TSGATE: TIMESTEP-AWARE GATED ATTENTION FOR DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion Transformers (DiTs) have emerged as the dominant architecture for high-fidelity image and video generation. Recent DiT systems increasingly use structured prompts for training, improving caption quality and prompt adherence. However, their generation quality can degrade severely under out-of-domain (OOD) prompts, including the free-form descriptions supplied by users at inference time. Although LLM-based rewriting can convert these prompts into structured formats, it does not guarantee that the rewritten prompts align with the training distribution. Our analysis links this degradation to attention sinks and reduced early-step image-to-text attention and shows that sink suppression alone is insufficient to restore generation quality. Despite effective sink suppression, models trained with standard gated attention exhibit reduced early-step image-to-text attention and suboptimal generation quality. Based on these insights, we propose Timestep-Aware Gated Attention (TSGate), which injects a timestep-conditioned bias into the gate signal so that gating behavior adapts across denoising steps. Extensive experiments show that TSGate consistently outperforms both the baseline and standard gated attention across multiple benchmarks, improving the raw-prompt DPG score by 9.5% over the baseline.

1 INTRODUCTION

Diffusion models have witnessed remarkable progress in visual content generation over the past few years (Ho et al., 2020; Rombach et al., 2022; Song et al., 2021). Among the most impactful architectural shifts is the adoption of Transformer backbones in place of convolutional U-Nets, giving rise to the Diffusion Transformer (DiT) family (Peebles & Xie, 2023; Bao et al., 2023). DiTs now underpin state-of-the-art systems for both image generation (e.g., SD3 (Esser et al., 2024), Flux (Black Forest Labs, 2024), and PixArt-α (Chen et al., 2024)) and video generation (e.g., Sora (Brooks et al., 2024), CogVideoX (Yang et al., 2025), and Wan (Team Wan et al., 2025)). A key enabler of their success is the joint attention paradigm introduced by MMDiT (Esser et al., 2024), which concatenates text and image tokens into a single sequence and lets them interact through shared self-attention layers. Concurrently, industrial practice (Lee et al., 2026; Happyhorse, 2026; Team Seedance et al., 2026) has converged on training with structured prompts, which are detailed, template-based captions organized under explicit section headings. This approach yields faster convergence and lower training loss and has become a widely adopted standard for improving text–image alignment during training.

Yet the impressive generation quality observed under in-domain structured prompts can obscure a fragile dependence on the specific prompt format. At inference time, real users write natural-language descriptions that are free-form, unstructured, and often terse. These descriptions deviate significantly from the training distribution. Although prompt engineering is standard practice (Team Seedance et al., 2026; MiniMax, 2026; Betker et al., 2023; Hao et al., 2023), it cannot fully eliminate the underlying distribution shift. We find an overall decline in quality across prompt conditions ranging from mild perturbations of the structured template to full OOD inputs, although the decline is not strictly monotonic across intermediate conditions. This is a critical practical bottleneck; no deployment-ready DiT can expect users to craft perfectly structured prompts.

To investigate this degradation, we systematically analyze attention dynamics within the DiT during denoising. Two related phenomena emerge. First, we observe that DiT attention maps exhibit *attention sinks*, characterized by a disproportionate concentration of attention mass on specific tokens,

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.

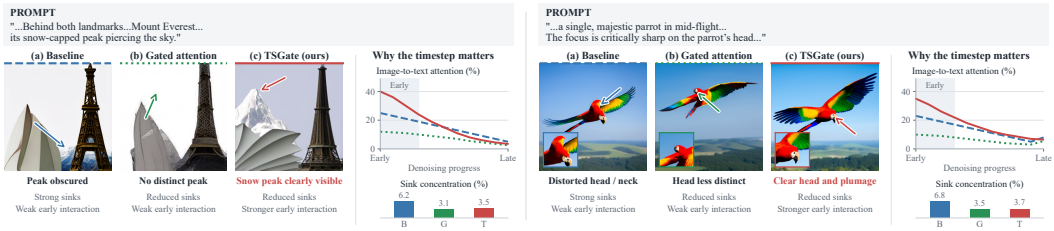


Figure 1: **Sink suppression is not enough; early text interaction matters.** For two out-of-domain prompts, TSGate renders the requested snow-capped peak and parrot's head more clearly than the baseline and standard gated attention. The schematic curves and bars illustrate the motivation for timestep-aware gating: reducing attention sinks must be accompanied by stronger text guidance during early denoising.

echoing discoveries in large language models (Xiao et al., 2024) and, more recently, in diffusion Transformers (Wu & Summa, 2026; Li et al., 2026). Sinks are exacerbated under OOD prompts, for which the absence of familiar header tokens disrupts established attention patterns. Second, OOD prompts induce significantly less image-to-text attention than in-domain inputs during early denoising, when noisy image tokens must acquire global semantic structure from the text condition. This early-step attention deficit limits how much textual guidance image tokens receive when global semantics are being established, potentially compromising prompt fidelity in the final image.

A natural remedy is gated attention, which has been shown to effectively eliminate attention sinks and improve robustness in large language models (Qiu et al., 2025). One would therefore expect it to address the sink problem in DiTs and consequently improve OOD generation quality. Despite substantially suppressing attention sinks, standard gated attention delivers suboptimal generation quality. Figure 1 illustrates this disconnect: reduced sinks coexist with weak early image-to-text interaction and unresolved errors in prompt-specified details. Sink suppression alone is therefore insufficient; the model must also preserve early text interaction. Although the input hidden states of the standard gate are timestep conditioned, the gate has no separately parameterized timestep offset. Because image tokens are dominated by noise in the early denoising phase, the content-dependent gate tends to suppress the image-to-text pathway. This finding exposes a fundamental mismatch: unlike autoregressive language models, diffusion models repeatedly process inputs whose statistics change dramatically with the noise level. Text–image interaction therefore requires a gate that is aware of the denoising stage.

Building on these insights, we propose **Timestep-Aware Gated Attention** (**TSGate**), a simple yet effective extension that explicitly conditions the gate on the current timestep. Concretely, we augment the standard gate with a timestep-conditioned bias. This design introduces only a small number of additional parameters per layer and is fully compatible with existing DiT architectures. Without explicit supervision of attention allocation, TSGate learns to maintain sink suppression while strengthening early image-to-text interaction, supporting semantic acquisition and more faithful rendering of prompt-specified details (Figure 1).

Our main contributions are as follows:

- We identify the mechanistic root causes of out-of-domain prompt degradation in diffusion transformers: the emergence of attention sinks and the collapse of early-step image-to-text attention. By demonstrating that naive sink suppression is fundamentally insufficient, we establish a compelling need for stage-aware attention mechanisms.
- We introduce TSGate, a lightweight, plug-and-play architectural enhancement that conditions attention gating on timestep information. This elegantly aligns gating behavior with the dynamic requirements of the denoising process while fully preserving token-specific content selection.
- Extensive experiments demonstrate that TSGate significantly enhances out-of-domain generalization and robustness, consistently outperforming standard baselines. Furthermore, we show that a first-layer-only TSGate variant achieves these substantial benefits with negligible parameter overhead.

2 Related Work

2.1 Diffusion Transformers

The transition from convolutional U-Net backbones (Ho et al., 2020; Rombach et al., 2022) to pure Transformer architectures marks a defining shift in diffusion model design. Peebles & Xie (2023) introduced DiT, replacing the U-Net denoiser with a Vision Transformer equipped with Adaptive Layer Normalization (AdaLN) for timestep and class conditioning. Around the same time, Bao et al. (2023) proposed U-ViT, which retains long skip connections within a ViT backbone. Subsequent work has rapidly scaled DiTs: PixArt-α (Chen et al., 2024) demonstrated efficient training through progressive resolution scaling, while Hunyuan-DiT (Li et al., 2024) and Lumina-T2X (Gao et al., 2024) pushed multilingual and multimodal capabilities. A pivotal architectural advance is the joint attention (or MMDiT) paradigm introduced by SD3 (Esser et al., 2024), in which text and image tokens are concatenated into a shared sequence for self-attention, superseding the earlier cross-attention design. Flux (Black Forest Labs, 2024) further streamlines this paradigm with flow-matching training. In the video domain, Sora (Brooks et al., 2024), CogVideoX (Yang et al., 2025), and Wan (Team Wan et al., 2025) extend DiTs to spatiotemporal modeling, while efficient variants such as DiT-MoE (Fei et al., 2024) explore mixture-of-experts scaling. In parallel, industrial practice has increasingly adopted structured, template-based captions for training (Lee et al., 2026; Betker et al., 2023), yet the implications of this practice for robustness, particularly vulnerability to OOD natural-language prompts at inference time, remain largely unexplored.

2.2 Attention Sinks in Transformers

The phenomenon of attention sinks, whereby a small number of tokens absorb a disproportionate share of attention mass regardless of their semantic relevance, was first systematically studied in autoregressive language models. Xiao et al. (2024) showed that the initial token in causal LLMs serves as an attention sink, and that preserving it is essential for stable streaming inference. In vision Transformers, Darcet et al. (2024) demonstrated that dedicating explicit register tokens eliminates attention artifacts and improves representation quality. More recently, attention sinks have been investigated in diffusion Transformers. Wu & Summa (2026) provided a causal analysis revealing that sink behavior in DiTs differs qualitatively from that in LLMs: it is coupled with the denoising timestep and varies across text versus image tokens. Li et al. (2026) discovered that certain text template tokens in the text encoder act as implicit semantic registers, absorbing excess attention mass within the joint attention framework. Su et al. (2026) offered a broader survey of sink phenomena across Transformer families, and Fu et al. (2026) proposed sink-aware training objectives for mitigating the issue. Despite these advances, existing studies have largely characterized sinks in isolation; they have not provided an in-depth analysis of the sink problem in DiTs, nor have they explored what practical issues the sink problem actually causes for DiTs in real-world usage. Our work fills this gap by connecting sink analysis with early-step attention allocation and demonstrating that effective solutions must be timestep-aware.

