Hierarchical Semantic Fusion with Timestep-Aware Conditioning for Stable T2I Diffusion Transformers

Anonymous ACL submission

Abstract

We propose a semantic-aware framework for text-to-image diffusion models to improve alignment between text prompts and generated images. Existing methods that rely on cross-attention often struggle to preserve hierarchical semantic structures throughout denoising timesteps. To address these issues, we propose HSF–TAC, which integrates hierarchical semantic fusion with timestep-aware conditioning. It decomposes text embeddings into global S_g, local S_l, and relational S_r components and fuses them through timestep-aware routing, residual-gated cross-attention, and affine modulation. We further introduce an SNR-guided training strategy using linear, logit-normal, and heavy-tailed mode sampling to better focus learning on informative denoising steps. Experiments on the MJHQ-30K and MS-COCO datasets show that HSF–TAC improves prompt alignment, relational understanding, and stability across different guidance scales and sampling settings. The model also achieves strong performance on several evaluation metrics, including GenEval, DPG, FID, and CLIP score. These results indicate that improvements primarily stem from the proposed conditioning mechanism, not merely from a larger model.

1 Introduction

Diffusion-based generative models have become a leading approach for high-quality text-to-image synthesis (Black Forest Labs, 2024; Esser et al., 2024; Xie et al., 2025a). By progressively denoising random noise into visually coherent images, these models can generate diverse and realistic outputs from natural language prompts. The latent diffusion paradigm further improves computational efficiency by performing the denoising process in a compressed latent space, while maintaining strong perceptual quality. Recent models such as SDXL (Podell et al., 2023), PixArt-α (Chen et al., 2024b), PixArt-Σ (Chen et al., 2024a), and Stable Diffusion 3M (Esser et al., 2024) have improved text-to-image generation through larger backbones, transformer-based denoisers, improved text encoders, and refined training strategies. However, despite these advances, semantic alignment between complex prompts and generated images remains a challenging problem.

A major limitation of many existing methods is their reliance on time-invariant text conditioning. In such designs, the same textual representation is repeatedly injected into the denoising network across all timesteps (Esser et al., 2024; Xie et al., 2025a; Chen et al., 2024b). This overlooks the dynamic nature of the reverse diffusion process, where different denoising stages require different types of semantic information. At high noise levels, global scene structure is more important; at intermediate stages, relational and compositional information becomes critical; and at low noise levels, fine-grained attributes and local details are required (Li et al., 2024; Chen et al., 2024a). Therefore, using a fixed prompt representation throughout denoising may reduce alignment quality, especially for prompts involving multiple objects, attributes, and relationships.

To address this issue, we propose Hierarchical Semantic Fusion with Timestep-Aware Conditioning (HSF–TAC), a semantic-aware conditioning framework for latent text-to-image diffusion models. HSF–TAC decomposes textual information into three complementary branches: global semantics, local semantics, and relational semantics. These branches are dynamically fused through a timestep-aware routing mechanism that adapts their contributions according to the current denoising state. The fused representation is then injected into the denoising network using multi-branch cross-attention, residual gating, and feature-wise affine modulation.

In addition, we propose an SNR-guided training strategy to improve learning across different noise

regimes. Specifically, we investigate linear, logit-normal, and heavy-tailed timestep sampling strategies to emphasize informative denoising regions while preserving coverage across the full diffusion trajectory. Overall, HSF–TAC aims

to improve prompt alignment, compositional consistency, and generation stability without relying solely on model scaling.

2 Literature Review

Recent text-to-image research has moved beyond visual quality toward transformer-based architectures, structured semantic conditioning, and optimized training strategies(Chen et al., 2024b). Current studies show that controllability, semantic alignment, and compositional reasoning are now central challenges (Chen et al., 2024a). Transformer-based diffusion models such as MMDiT and diffusion transformers improve scalability and text–image interaction, but many still treat text conditioning as a static input across denoising steps, which limits fine-grained semantic control (Black Forest Labs, 2024; Esser et al., 2024; Meral et al., 2024).

