\backslash s$	1.2	1.3	1.4	1.5	1.2	1.3	1.4	1.5
15	88.47	88.70	88.60	88.73	87.28	87.63	87.57	87.76
20	88.63	**88.80**	88.73	88.70	87.50	**87.76**	87.74	**87.76**
25	88.53	88.53	88.27	88.00	87.40	87.46	87.22	87.01

MME. Table 16 sweeps w of VCD on the hallucination subset of MME: the score jumps from $w = 2$ to $w = 4$ and stays on a plateau from $w = 4$ on; the best score is reached at $w = 10$, with identical predictions at $w = 12$, so we take $w = 10$ as the baseline. Tables 17, 18, and 19 give the search of the head pattern on top of VCD at $w = 6$, 8, and 10. The score decreases from $s = 1.3$ on for every w and k; the best point is $w = 8$, $k = 15$, $s = 1.1$, where VCD itself is not yet saturated. Adjacent points differ by two to four questions, which is the resolution of the benchmark.

Table 16: VCD on Janus, hallucination subset of MME: score per subtask and in total over the guidance scale $w = 1 + \alpha$ ($\beta = 0.2$, $T = 500$). Regular decoding: 138.33 / 56.67 / 90.00 / 93.33, total 378.33.

w	2	4	6	8	10	12
existence	145.00	160.00	165.00	165.00	170.00	170.00
count	58.33	76.67	76.67	76.67	81.67	81.67
position	80.00	93.33	93.33	91.67	91.67	91.67
color	115.00	130.00	130.00	123.33	123.33	123.33
total	398.33	460.00	465.00	456.67	**466.67**	**466.67**

Table 17: HeadTuned-VCD search on Janus, hallucination subset of MME, on top of VCD at $w = 6$: total score over the number of tuned heads k (rows) and strength s (columns). VCD at $w = 6$: 465.00.

$k\backslash s$	1.05	1.10	1.15	1.20	1.25	1.30
10	461.67	466.67	458.33	460.00	468.33	445.00
15	465.00	471.67	460.00	470.00	465.00	458.33
20	468.33	463.33	466.67	468.33	461.67	455.00

Table 18: HeadTuned-VCD search on Janus, hallucination subset of MME, on top of VCD at $w = 8$: total score over the number of tuned heads k (rows) and strength s (columns). VCD at $w = 8$: 456.67.

$k\backslash s$	1.05	1.10	1.15	1.20	1.25	1.30
10	465.00	458.33	461.67	460.00	461.67	443.33
15	473.33	**475.00**	460.00	465.00	463.33	458.33
20	466.67	466.67	463.33	468.33	460.00	450.00

Table 19: HeadTuned-VCD search on Janus, hallucination subset of MME, on top of VCD at $w = 10$: total score over the number of tuned heads k (rows) and strength s (columns). VCD at $w = 10$: 466.67.

$k\backslash s$	1.05	1.10	1.15	1.20	1.25	1.30
10	470.00	463.33	461.67	460.00	461.67	443.33
15	468.33	470.00	465.00	465.00	458.33	448.33
20	466.67	466.67	458.33	463.33	460.00	445.00