2.3 Gated Attention Mechanisms

Gating mechanisms have a long history in sequence modeling, from LSTM (Hochreiter & Schmidhuber, 1997) gates to modern attention variants. Qiu et al. (2025) recently conducted a comprehensive study of gated attention in large language models, showing that introducing a learned sigmoid gate on attention outputs simultaneously adds beneficial nonlinearity, promotes sparsity, and eliminates attention sinks, yielding consistent improvements in LLM pre-training and downstream performance. Several efficient-attention architectures incorporate related gating ideas: Gated Linear Attention (GLA) (Yang et al., 2024) integrates data-dependent gating into linear-complexity attention, RetNet (Sun et al., 2023) employs exponential decay as an implicit gate, and RWKV (Peng et al., 2023) uses channel-wise gating in a recurrent framework. In the diffusion domain, Zhu et al. (2025) applied GLA to build an efficient diffusion backbone, and Liu et al. (2025) analyzed attention from a temporal perspective, decomposing attention contributions across denoising stages. Zhao et al. (2025) proposed dynamic architectures that adapt computation per timestep, though without modifying the attention gating mechanism itself. Content-only output gating computes its logits from the current hidden representation, without a separate timestep offset in those logits. In DiTs, this representation can already be timestep conditioned; TSGate complements that implicit path with an explicit bias branch, aligning content selection with the evolving requirements of denoising.

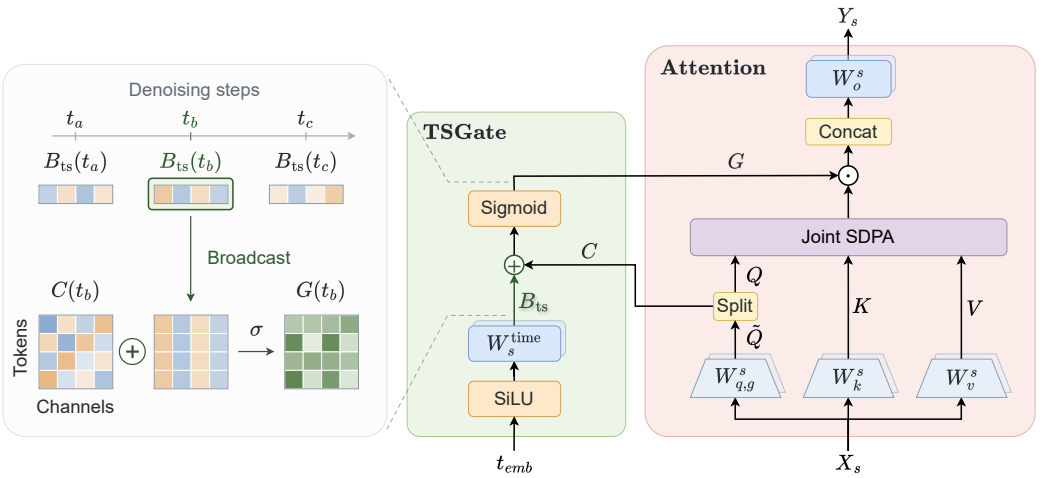


Figure 2: **Overview of TSGate.** At each denoising step t, a channel-wise time bias $B_{\text{ts}}(t)$ is shared across tokens and added to the content logits C; a sigmoid produces the gate G. The gate scales the attention features before the output projection.

3 METHOD

TSGate combines token-specific feature selection with explicit stage-dependent retention. We review content-only gating (standard gated attention (Qiu et al., 2025)), introduce the time branch, and explain how it complements the backbone's existing timestep conditioning.

3.1 PRELIMINARIES: GATED ATTENTION

Let $X \in \mathbb{R}^{N\times D}$ contain N token features, with H attention heads of width d_h and $D = Hd_h$. For per-head queries, keys, and values, attention computes $A_h = \text{softmax}(Q_h K_h^\top/\sqrt{d_h})V_h$ and concatenates the heads into $A \in \mathbb{R}^{N\times D}$. *Gated attention* scales these aggregated features before the output projection. A widened query projection jointly produces queries Q and content logits C:

$$(Q,\, C) = \text{split}(XW_{q,g} + b_{q,g})\,, \tag{1}$$

$$Y = (A \odot \sigma(\text{expand}(C)))\, W_o + b_o. \tag{2}$$

Here $W_{q,g} \in \mathbb{R}^{D\times(D+D_g)}$, $Q \in \mathbb{R}^{N\times D}$, and $C \in \mathbb{R}^{N\times D_g}$. For *element-wise* gating ($D_g = D$), expand is the identity; for *head-wise* gating ($D_g = H$), it repeats each head's logit across its d_h channels. We use σ for the element-wise sigmoid and $\odot$ for element-wise multiplication. We call this gate *content-only*: its logits are computed from X, although X itself can already depend on the timestep through the DiT backbone (Peebles & Xie, 2023).

3.2 TIMESTEP-AWARE GATED ATTENTION

We describe the default element-wise form for one sample and layer, omitting batch and layer indices. At timestep t, stream s has input $X_s(t) \in \mathbb{R}^{N_s\times D}$ and aggregated attention features $A_s(t) \in \mathbb{R}^{N_s\times D}$. Following MMDiT (Esser et al., 2024), queries from each stream attend jointly to keys and values from all streams. We represent features as row vectors and use separate projection parameters for each layer and stream.

Content branch. As in content-only gating, the content branch in Figure 2 uses a widened query projection to produce token-dependent logits:

$$(Q_s(t), C_s(t)) = \text{split}\big(X_s(t)W_{q,g}^s + b_{q,g}^s\big)\,, \tag{3}$$

where $W_{q,g}^s \in \mathbb{R}^{D\times 2D}$ and $Q_s(t), C_s(t) \in \mathbb{R}^{N_s\times D}$. The split separates queries and logits within each head. Queries and keys follow the backbone's RMSNorm (Zhang & Sennrich, 2019) and



Figure 3: **Paired generations from in-domain structured prompts and OOD prompts.** Across three representative cases, departing from the training-time template substantially degrades object fidelity, compositionality, and scene coherence.

RoPE (Su et al., 2024) operations; content logits do not. Appendix A.2 gives the tensor layouts and attention-path details.

Time branch. As shown in Figure 2, a separate projection maps the backbone timestep embedding $t_{\text{emb}}(t) \in \mathbb{R}^{1\times d_t}$ to a channel-wise offset:

$$B_{\text{ts},s}(t) = \text{SiLU}(t_{\text{emb}}(t))\, W_s^{\text{time}} + b_s^{\text{time}}, \tag{4}$$

Here SiLU (Elfwing et al., 2018) is the activation function, $W_s^{\text{time}} \in \mathbb{R}^{d_t\times D}$ and $b_s^{\text{time}} \in \mathbb{R}^{1\times D}$. Unlike $C_s(t)$, the bias $B_{\text{ts},s}(t)$ depends on the timestep embedding rather than on individual token features and is shared across the stream's tokens. Appendix A.6 explains why this separate, noise-free branch is useful when the content logits are dominated by noise at high timesteps.

We add the time bias to the content logits *before* the sigmoid, then gate the aggregated attention features:

$$G_s(t) = \sigma(C_s(t) + \text{broadcast}(B_{\text{ts},s}(t)))\,, \tag{5}$$
$$Y_s(t) = (G_s(t) \odot A_s(t))\, W_o^s + b_o^s. \tag{6}$$

Here broadcast repeats the time bias over N_s tokens, so $G_s(t), Y_s(t) \in \mathbb{R}^{N_s\times D}$ and $W_o^s \in \mathbb{R}^{D\times D}$. For fixed content logits, increasing a channel's time bias raises its retention while preserving the ordering of gates across tokens; stage modulation thus remains content dependent. Appendix A.3 summarizes these local properties.

The gate acts on the value-aggregated attention features rather than on the softmax matrix. For a fixed query and channel, the same gate scales contributions from every key, so it neither renormalizes attention probabilities nor directly favors text keys over image keys within the current operation. Changes in the measured text-attention fraction instead emerge through training, subsequent layers, and denoising updates. This placement also differs from the backbone's adaLN-Zero conditioning (Peebles & Xie, 2023): its residual gate scales the projected attention branch, whereas $G_s(t)$ acts before W_o^s and combines token-dependent content logits with an explicit timestep offset. Because pre-projection channel gating and post-projection residual scaling are generally not interchangeable, TSGate provides a distinct conditioning path. Appendices A.4 and A.5 give the formal details.

4 Experiments

4.1 Attention Sinks in Diffusion Transformers

We begin by examining DiTs (Peebles & Xie, 2023) trained on structured captions. Although these models generate high-quality images from in-domain prompts, their performance degrades severely on out-of-domain (OOD) inputs. Figure 3 illustrates this contrast with three paired examples: generations from full-recaption structured prompts depict coherent scenes, whereas those from raw prompts omit requested objects, distort spatial relations, or lose semantic coherence.

Such failures are not surprising in isolation: free-form prompts are absent from the training distribution, so the model has no guarantee of preserving its behavior under this format shift. We therefore investigate the internal mechanism through which prompt format affects generation. Inspection of the joint-attention maps reveals a pronounced *attention sink* phenomenon (Xiao et al., 2024; Wu & Summa, 2026). During inference with in-domain structured prompts, a large fraction of attention is absorbed by a small set of fixed template tokens, particularly recurring section-heading markers.

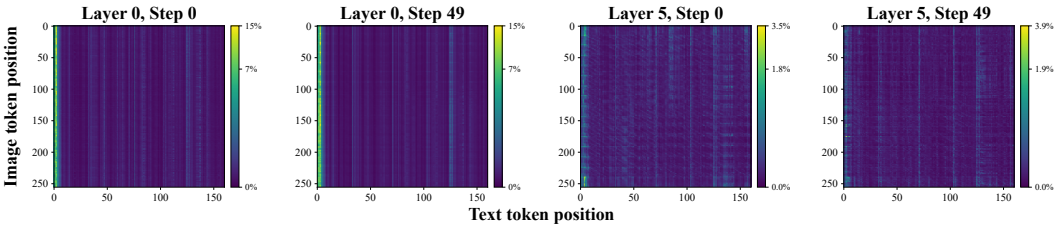


Figure 4: **Attention sinks persist across depth and denoising time.** Image-to-text attention maps for an in-domain example at two layers and two denoising steps. Bright vertical bands show attention concentrated on the same text tokens across image-token positions.