Recent works address this issue through structured conditioning, such as splitting prompts into semantic units, using global text embeddings, and injecting information progressively during denoising (Greenberg, 2025; Qin et al., 2025; Rombach et al., 2022). However, compositional generation remains difficult, especially for prompts involving multiple objects, attributes, and spatial relationships (Wada et al., 2025; Pearl and Arefin, 2025). Existing solutions improve object-level control and scene grounding but do not fully solve relational reasoning.

Another important direction is timestep- and SNR-aware training. Studies show that prioritizing informative timesteps, using logit-normal sampling, and applying SNR-based strategies can improve training efficiency and generation quality (Ghosh et al., 2023). Overall, prior work suggests that strong text-to-image generation requires both adaptive semantic conditioning and noise-aware training (Esser et al., 2024; Black Forest Labs, 2024). However, few methods jointly integrate these two aspects, motivating the proposed HSF-TAC framework, which combines hierarchical semantic fusion, timestep-aware conditioning, and SNR-guided sampling in a unified model.

3 Methodology

3.1 Problem Formulation and Method Overview

In this section, we propose HSF-TAC, a latent text-to-image generation framework that improves semantic alignment during iterative denoising. While modern diffusion transformers achieve strong visual quality, most methods reuse the same text representation across all timesteps, ignoring that different denoising stages require different semantic information (Black Forest Labs, 2024; Chen et al., 2024b; Esser et al., 2024; **?**).

HSF-TAC addresses this limitation by decomposing text into global, local, and relational features and dynamically routing them according to the denoising timestep. Global context guides early high-noise stages, relational information supports intermediate composition, and local details refine low-noise outputs. The framework combines hierarchical semantic factorization, timestep-aware routing, residual gated fusion, and SNR-aware timestep sampling to enhance training efficiency, compositional consistency, and generation quality.

3.2 Preliminaries

We adopt the standard latent diffusion formulation, where an encoder $\mathcal{E}(\cdot)$ maps an image x into a latent representation $z_0 = \mathcal{E}(x)$. The forward diffusion process gradually corrupts z_0 into a noisy latent z_t:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad (1)$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. The denoising network ϵ_θ is trained to predict the injected noise:

$$\mathcal{L}_{diff} = E_{z_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, y)\|_2^2\right]. \quad (2)$$

To describe the noise level, we define

$$SNR(t) = \frac{\bar{\alpha}_t}{1 - \bar{\alpha}_t}, \quad \lambda_t = \log SNR(t). \quad (3)$$

The log-SNR λ_t indicates whether the current denoising stage is noise-dominated, transitional, or signal-dominated.

3.3 Hierarchical Semantic Factorization

A single prompt embedding may not provide suitable semantic guidance for all denoising stages. Therefore, HSF–TAC decomposes the text embedding H into global, local, and relational semantic