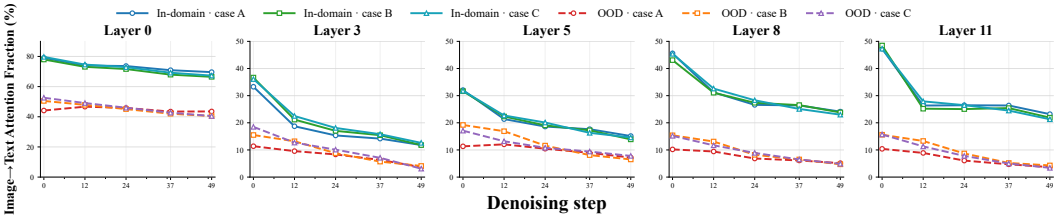


Figure 5: **Image-to-text attention across denoising for in-domain and OOD prompts.** Across five layers and three paired cases, OOD prompts (dashed curves) exhibit an early-step attention deficit relative to in-domain prompts (solid curves).

Figure 4 visualizes image-to-text blocks from an in-domain example at two layers and two denoising steps, showing that these vertical sink bands persist across both depth and time.

Dataset-level statistics confirm that this concentration is substantial: recurring heading markers and separators such as "`\n`" and "`,`" receive more than $20\times$ the attention assigned to an average token. The model thus appears to rely on stable template markers with limited standalone semantic content as attention sinks. When these anchors disappear under OOD prompts, the learned attention organization is disrupted.

The timestep dimension exposes a second, complementary failure mode. As shown in Figure 5, image tokens allocate a large fraction of their attention to text early in denoising for in-domain prompts, then gradually reduce this interaction as visual structure emerges. OOD prompts lack this elevated initial level of text attention: their image-to-text attention fraction remains substantially below the in-domain trajectory precisely in the high-noise regime, when global semantics must be established. This deficit extends beyond the sink-dominated first layer: Layers 3, 5, 8, and 11 also lack the strong initial text interaction observed with in-domain prompts. The prompt shift therefore affects cross-modal information allocation throughout the backbone.

4.2 Text-to-Image Experiments

The preceding analysis suggests that robust prompt generalization requires both controlling attention sinks and preserving early image-to-text interaction. We design TSGate to address these objectives jointly and evaluate it at scale on text-to-image generation. A natural alternative is standard gated attention (Qiu et al., 2025), which is effective at suppressing sinks. However, it lacks a separately parameterized timestep offset. The results below confirm that directly transferring content-only gated attention to a DiT is suboptimal.

Introducing gated attention mid-training yields performance inferior to the baseline, since gate-controlled attention must be learned at the start (Qiu et al., 2025). Therefore, we validate TSGate by training from scratch. Our training data comprise a curated 100M-image subset of LAION-5B (Schuhmann et al., 2022), re-captioned with Qwen2.5-VL 72B (Bai et al., 2025) in a structured format detailed in Appendix D. Further training and optimization details appear in Appendix B.

4.2.1 Experimental Setup

We use a vanilla DiT as Baseline. Notably, to avoid the influence of differing parameter counts, Baseline should be scaled up to match the parameter of TSGate. Therefore, our primary experiments use a MoE architecture, which provides a simple way to align the total parameter count with

Table 1: **TSGate improves generation quality across prompt formats and benchmarks.** We compare Baseline, Gated attention, and TSGate on DPG-Bench with raw and full-reception prompts, and on GenEval and CompBench. Higher is better; DPG scores use a 0–100 scale, while the others use 0–1. Bold marks the best result among the three models shown; complete results and confidence intervals for all five variants appear in Appendix C.

Model	DPG raw ↑	DPG full reception ↑	GenEval ↑	CompBench ↑
Baseline	55.483	78.637	0.7861	0.5224
Gated attention	58.976	78.327	0.7862	0.5155
TSGate	**60.749**	**79.355**	**0.8064**	**0.5236**

TSGate by widening its shared experts. We use a backbone of 12-layer MoE DiT with 12 attention heads per layer and a head dimension of 128. Following the gate definitions in Section 3, the resulting configurations are Baseline (1.25B parameters), Gated attention (standard element-wise gated attention, 1.16B), and element-wise TSGate (1.25B).

We evaluate on DPG-Bench (Hu et al., 2024) using both raw prompts and full-reception prompts, GenEval (Ghosh et al., 2023) with 553 prompts, and T2I-CompBench (Huang et al., 2023) with 900 prompts. Unless stated otherwise, images are generated with a classifier-free guidance (Ho & Salimans, 2022) scale of 5, a resolution of 320×320, and 50 denoising steps. Additional architecture, optimization, and evaluation details appear in Appendix B.

4.2.2 Main Results

Table 1 shows that TSGate consistently improves mean generation scores. On raw DPG-Bench prompts, TSGate scores 60.749, improving over Baseline by 5.266 points (9.5%) and over Gated attention by 1.773 points. TSGate also obtains the highest full-reception score, indicating that its benefit is not restricted to short OOD inputs. On compositional evaluation, TSGate improves GenEval from 0.7861 to 0.8064 and CompBench from 0.5224 to 0.5236 relative to Baseline. The TSGate−Gated-attention difference is positive on both external benchmarks (Appendix Table 7). The category breakdowns identify where these gains are concentrated (Appendix Tables 8 and 9).

A learned shift from text acquisition to visual refinement. The attention trajectories in Figure 6 help characterize this quality difference. TSGate exhibits stronger early image-to-text interaction: at $t = 1000$, its attention fraction exceeds that of content-only gating by approximately 15.9 percentage points (103% relative), and that of Baseline by 7.8 points. The early TSGate−Gated-attention gain appears in both equal-count length groups: 16.6 points for short prompts and 15.2 for long prompts. Thus, the aggregate increase is not driven solely by one prompt-length group.

More importantly, TSGate does not simply increase text attention throughout denoising. In all three panels, the red TSGate mean curve starts above the blue Baseline, crosses below it between the displayed $t = 520$ and $t = 260$ positions, and remains lower at $t = 20$. This shared crossover indicates stage-dependent reallocation: image tokens attend more to text early, while their reduced late-stage text fraction is consistent with greater reliance on image–image interactions relative to Baseline. This pattern admits an intuitive coarse-to-fine interpretation of joint text–image interaction during denoising: noisy image tokens initially consult text to establish object identities and global layout; as visual structure emerges, interactions among image tokens become more useful for refining local appearance and detail. As noted in SD3 (Esser et al., 2024), generation progresses differently across denoising stages: early steps establish coarse structure, whereas later steps refine details.

Neither a crossover nor a prescribed early/late attention schedule is built into TSGate. Moreover, its output gate does not directly renormalize text versus image probabilities (Appendix A.4). The observed pattern therefore reflects an emergent stage-dependent allocation in the learned network, rather than a manually imposed switching rule.

Indeed, measured first-layer sink suppression follows Gated attention > TSGate > Baseline (Appendix F), whereas raw-prompt DPG quality follows TSGate > Gated attention > Baseline. Sink suppression alone therefore does not explain OOD robustness in DiTs. Content-only gating has the lowest text-attention fraction throughout Figure 6, including the early semantic-acquisition stage. TSGate instead combines reduced sink concentration with stronger early text interaction and a later

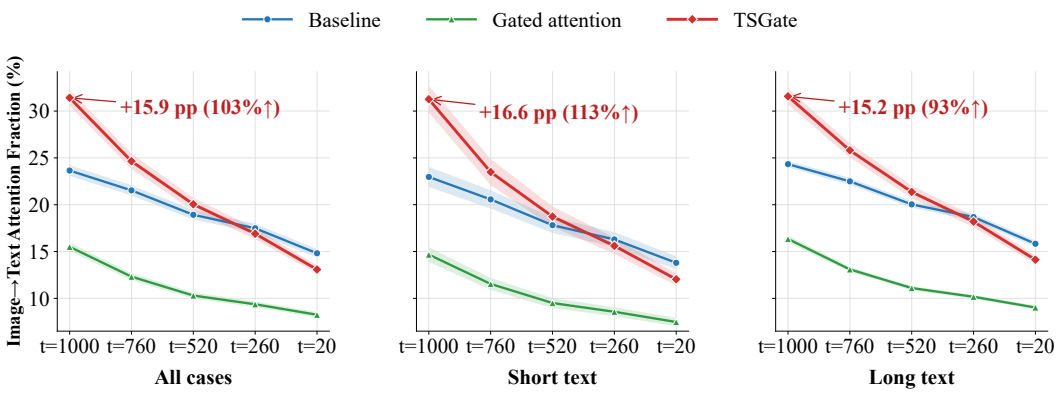


Figure 6: **TSGate exhibits stronger early text attention and lower late text attention than Baseline.** Panels show Layer 5 image-to-text attention for all prompts, short prompts, and long prompts, from left to right. Shaded bands show 95% confidence intervals. Red annotations report the early TSGate–Gated-attention gap in percentage points (pp), with the relative increase in parentheses.

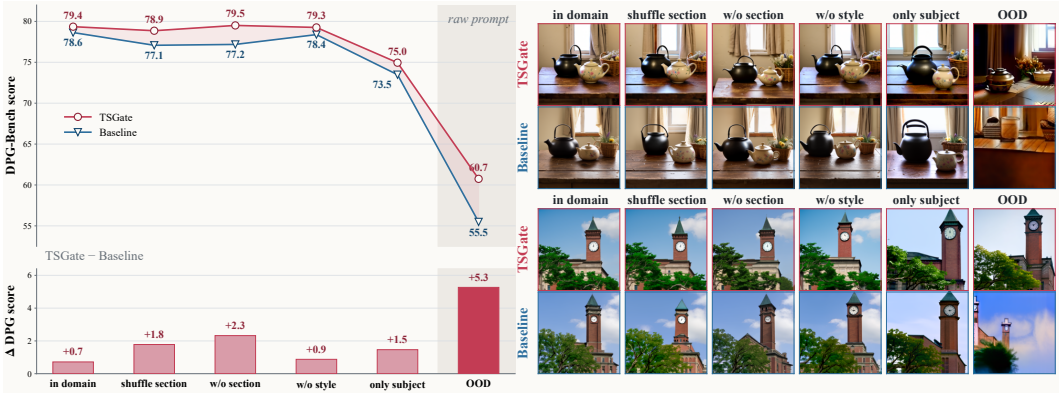


Figure 7: **TSGate is more robust as prompt structure is removed.** Left: DPG-Bench scores and the TSGate–Baseline gap across six prompt conditions, from the structured training format to the raw OOD prompt. Right: examples under the same conditions, with TSGate above Baseline.

shift away from text relative to Baseline. The useful objective is thus stage-appropriate allocation, not minimizing either sink concentration or text attention in isolation.

Progressive prompt degradation. To isolate robustness to prompt format and descriptive context, we progressively move from the training distribution to raw benchmark inputs using six conditions. *In domain* retains the complete structured prompt with section headings. *Shuffle section* reorders the section headings while preserving the descriptive content, thereby disrupting its organization. *W/o section* removes all heading markers (e.g., `#Summary` and `#Setting`) but retains the prose. *W/o style* removes the mood and visual-style sections. *Only subject* retains only the subject-description section and discards all other sections. Finally, *OOD* replaces the structured description with the original short prompt from DPG-Bench, fully departing from the training-time format.

Figure 7 shows that both models decline overall as structure and descriptive context are removed. TSGate remains better throughout, and its advantage is largest under the strongest shift: using unrounded scores, the gap grows from 0.7 points on in-domain prompts to 5.3 points on OOD prompts.

The intermediate conditions distinguish sensitivity to template organization from reliance on descriptive content. Shuffling sections preserves scene information yet lowers Baseline's score from 78.6 to 77.1, suggesting sensitivity to ordering beyond the facts supplied. Removing headings leaves TSGate near its in-domain score (79.5 versus 79.4), while Baseline's score falls to 77.2. This is consistent with TSGate making less use of template markers as indispensable anchors. Both models deteriorate substantially when only the subject description remains and when it is replaced by the raw prompt. Further discussion is provided in Appendix G.

The qualitative examples reinforce this trend. Under OOD prompts, Baseline exhibits pronounced compositional and color distortions in both the teapot and clock-tower scenes, whereas TSGate preserves recognizable subjects and produces substantially more coherent compositions.

4.2.3 TSGATE FAMILY

We further explore TSGate across gating configurations and backbone architectures, considering head-wise modulation, a dense backbone, and a first-layer time branch.

The head-wise variant, TSGate-head, replaces element-wise gates with one shared gate value per attention head in each stream and layer. Both the content and time branches use this coarser granularity. On raw DPG-Bench, TSGate-head scores 55.069, slightly below Baseline at 55.483 and substantially below element-wise TSGate at 60.749. Appendix Tables 2 and 4 provide the configurations and complete results.

Experiments on a dense backbone further test whether TSGate's benefits extend beyond MoE architectures (Appendix C.1). On raw prompts, TSGate improves DPG from 41.619 to 64.159, a gain of 22.540 points. Under full recaption, the mean increases from 81.8 to 82.3. These results support the applicability of TSGate beyond MoE, particularly for raw prompts. Note that the dense comparison is not parameter-matched. Architecture details appear in Appendix Table 3.

Finally, we examine whether the timestep branch is needed in every layer. The sink score measurements in Appendix Figure 16 show the strongest concentration at the first transformer layer among the sampled layers. We also observe more pronounced cross-modal attention allocation in the first layer than in the other layers. This motivates TSGate-L0, which retains content gates in all layers but applies the timestep branch only at Layer 0. TSGate-L0 reaches 60.4 on raw DPG-Bench, compared with 60.7 for all-layer TSGate, retaining 93.8% of its improvement over Baseline. Meanwhile, the time-branch parameter count drops from 85.0M to 7.1M (Appendix Table 2). Thus, restricting the time branch to the first layer preserves most of the observed gain with fewer additional parameters.

4.2.4 TIME-BIAS INTERVENTIONS

To test the importance of temporal alignment, we reverse the learned time-bias schedule while keeping the checkpoint and backbone timestep conditioning fixed. This intervention reduces the raw DPG-Bench score by 2.6 points. Reversal preserves the learned bias values and their schedule average, changing only their assignment to denoising stages. The resulting degradation supports the functional importance of temporal alignment beyond the presence of an additive bias. Full intervention definitions and results are provided in Appendix F.

5 CONCLUSION

We investigate why Diffusion Transformers (Peebles & Xie, 2023) trained on structured prompts degrade when given free-form user prompts. Our analysis links this fragility to attention sinks on template markers and reduced early-step image-to-text attention during high-noise denoising. Standard gated attention (Qiu et al., 2025) suppresses sinks but exhibits weaker early text interaction, showing that sink suppression alone is insufficient to restore generation quality. To address this mismatch, we propose TSGate, Timestep-Aware Gated Attention, a lightweight, plug-and-play modification that adds a timestep-conditioned bias to the gate signal. TSGate jointly reduces attention sinks and restores early text guidance, spontaneously learning a coarse-to-fine shift from semantic acquisition to visual refinement. It improves the raw-prompt DPG-Bench (Hu et al., 2024) score by 9.5% over the baseline while also achieving gains on GenEval (Ghosh et al., 2023) and CompBench (Huang et al., 2023). Moreover, applying the timestep branch only to the first layer retains 93.8% of the improvement with minimal parameter overhead, offering an efficient path toward improving DiT performance.

While our evaluation centers on text-to-image generation, extending TSGate to broader multimodal tasks (e.g., text-to-audio and text-to-video generation) remains to be explored. Moreover, visual tasks such as image editing require more nuanced spatial relationships between tokens than those typically modeled by LLMs. Combining TSGate with spatial positional encodings could thus be a worthwhile direction to explore.

Reproducibility Statement

Section 3 specifies the computation; Appendices A.3 and A give the gate properties and MoE configuration details; Appendix B specifies generation settings. Appendix C collects the overall scores, category breakdowns, and reported confidence intervals for all three benchmarks. Appendix F reports the complete bias interventions.

AI Use Statement

OpenAI Codex assisted with writing code, retrieving literature, revising prose, and formatting the LaTeX document. We use Qwen2.5-VL 72B to generate structured captions for the training images, and deepseek-v4.1-flash and Gemini-3.1-pro to rewrite the original evaluation prompts into in-domain structured prompts. We have not used generative AI tools for synthesizing training images, developing theoretical models, proposing hypotheses, designing research methodology, interpreting results, or any other tasks requiring mandatory disclosure beyond those stated above.

All AI-assisted outputs were rigorously reviewed by the authors. All AI-polished text was carefully reviewed to ensure it faithfully represents our original ideas and claims. We take full responsibility for the final content of this work, including all text, claims, code, mathematical proofs, and artifacts produced with the aid of generative AI.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A ViT backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22669–22679, 2023.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. Technical report, OpenAI, 2023. URL https://cdn.openai.com/papers/dall-e-3.pdf.

Black Forest Labs. FLUX.1: Text-to-image generation model. https://github.com/black-forest-labs/flux, 2024.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. https://openai.com/index/video-generation-models-as-world-simulators/, 2024.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/fe989bb038b5dcc44181255dd6913e43-Abstract-Conference.html.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0b408293619f725fd30162af057e531a-Abstract-Conference.html.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993. doi: 10.1201/9780429246593.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018. doi: 10.1016/j.neunet.2017.12.012. URL https://www.sciencedirect.com/science/article/pii/S0893608017302976.

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, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633, 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

Zhengcong Fei, Mingyuan Fan, Changqian Yu, Debang Li, and Junshi Huang. Scaling diffusion transformers to 16 billion parameters. *arXiv preprint arXiv:2407.11633*, 2024. URL https://arxiv.org/abs/2407.11633.

Zizhuo Fu, Wenxuan Zeng, Runsheng Wang, and Meng Li. Attention sink forges native MoE in attention layers: Sink-aware training to address head collapse. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.01203.

Peng Gao, Le Zhuo, Dongyang Liu, Ruoyi Du, Xu Luo, Longtian Qiu, Yuhang Zhang, Chen Lin, Rongjie Huang, Shijie Geng, Renrui Zhang, Junlin Xi, Wenqi Shao, Zhengkai Jiang, Tianshuo Yang, Weicai Ye, He Tong, Jingwen He, Yu Qiao, and Hongsheng Li. Lumina-T2X: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945*, 2024. URL https://arxiv.org/abs/2405.05945.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.nips.cc/paper_files/paper/2023/hash/a3bf71c7c63f0c3bcb7ff67c67b1e7b1-Abstract-Datasets_and_Benchmarks.html.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d346d91999074dd8d6073d4c3b13733b-Abstract.html.

Team Happyhorse. Happyhorse 1.0. Video Generation Platform, 2026. URL https://www.happyhorse.com/.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://direct.mit.edu/neco/article/9/8/1735/6109/Long-Short-Term-Memory.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: Equip diffusion models with LLM for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. URL https://arxiv.org/abs/2403.05135.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f8ad010cdd9143dbb0e9308c093aff24-Abstract.html.

Sangwu Lee, Erwann Millon, Le Zhuo, Matthew Newton, Andrei Filatov, Naga Sai Abhinay Devarinti, Dazhi Zhong, Avram Djordjevic, Gabriel Menezes, Will Beddow, Titus Ebbecke, Mihai Petrescu, Owen Fahey, Gian Saß, Felix Gil, Albert Salgueda, and Victor Perez. Krea 2 technical report. Krea, 2026. URL https://www.krea.ai/blog/krea-2-technical-report.

Maohua Li, Qirui Li, Yanke Zhou, Yiduo Li, Zhaosheng Chi, Chao Xu, Cuifeng Shen, Yixuan Xu, Hanlin Tang, Kan Liu, Tao Lan, Lin Qu, and Shao-Qun Zhang. Text template tokens are implicit semantic registers in diffusion transformers. *arXiv preprint arXiv:2607.19139*, 2026. URL https://arxiv.org/abs/2607.19139.