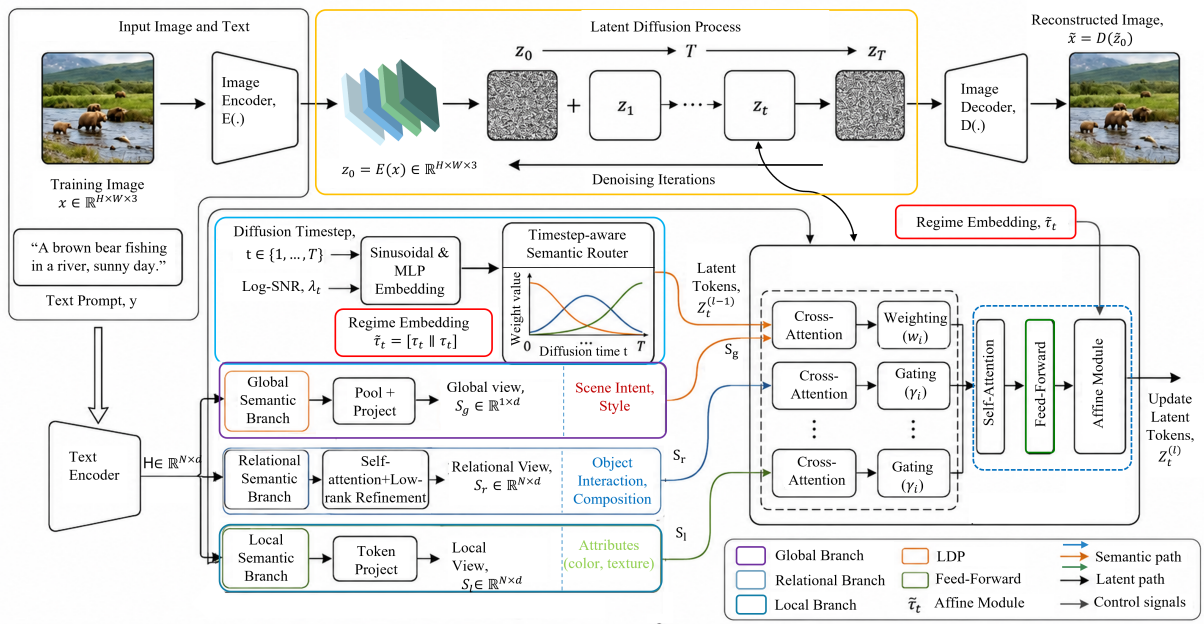

Figure 1: Overview of the proposed HSF-TCA architecture. Zoom in for better clarity

branches, as shown in Fig. 1. The global branch captures overall scene intent:

$$S_g = Pool(H)W_g. \tag{4}$$

The local branch preserves token-level attributes:

$$S_l = HW_l. \tag{5}$$

The relational branch models dependencies among prompt tokens:

$$S_r = Softmax\left(\frac{Q_r(H)K_r(H)^\top}{\sqrt{d}}\right)V_r(H). \tag{6}$$

To improve relational expressiveness with low cost, we apply a low-rank residual refinement:

$$S_r' = S_r + U(V^\top S_r), \quad k \ll d. \tag{7}$$

3.4 Hierarchical Branch Cross-Attention

For each transformer block l, branch-specific cross-attention is computed between latent tokens $Z_t^{(l)}$ and each semantic branch S_i, where $i \in \{g, l, r\}$:

$$A_i^{(l)} = Softmax\left(\frac{Q_i^{(l)}(Z_t^{(l)})K_i^{(l)}(S_i)^\top}{\sqrt{d}}\right)V_i^{(l)}(S_i). \tag{8}$$

The outputs are fused using timestep-aware routing weights $w_i(t)$ and residual gates $\gamma_i(t)$:

$$\tilde{Z}_t^{(l)} = Z_t^{(l)} + \sum_{i\in\{g,l,r\}} w_i(t)\gamma_i(t)A_i^{(l)}. \tag{9}$$

Here, $w_i(t)$ controls the relative importance of each branch, while $\gamma_i(t)$ controls the absolute injection strength.

3.5 Timestep-Aware Semantic Routing

HSF–TAC learns when each semantic branch should be emphasized during denoising. We augment the timestep embedding τ_t with log-SNR:

$$\tilde{\tau}_t = [\tau_t \| \lambda_t]. \tag{10}$$

The routing weights are then predicted as

$$w(t) = [w_g(t), w_l(t), w_r(t)]$$

$$= \text{Softmax}\left(MLP_w(\tilde{\tau}_t)\right), \tag{11}$$

with

$$\sum_{i\in\{g,l,r\}} w_i(t) = 1. \tag{12}$$

This allows global semantics to guide early high-noise stages, relational semantics to support intermediate composition, and local semantics to refine low-noise details.

Timestep-Aware Affine Conditioning. To provide direct channel-wise semantic modulation, we introduce affine conditioning. At each transformer block, scale and bias parameters are generated from the fused semantic representation $\tilde{s}_t$ and augmented timestep embedding $\tilde{\tau}_t$:

$$[\Gamma_t^{(l)}, B_t^{(l)}] = f_{aff}^{(l)}(\tilde{s}_t, \tilde{\tau}_t). \tag{13}$$

The latent feature is then modulated as

$$\hat{Z}_t^{(l)} = (1 + \Delta\Gamma_t^{(l)}) \odot \tilde{Z}_t^{(l)} + B_t^{(l)}. \tag{14}$$

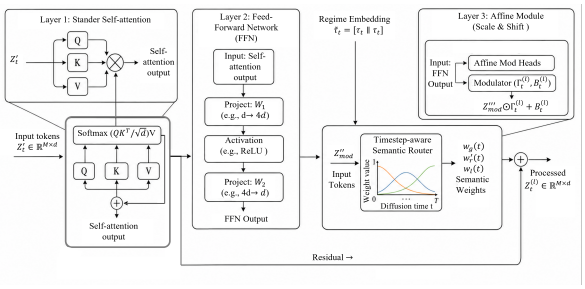

Figure 2: The framework combines self-attention, FFN, and timestep-aware affine modulation with semantic routing. Zoom in for better clarity.