Zhimin Li et al. Hunyuan-DiT: A powerful multi-resolution diffusion transformer with fine-grained Chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Haozhe Liu, Wentian Zhang, Jinheng Xie, Francesco Faccio, Mengmeng Xu, Tao Xiang, Mike Zheng Shou, Juan-Manuel Perez-Rua, and Jürgen Schmidhuber. Faster diffusion via temporal attention decomposition. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=xXs2GKXPnH.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Oscar Mañas, Pietro Astolfi, Melissa Hall, Candace Ross, Jack Urbanek, Adina Williams, Aishwarya Agrawal, Adriana Romero-Soriano, and Michal Drozdzal. Improving text-to-image consistency via automatic prompt optimization. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=g12Gdl6aDL.

MiniMax. MiniMax H3: An open model breaking the boundaries between tasks and modalities. MiniMax Research Blog, July 2026. URL https://www.minimax.io/blog/minimax-h3. Published July 31, 2026. Official model release.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, et al. RWKV: Reinventing RNNs for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14048–14077, 2023.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/904e89bb4e632e75fb47f093b620b257-Abstract-Conference.html.

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*, pp. 10684–10695, 2022.

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. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063. URL https://www.sciencedirect.com/science/article/pii/S0925231223011864.

Zunhai Su et al. Attention sink in transformers: A survey on utilization, interpretation, and mitigation. *arXiv preprint arXiv:2604.10098*, 2026. URL https://arxiv.org/abs/2604.10098.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

Team Seedance et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026. URL https://arxiv.org/abs/2604.14148.

Team Wan, Ang Wang, Baole Ai, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Fangzheng Wu and Brian Summa. Attention sinks in diffusion transformers: A causal analysis. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.09313.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56501–56523, 2024. URL https://proceedings.mlr.press/v235/yang24ab.html.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ce31378e9f41d8907e97dab172b6c559-Abstract-Conference.html.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1910.07467.

Wangbo Zhao, Yizeng Han, Jiasheng Tang, Kai Wang, Yibing Song, Gao Huang, Fan Wang, and Yang You. Dynamic diffusion transformer. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a44a70acd5d0abc1a252ada9719dd06d-Abstract-Conference.html.

Lianghui Zhu, Zilong Huang, Bencheng Liao, Jun Hao Liew, Hanshu Yan, Jiashi Feng, and Xinggang Wang. DiG: Scalable and efficient diffusion models with gated linear attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhu_DiG_Scalable_and_Efficient_Diffusion_Models_with_Gated_Linear_Attention_CVPR_2025_paper.html.

A GATE IMPLEMENTATION AND CONDITIONING

A.1 GRANULARITY, LAYER SCOPE, AND INITIALIZATION

Gate granularity. Section 3.2 gives the default element-wise form of gated attention (Qiu et al., 2025). More generally, let $D_g = D$ for element-wise gating and $D_g = H$ for head-wise gating. The content projection has $W^s_{q,g} \in \mathbb{R}^{D\times(D+D_g)}$ and $b^s_{q,g} \in \mathbb{R}^{1\times(D+D_g)}$, yielding $Q_s(t) \in \mathbb{R}^{N_s\times D}$ and $C_s(t) \in \mathbb{R}^{N_s\times D_g}$. The time projection has $W^{\text{time}}_s \in \mathbb{R}^{d_t\times D_g}$ and $b^{\text{time}}_s \in \mathbb{R}^{1\times D_g}$. The two forms share the computation

$$G_s(t) = \sigma(\text{expand}(C_s(t) + \text{broadcast}(B_{\text{ts},s}(t)))) \,. \tag{7}$$

Here broadcast repeats the time bias along the token dimension, and expand maps the D_g gate channels to D attention channels as in Equation (1). Thus $G_s(t) \in \mathbb{R}^{N_s\times D}$ in either case, and Equation (6) is unchanged. Head-wise gating shares one sigmoid value across each head's d_h channels; it changes the granularity of both branches.

Layer scope and evaluated configurations. We evaluate five MoE configurations (Table 2). Baseline has no output gate. Gated attention gates content element-wise in all 12 layers. TSGate and TSGate-L0 retain this content gating, adding time biases in all layers and Layer 0, respectively. TSGate-head gates both branches head-wise in all layers; TSGate versus TSGate-head thus compares both branches' granularity. Dense experiments compare a no-output-gate baseline with full TSGate on a 24-layer backbone.

For layers without a time branch, Equation (7) uses $B_{\text{ts},s}(t) = 0$ and no time-projection parameters. The first-layer configuration therefore retains content-only gating in Layers 1–11; it does not remove their output gates.

Initialization. For the MoE configurations, the content-logit entries of $b^s_{q,g}$ are initialized to 4, while W^{time}_s and b^{time}_s are initialized to zero. Consequently, $B_{\text{ts},s}(t) = 0$ at initialization, and the gate is exactly the corresponding content-only gate for the same content-branch parameters. If the token-dependent contribution to a logit is zero, its initial gate value is $\sigma(4) \approx 0.982$; in general, that value also depends on the projected token features. This initialization favors retention but does not make the gated model exactly equivalent to an ungated baseline.

A.2 JOINT ATTENTION LAYOUT

For element-wise gating, the projection in Equation (3) is reshaped to $N_s \times H \times 2d_h$ and split within each head into queries and content logits. Head-wise gating instead uses $N_s \times H \times (d_h + 1)$, with one content logit per head. Keys and values use independent affine projections. Queries and keys undergo per-head RMSNorm (Zhang & Sennrich, 2019) followed by RoPE (Su et al., 2024); content logits undergo neither operation. Each stream's queries attend to the concatenated keys and values of all streams, following MMDiT (Esser et al., 2024), using the standard softmax attention computation (Vaswani et al., 2017) in Section 3.1. Head concatenation gives $A_s(t)$, which is gated before the output projection. Gating before head concatenation is equivalent when the head–channel layout is preserved.

A.3 SHARED BIAS AND CONTENT SELECTION

For a fixed layer, stream, channel, and step, write $g_i(b) = \sigma(c_i + b)$. The sigmoid is strictly increasing, so a shared bias preserves the ordering of the given content logits across tokens. Also, $g_i(b) > 1/2$ exactly when $c_i > -b$, and $\partial g_i/\partial b = g_i(1-g_i) \le 1/4$. Thus, a common offset changes retention most strongly for gates near $1/2$, while saturated gates respond less. These are local properties at fixed content logits; they do not require a monotone timestep schedule or guarantee an increase in text attention.

A.4 Output gating does not renormalize attention probabilities

Let P_{ij}^h be the pre-gate softmax probability from query i to key j in head h. For channel c, output gating gives

$$\widetilde{A}_{s,i,c}^h = G_{s,i,c}^h A_{s,i,c}^h = \sum_j G_{s,i,c}^h P_{ij}^h V_{j,c}^h. \tag{8}$$

The gate is shared across keys for a fixed query and channel, so these coefficients sum to $G_{s,i,c}^h$ rather than one. They do not define a new softmax distribution or selectively rescale text keys relative to image keys. Our attention metrics use the head-averaged P (Equation (9)). Output gating affects subsequent layers and denoising updates, and training can change the learned projections; it does not alter probabilities already computed in the same operation.

Using the head-averaged, pre-gate probability P_{ij} from query i to key j, we define the reported attention metrics as follows, where $\mathcal{I}$ is the image-query set, $\mathcal{J}$ is the text-key set, and $\mathcal{K}$ is the set of all keys:

$$\begin{aligned} \text{Image-to-text attention fraction} &= \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} P_{ij}, \\ \text{Peak attention mass} &= \max_{j \in \mathcal{K}} \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} P_{ij}. \end{aligned} \tag{9}$$

Both statistics use image queries and take values in $[0, 1]$. Peak attention mass searches all key modalities, whereas the image-to-text attention fraction sums over text keys. The image-to-text attention plots in the main text report this fraction as a percentage, multiplying it by 100.

A.5 Relationship to adaLN-Zero and residual gating

Following the adaLN-Zero formulation of DiT (Peebles & Xie, 2023), let U_s be the residual-stream input and $\gamma_s(t)$, $\beta_s(t)$, and $\alpha_s(t)$ the backbone's channel-wise scale, shift, and residual gate. Omitting the MLP branch and token broadcasting,

$$\begin{aligned} X_s(t) &= \big(1 + \gamma_s(t)\big) \odot \mathrm{LN}(U_s) + \beta_s(t), \\ Y_s(t) &= \big(G_s(t) \odot A_s(t)\big) W_o^s + b_o^s, \\ U_s^+ &= U_s + \alpha_s(t) \odot Y_s(t). \end{aligned} \tag{10}$$

The content logits already depend on t through $X_s(t)$. TSGate adds a separately parameterized offset to those logits, rather than making an otherwise timestep-independent model time aware. Its token-dependent gate acts before the channel-mixing projection W_o^s, whereas $\alpha_s(t)$ scales the projected residual branch; these operations are generally not interchangeable. The distinction is placement and combination with content logits, not the use of SiLU (Elfwing et al., 2018). The two conditioners also differ in modulation type: adaLN-Zero applies an affine map with unbounded scale and shift $\gamma_s(t), \beta_s(t) \in \mathbb{R}^D$, whereas the sigmoid restricts each gate entry to $(0, 1)$, so the time branch can only attenuate or pass an attention channel. This bounded, multiplicative form gives it an inductive bias toward selective suppression rather than arbitrary feature rescaling.

A.6 Signal-to-noise view of the time bias

The content and time branches consume different inputs, which motivates an explicit time bias even though the content logits already depend on t through $X_s(t)$ (Appendix A.5). Write the content logits as $C_s(t) = X_s(t) W_{q,g}^s[\text{gate}] + b_{q,g}^s[\text{gate}]$. In flow/diffusion training (Lipman et al., 2023; Ho et al., 2020), the stream input carries a noisy latent $x_t = \alpha_t x_0 + \sigma_t \epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$, so the layer input decomposes as $X_s(t) = X_s^{\text{signal}} + X_s^{\text{noise}}$, where the first term carries semantic content and the second reflects residual noise. During the early, high-noise steps (large t, low SNR), the noise term dominates, $\|X_s^{\text{signal}}\| \ll \|X_s^{\text{noise}}\|$, and

$$C_s(t) \approx X_s^{\text{noise}} W_{q,g}^s[\text{gate}] + b_{q,g}^s[\text{gate}], \tag{11}$$

so the token-dependent logits that should steer retention are themselves contaminated by noise precisely when reliable gating matters most.

The time bias avoids this pathway. Since $B_{\mathrm{ts},s}(t) = \mathrm{SiLU}(t_{\mathrm{emb}}(t))W_s^{\mathrm{time}} + b_s^{\mathrm{time}}$ is a deterministic function of the scalar timestep and never passes through the noisy latent, it supplies a noise-free, token-shared prior that is well defined at every noise level. Because $B_{\mathrm{ts},s}(t)$ is added before the sigmoid (Equation (5)), it can set a stage-appropriate operating point for the gate when $C_s(t)$ is unreliable, then recede in relative influence as the SNR of $X_s(t)$ rises during late steps and the content logits recover their fidelity. This yields an implicit, learned curriculum in which the timestep prior governs retention early and cedes control to content selection later, without any explicit schedule.

B Experimental configuration and evaluation protocol

B.1 Architecture and training

Table 2 brings together the architecture, gate settings, and parameter counts of all five evaluated MoE variants. Widening the Baseline's shared expert approximately matches TSGate's total parameter count, but the capacity allocation differs across variants. The fixed-weight interventions in Appendix F all use the same trained MoE TSGate checkpoint.

Table 2: **Architecture and gate settings of the five MoE configurations.** Gated attention (Qiu et al., 2025) uses content-only gating; TSGate-L0 adds the time branch only at Layer 0. EW and HW denote element-wise and head-wise gating. Parameter counts are in millions (M) and include all saved streams; totals are rounded to 0.001M.

Configuration	Baseline	Gated attention	TSGate	TSGate-L0	TSGate-head
Transformer layers	12	12	12	12	12
Hidden width D	1,536	1,536	1,536	1,536	1,536
Attention heads	12	12	12	12	12
Head dimension	128	128	128	128	128
Text input width	3,584	3,584	3,584	3,584	3,584
Routed experts E	8	8	8	8	8
Active experts k	2	2	2	2	2
Routed-expert FFN width	1,024	1,024	1,024	1,024	1,024
Shared experts	1	1	1	1	1
Shared-expert width	5,120	2,048	2,048	2,048	2,048
Text/audio FFN width	1,536	1,536	1,536	1,536	1,536
Content-gate granularity	–	EW	EW	EW	HW
Content-gated layers	–	0–11	0–11	0–11	0–11
Time-gate granularity	–	–	EW	EW	HW
Time-conditioned layers	–	–	0–11	0	0–11
Content-gate bias init.	–	4	4	4	4
Time-branch parameters (M)	0.000	0.000	84.990	7.082	0.664
Total parameters (M)	1,250.706	1,165.826	1,250.816	1,172.909	1,082.164

Table 3 lists the two dense checkpoints compared in Appendix C.1. They share the 24-layer backbone dimensions; TSGate adds element-wise content gates in all 24 layers with biases initialized to 4 and a zero-initialized time projection in every layer. As in the MoE models, the saved checkpoints register image, text, and audio stream projections, so the dense time branch contributes $3 \times 24 \times (1{,}536 \times 1{,}536 + 1{,}536) = 169{,}979{,}904$ parameters, the same amount as its widened content-gate projections. No dense content-only gated-attention results are reported.

All configurations are trained on the curated 100M-image subset of LAION-5B (Schuhmann et al., 2022) re-captioned with Qwen2.5-VL 72B (Bai et al., 2025) into the structured-prompt format of Appendix D, using 32 NVIDIA H200 GPUs. Training uses the AdamW optimizer (Loshchilov & Hutter, 2019) with a learning rate of 1×10^{-4}, a weight decay of 0.01, and a maximum gradient norm of 1.0. The learning rate is warmed up over the first 1% of training and then held constant. Training uses a text-drop ratio of 0.1 for classifier-free guidance (Ho & Salimans, 2022) and a resolution of

Table 3: **Dense backbone configurations for the ungated baseline and TSGate.** All 24 image FFNs are dense, so routed/shared-expert settings are inactive. Audio FFN width is a configured value, although generation uses image and text streams. Parameter counts are in millions (M) and include all saved streams; totals are rounded to 0.001M.

Configuration	Baseline	TSGate
Transformer layers	24	24
Hidden width D	1,536	1,536
Attention heads	12	12
Head dimension	128	128
Text input width	4,096	4,096
Image FFN width	3,072	3,072
Text FFN width	1,024	1,024
Audio FFN width (configured)	128	128
Content-gate granularity	–	Element-wise
Content-gated layers	–	0–23
Time-gate granularity	–	Element-wise
Time-conditioned layers	–	0–23
Content-gate bias init.	–	4
Time-branch parameters (M)	0.000	169.980
Total parameters (M)	1,167.000	1,506.960

320×320. SNR sampling follows a log-normal distribution with mean 0 and standard deviation 1 in log space. All models are trained on the same total of 100M images.

B.2 Inference

MoE inference uses CFG=5, 320×320 resolution, 50 sampling steps, timeshift=3, and seeds 42, 1234, 7777, and 8888. DPG-Bench (Hu et al., 2024) scores the four-image grid for each source and averages its four member scores before averaging sources, on a 0–100 scale. GenEval (Ghosh et al., 2023) and T2I-CompBench (Huang et al., 2023) first average seeds within each source and then equally weight the six category means. Dense inference uses the same CFG, resolution, and sampling-step settings.

C Complete benchmark results

This section collects the reported overall scores, category breakdowns, and paired confidence intervals for DPG-Bench (Hu et al., 2024), GenEval (Ghosh et al., 2023), and T2I-CompBench (Huang et al., 2023). The five variants are Baseline, Gated attention (Qiu et al., 2025) (content-only gating), TSGate, TSGate-L0 (first-layer time branch), and TSGate-head (head-wise gating) (Appendix A). DPG scores use a 0–100 scale; GenEval and CompBench use a 0–1 scale. Higher scores indicate better performance on all benchmarks. Generation and aggregation protocols are specified in Appendix B, and parameter settings are collected in Tables 2 and 3. Differences use unrounded, source-paired scores after averaging scores over four seeds within each source. Reported 95% confidence intervals (CIs) are pointwise percentile intervals (Efron & Tibshirani, 1993) from 10,000 source-bootstrap resamples (seed 20260909), with within-category resampling for external benchmarks.

C.1 DPG-Bench

Table 4 collects the raw and full-recaption scores for all five MoE configurations on the same 1,065 D-main sources. TSGate has the highest mean for both inputs, followed by the first-layer variant TSGate-L0. The TSGate−Gated-attention difference is positive for both inputs, whereas the paired TSGate-L0−TSGate intervals include zero. These are source-paired comparisons of the checkpoints in Table 2, not estimates of training-seed variation.

Table 4: **MoE DPG-Bench scores and paired differences.** Raw and full-recaption prompts use the same 1,065 sources, with scores averaged over four seeds per source. Scores use a 0–100 scale. Bold and underlining mark the highest and second-highest means in each prompt condition; intervals are pointwise 95% source-bootstrap confidence intervals (CIs).

	Raw		Full recaption	
Model	DPG score		DPG score	
Baseline	55.483		78.637	
Gated attention	58.976		78.327	
TSGate	**60.749**		**79.355**	
TSGate-L0	<u>60.424</u>		<u>79.264</u>	
TSGate-head	55.069		78.858	
Contrast	Δ	95% CI	Δ	95% CI
TSGate − Gated attention	+1.773	$[0.709, 2.870]$	+1.028	$[0.239, 1.795]$
TSGate-L0 − TSGate	-0.325	$[-1.291, 0.615]$	-0.091	$[-0.727, 0.530]$

Both dense models use the architecture and inference settings in Appendix B. Table 5 compares the baseline and TSGate on common sources. The larger raw-prompt gain is consistent with the MoE pattern, while the full-recaption paired interval includes zero.

Table 5: **Dense DPG-Bench scores on common sources.** Scores use a 0–100 scale. N counts paired sources; Δ is TSGate minus the baseline, computed before rounding. Intervals are pointwise 95% source-bootstrap CIs for the paired differences.

Input	N	Baseline	TSGate	Δ	95% CI
Raw	1,064	41.619	64.159	+22.540	$[21.175, 23.884]$
Full recaption	1,062	81.809	82.297	+0.489	$[-0.084, 1.068]$

Raw scores cover 1,064 sources per model. Full recaption covers 1,064 baseline and 1,062 TSGate sources; the saved-coverage means are 81.753 and 82.297 (difference 0.544). The paired table uses their 1,062 common sources; the category table retains the evaluator's original coverage and aggregation.

Table 6 gives the dense backbone's five L1 categories (Global, Entity, Attribute, Relation, Other) and 13 L2 entries. Global repeats at L2 because it has no subdivisions. These category aggregates differ from source-level overall scores; MoE results are available only at the overall-score level.

C.2 GenEval and T2I-CompBench: overall scores

Table 7 reports all five MoE variants, with intervals for absolute scores and paired differences. TSGate outperforms Gated attention on both benchmarks, with the highest GenEval mean and the second-highest CompBench mean, where TSGate-head leads. The CompBench TSGate−Baseline interval includes zero.

C.3 GenEval: category breakdown

Table 8 reports all six GenEval categories. TSGate's means exceed those of Gated attention in every category, with the largest differences in position and color attributes. The T2I-CompBench breakdown follows in Appendix C.4. Category means describe where the aggregate gains occur; paired uncertainty for the macro scores appears in Table 7.

C.4 T2I-CompBench: category breakdown

Table 9 reports all six T2I-CompBench categories. The TSGate−Gated-attention difference is largest in spatial relations and color, while Gated attention has the highest complex-category mean and TSGate-head leads the overall macro score.

Table 6: **Dense DPG-Bench category scores.** Raw and full-reception scores (0–100) retain the evaluator's original aggregation and source coverage for first-level (L1) and second-level (L2) categories.

	Raw		Full reception	
Category	Baseline	TSGate	Baseline	TSGate
L1 categories				
Global	76.287	75.758	89.914	85.226
Entity	68.612	77.344	89.647	90.350
Attribute	65.061	78.597	89.015	89.309
Relation	75.489	85.862	89.395	91.396
Other	66.092	80.397	90.113	90.575
L2 categories				
Global / –	76.287	75.758	89.914	85.226
Entity / whole	62.459	75.391	90.496	90.068
Entity / part	71.914	75.591	86.853	81.919
Entity / state	77.866	79.223	90.194	92.351
Attribute / color	65.570	76.822	90.599	90.564
Attribute / shape	57.505	82.126	90.858	81.641
Attribute / size	61.628	76.580	85.992	85.515
Attribute / texture	65.952	82.421	87.500	90.249
Attribute / other	73.557	80.556	89.376	89.733
Relation / spatial	79.070	87.211	91.476	92.308
Relation / non-spatial	71.835	83.547	84.173	90.336
Other / count	59.878	86.484	90.650	84.167
Other / text	69.871	75.499	89.205	92.179

Table 7: **GenEval and T2I-CompBench macro scores and paired differences.** Scores use a 0–1 scale. The six categories are equally weighted after averaging scores over four seeds within each source; pointwise 95% CIs use 10,000 within-category bootstrap resamples. All models use full recaption. Bold and underlining mark the highest and second-highest means for each benchmark.

	GenEval		T2I-CompBench	
Model	Score	95% CI	Score	95% CI
Baseline	0.7861	$[0.7583, 0.8124]$	0.5224	$[0.5057, 0.5394]$
Gated attention	0.7862	$[0.7598, 0.8119]$	0.5155	$[0.4984, 0.5326]$
TSGate	**0.8064**	$[0.7796, 0.8318]$	<u>0.5236</u>	$[0.5070, 0.5404]$
TSGate-L0	0.7887	$[0.7620, 0.8147]$	0.5231	$[0.5066, 0.5398]$
TSGate-head	<u>0.7972</u>	$[0.7703, 0.8231]$	**0.5264**	$[0.5097, 0.5432]$
Contrast	Δ	95% CI	Δ	95% CI
TSGate − Gated attention	+0.0202	$[0.0007, 0.0401]$	+0.0081	$[0.0023, 0.0141]$
TSGate − Baseline	+0.0203	$[0.0025, 0.0389]$	+0.0012	$[-0.0046, 0.0067]$

Table 8: **GenEval category means.** Scores use a 0–1 scale. N is the number of source prompts and Δ denotes TSGate−Gated attention, computed before rounding. Bold marks the highest mean in each row.

Category	N	Baseline	Gated attention	TSGate	TSGate-L0	TSGate-head	Δ
Single object	80	0.9500	0.9563	**0.9625**	0.9563	0.9437	+0.0062
Two objects	99	0.8283	0.8308	0.8359	**0.8485**	0.8409	+0.0051
Counting	80	0.6281	0.6375	**0.6594**	0.6250	0.6531	+0.0219
Colors	94	0.8404	0.8404	**0.8457**	0.8351	0.8431	+0.0053
Position	100	0.8175	0.7800	0.8250	0.7850	**0.8325**	+0.0450
Color attribute	100	0.6525	0.6725	**0.7100**	0.6825	0.6700	+0.0375

Table 9: **T2I-CompBench category means.** Scores use a 0–1 scale. N is the number of source prompts and Δ denotes TSGate−Gated attention, computed before rounding. Bold marks the highest mean in each row.

Category	N	Baseline	Gated attention	TSGate	TSGate-L0	TSGate-head	Δ
Color	150	0.8145	0.8092	**0.8264**	0.8217	0.8200	+0.0172
Shape	150	0.5912	0.5876	0.5903	0.6006	**0.6027**	+0.0027
Texture	150	0.6954	0.6802	0.6852	**0.6998**	0.6913	+0.0050
Spatial	150	0.3409	0.3193	0.3444	0.3217	**0.3480**	+0.0251
Non-spatial	150	0.3043	0.3033	0.3044	0.3038	**0.3053**	+0.0011
Complex	150	0.3879	**0.3932**	0.3907	0.3913	0.3910	−0.0026

D STRUCTURED PROMPT FORMAT

Our text-to-image data use long English captions organized under Markdown headings. The examples below retain the five fields of their paired source captions: a summary, scene context, subject details, visual style, and atmosphere. Within each field, the description remains ordinary prose. The three image–caption pairs below are data examples.



(a) Lavender

Data sample 0439

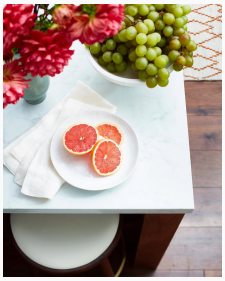
```
# Summary
A close-up of purple lavender blooms, with a few
sharply focused stalks rising against a softly blurred,
warm-toned background.
# Setting
An outdoor lavender patch in soft, diffused daylight;
the foreground stalks are crisp while the background
dissolves into a smooth bokeh.
# Subject
Slender light-green stems carry whorls of small violet
and lilac flowers; a few stalks are in sharp focus
while the rest fade into blur, densely packed near the
base.
# Style
Botanical macro close-up; a pronounced shallow depth of
field isolates the flowers against a warm beige and
sandy bokeh, with cool purples contrasting the earthy
backdrop.
# Mood
Quiet, delicate, and slightly dreamy; the soft light
and gentle blur evoke a calm, intimate moment in
nature.
```



(b) Mountain light

Data sample 0454

```
# Summary
Golden low-sun light falls across rolling hills and a
wooded ridge beneath a range of snow-capped mountains.
# Setting
A mountainous landscape at sunrise or sunset; patchy
snow and dry brush fill the foreground, a
conifer-dotted ridge crosses the midground, and a
broad, partly cloudy sky spans the horizon.
# Subject
Sunlit conifers glow golden on the right of the ridge
while the left falls into shadow; distant peaks show
snow on their upper slopes against a gradient sky from
warm yellow to deep blue.
# Style
Wide landscape photograph with deep focus; a strong
warm-cool contrast pairs golden highlights on the trees
and hillside with cool blue sky and shadowed terrain.
# Mood
Serene, expansive, and quietly dramatic; the interplay
of warm light and cool shadow conveys peaceful
isolation and the scale of nature.
```



(c) Fresh still life

Data sample 0861

```
# Summary
A high-angle still life of pink grapefruit slices,
green grapes, and deep red dahlias arranged on a white
marble countertop.
# Setting
A bright interior lit by soft window light from the
upper left; the marble counter fills the frame, with a
stool and wooden floor visible below and a patterned
rug in the corner.
# Subject
Three coral-pink grapefruit slices rest on a white
plate beside a folded off-white linen napkin; a white
bowl of glossy green grapes sits at the top right and
crimson dahlias in a sage-green vase at the top left.
# Style
High-angle lifestyle photograph; vivid coral, green,
and crimson accents stand out against a high-key
palette of white marble and pale linen, with soft
shadows revealing each texture.
# Mood
Fresh, calm, and inviting; the bright light and clean,
casual arrangement suggest a relaxed, healthy morning.
```

E ADDITIONAL PAIRED GENERATION COMPARISONS

We provide additional qualitative comparisons across all three models, laid out two per row. In each panel the top row uses the full-recaption structured prompt (in domain) and the bottom row uses the raw prompt (out of domain); columns are Baseline, Gated attention, and TSGate. Within each example, all three models and both prompt forms use the same random seed.



Figure 8: Blue washing machines in a laundry room.



Figure 9: A white lighthouse on a rocky coast.

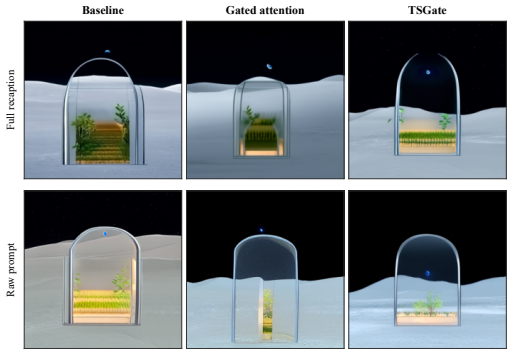


Figure 10: A glass greenhouse on the Moon.

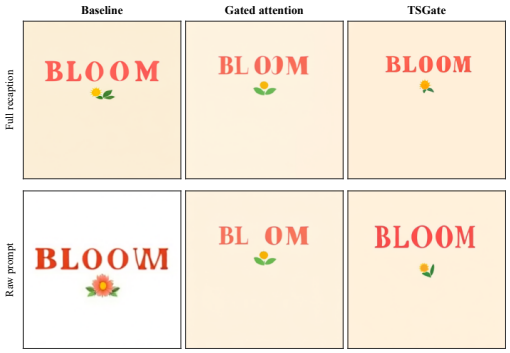


Figure 11: A "BLOOM" poster with a small flower.

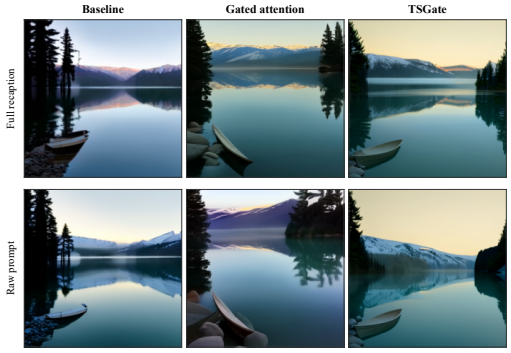


Figure 12: A small boat beside a calm mountain lake.

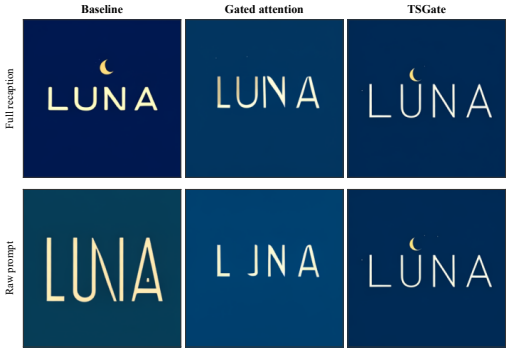


Figure 13: A "LUNA" logo with a small crescent moon.



Figure 14: A blue gemstone pendant on white fabric.

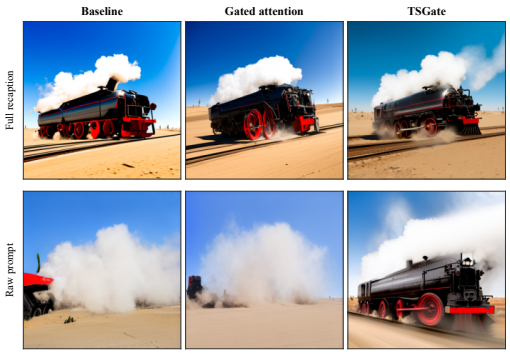


Figure 15: A steam locomotive crossing a desert.

F STAGE-DEPENDENT TEXT INTERACTION AND TIME-BIAS INTERVENTIONS

The strongest sink suppression does not yield the highest generation quality. Across layers and denoising steps, the Sink Score concentrates at Layer 0, where it follows Gated attention (Qiu et al., 2025) $<$ TSGate $<$ Baseline, whereas TSGate outperforms Gated attention on DPG (Hu et al., 2024) (Figure 16, Table 1). The Sink Score is the maximum normalized column-sum of the joint-attention map, i.e., the largest fraction of attention received by any single token; it is measured before the output gate, so it characterizes learned interactions and the hidden states propagated through the network rather than a same-operation renormalization by the gate.

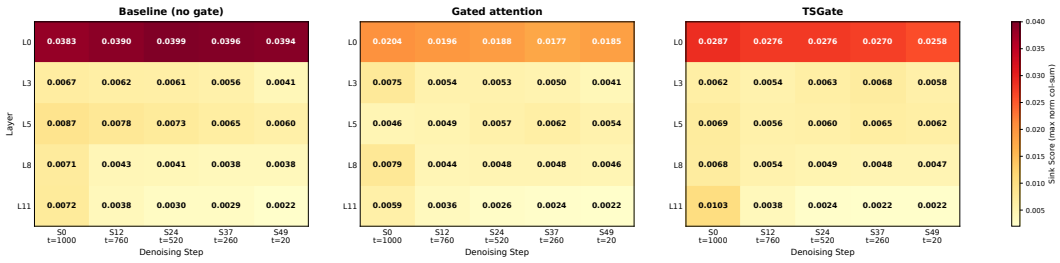


Figure 16: **Attention sinks concentrate at Layer 0 across all three variants.** Heatmaps show prompt-averaged Sink Scores across layers and denoising steps. The Sink Score is the maximum normalized column-sum of the joint-attention map; darker cells indicate stronger sinks. Attention weights are from the conditional branch and measured before the output gate.

The Layer 5 trajectories in Figure 6 show that TSGate maintains stronger early text interaction than content-only gating. Its image-to-text attention fraction exceeds that of the baseline during the early high-noise steps before falling below it later. This combination motivates the fixed-weight test of the explicit time bias.

Reverse assigns the bias from position $49 - k$ to position k, matched-mean applies each channel's average over the 50-step schedule at every position, and zero removes the additive bias. With the TSGate checkpoint, generation seeds, and backbone timestep conditioning fixed, reverse reduces the raw DPG score by 2.614 points (95% CI $[-4.281, -0.994]$), matched-mean changes it by -0.929 points with an interval including zero, and zero changes it by -42.831 points. The reverse result demonstrates that correspondence between the learned bias and denoising stage matters.

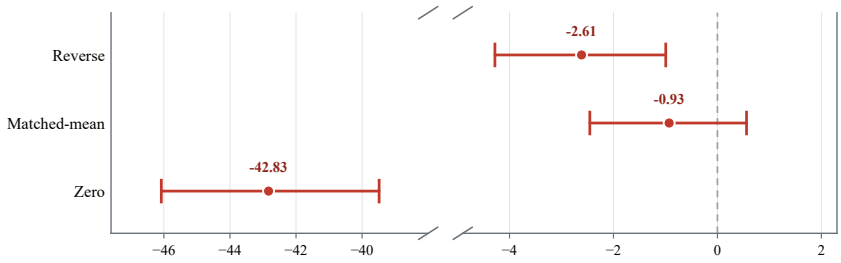


Figure 17: **Effects of time-bias interventions with model weights fixed.** On 200 raw D-mech sources, points show mean DPG changes relative to native inference; whiskers show paired source-bootstrap 95% CIs. The x-axis omits $(-38, -5)$ with equal unit spacing on the two segments; the dashed line marks zero.

F.1 Fixed-weight interventions on time biases

To test the temporal role of TSGate in Section F, we replace its explicit bias sequence while keeping the trained weights and backbone timestep conditioning fixed. The content branch remains active and responds to the hidden states of each intervened trajectory. For the complete schedule $t_0, \ldots, t_{K-1}$ with $K = 50$, the bias used at step k is

$$\begin{aligned} \text{native}: &\quad \widetilde{B}(t_k) = B_{\text{ts}}(t_k), \\ \text{reverse}: &\quad \widetilde{B}(t_k) = B_{\text{ts}}(t_{K-1-k}), \\ \text{matched-mean}: &\quad \widetilde{B}(t_k) = \frac{1}{K}\sum_{j=0}^{K-1} B_{\text{ts}}(t_j), \\ \text{zero}: &\quad \widetilde{B}(t_k) = 0. \end{aligned} \tag{12}$$

Layer and stream indices are omitted; each replacement is applied separately to every enabled layer, stream, and channel in Equation (5). Reverse preserves the bias values but reverses their order, testing alignment with the current stage. Matched-mean preserves the average bias and removes variation across steps. Zero measures reliance on the entire branch in the current model. Although zero recovers the content-only gate formula, it does not recover the independently trained Gated attention checkpoint: TSGate's weights are retained, and its content logits evolve along the intervened trajectory. Each intervention is paired with native inference using the same source prompt and generation seed.

F.2 Effects on raw and structured inputs

Table 10 reports nine contrasts on 200 paired sources. Reverse reduces the raw DPG score, with a confidence interval for the change that lies below zero, supporting the functional role of stage assignment. Zero produces a large reduction under both input constructions, showing reliance on the explicit branch in the trained model. Matched-mean preserves generation quality substantially better than zero. All table intervals are pointwise and use the bootstrap protocol in Appendix C.

G Input representations and prompt robustness

We examine whether TSGate's gains depend on prompt rewriting or additional scene information, complementing work on prompt optimization (Mañas et al., 2024). For this experiment, we sample 300 source prompts from DPG-Bench (Hu et al., 2024) (D-repr) and evaluate Baseline, Gated attention (Qiu et al., 2025), and TSGate on six input constructions, using the same sources and original scoring questions throughout.

Raw retains the original prompt. Semantic repeat duplicates that text without adding scene information. Neutral padding appends low-information text. Structured reorganizes the original description using headings while preserving its scene information; header only adds fixed headings. Full recaption rewrites the description into the structured format used for the main benchmark comparisons

Table 10: **DPG changes under time-bias interventions.** Δ is the score change relative to native inference. Structured denotes the content-preserving reorganization defined in Appendix G. The last three rows subtract the structured-prompt change from the raw-prompt change. All rows use 200 paired sources; intervals are pointwise 95% source-bootstrap CIs.

Input / contrast	Intervention	DPG Δ	95% CI
Raw	Reverse	-2.614	$[-4.281, -0.994]$
Raw	Matched-mean	-0.929	$[-2.455, 0.560]$
Raw	Zero	-42.831	$[-46.078, -39.485]$
Structured	Reverse	-0.919	$[-2.854, 0.948]$
Structured	Matched-mean	-0.027	$[-1.625, 1.619]$
Structured	Zero	-39.987	$[-43.457, -36.513]$
Raw minus structured	Reverse	-1.695	$[-4.329, 0.989]$
Raw minus structured	Matched-mean	-0.902	$[-3.131, 1.347]$
Raw minus structured	Zero	-2.844	$[-5.929, 0.284]$

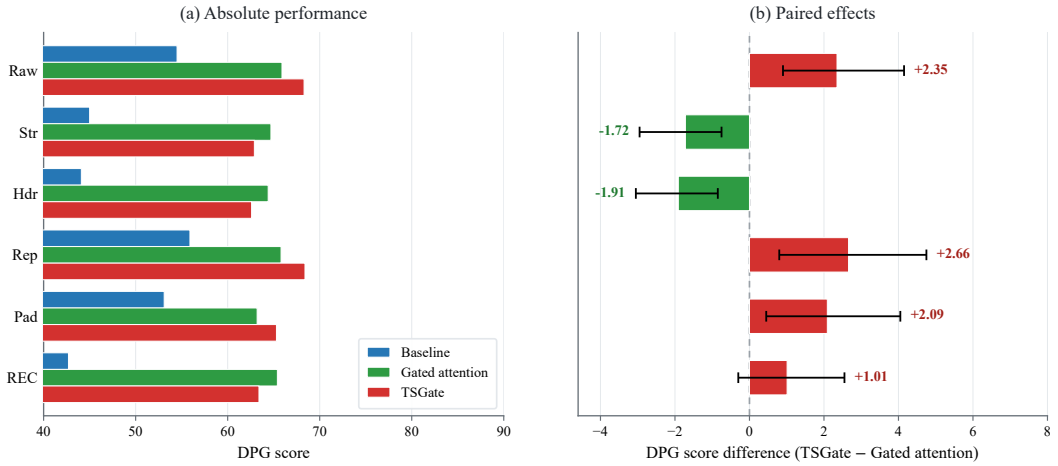


Figure 18: **Generation quality across six prompt representations.** Both panels use the same 300 sources. **(a)** Mean DPG scores. **(b)** Paired differences (TSGate minus Gated attention) with pointwise source-bootstrap 95% CIs; the dashed line marks zero. Raw: original; Str: structured; Hdr: header only; Rep: semantic repeat; Pad: neutral padding; REC: full recaption.

and may add or change scene facts, as illustrated in Appendix E. Thus, the Structured condition here and in Table 10 is distinct from the full-recaption structured prompts in the main text. These constructions probe sensitivity to repetition, appended text, organization, and rewriting; they are distinct input categories rather than successive levels of degradation.

Figure 18 shows that TSGate achieves higher mean DPG scores than Baseline under all six constructions. Relative to Gated attention, its gains are clearest on raw and semantic repeat, where the paired 95% confidence intervals lie above zero. The intervals for the other four constructions include zero, and the mean differences for structured and header only are negative. These results show that TSGate can improve over content-only gating without recaptioning or additional scene information, but do not establish an advantage under every input construction.