This formulation preserves the original denoising behavior when the modulation is small, while allowing stronger semantic adaptation when required.

Feed-Forward Network. The feed-forward network (FFN) refines the conditioned latent tokens after semantic-aware cross-attention in Fig 2. Let $\tilde{Z}_t^{(l)} \in R^{M\times d}$ denote the conditioned latent tokens at layer l. The conditional FFN is defined as

$$\tilde{H}_t^{(l)} = \Gamma_{t,\text{ffn}}^{(l)} \odot \text{LN}\left(\tilde{Z}_t^{(l)}\right) + B_{t,\text{ffn}}^{(l)}. \tag{15}$$

$$\text{FFN}_{\text{cond}}\left(\tilde{H}_t^{(l)}\right) = W_2^{(l)}\phi\left(W_1^{(l)}\tilde{H}_t^{(l)}+b_1^{(l)}\right)+b_2^{(l)}. \tag{16}$$

$$Z_{t,\text{out}}^{(l)} = \tilde{Z}_t^{(l)}+\rho_t^{(l)}\,\text{FFN}_{\text{cond}}\left(\tilde{H}_t^{(l)}\right), \rho_t^{(l)} = \sigma\left({v_l}^\top\tilde{\tau}_t\right). \tag{17}$$

Here, $W_1^{(l)}$ and $W_2^{(l)}$ are learnable projection matrices, $\phi(\cdot)$ denotes a nonlinear activation function such as GELU, and $\text{LN}(\cdot)$ denotes layer normalization. Unlike a standard transformer FFN, this module is conditioned on both fused semantics and diffusion state, enabling adaptive nonlinear refinement within each denoising block.

Cross-Scale Semantic Scheduling. Diffusion transformers may operate across multiple token resolutions or latent stages (Podell et al., 2023). Since low-resolution features mainly encode global layout and high-resolution features emphasize fine details, semantic emphasis should vary across scales. Let $r \in \mathcal{R}$ denote the feature scale. We introduce a learned scale prior $\pi^{(r)} \in R_+^3$ and define

$$w^{(r)}(t) = \text{Normalize}\left(w(t) \odot \pi^{(r)}\right). \tag{18}$$

Here, $\text{Normalize}(\cdot)$ renormalizes the vector to sum to one. Thus, each scale follows the global routing policy $w(t)$ while biasing the semantic focus toward the most suitable branch. Low-resolution stages may emphasize S_g, whereas high-resolution stages may assign greater weight to S_l and S_r'.

3.6 Signal-to-Noise Ratio Sampling Strategies

Timestep sampling affects which denoising regimes receive more training updates (Xie et al., 2025b). Since uniform timestep sampling does not guarantee uniform coverage in noise space, we use SNR-aware sampling to better align training with diffusion dynamics. Under a general sampling distribution $p(t)$, the diffusion loss is written as

$$\mathcal{L}_{\text{diff}}^{(p)} = E_{t\sim p(t),z_0,\epsilon}\left[w(t)\,\|\epsilon-\epsilon_\theta(z_t,t,y)\|_2^2\right], \tag{19}$$

where $w(t)$ is an optional correction weight. For unbiased training, $w(t) = \frac{1/T}{p(t)}$, while for targeted optimization we set $w(t) = 1$.

LDM-Linear Sampling. The baseline strategy follows a linear variance schedule:

$$\beta_t = \beta_{\min}+\frac{t-1}{T-1}\left(\beta_{\max}-\beta_{\min}\right), \quad t = 1,\ldots,T. \tag{20}$$

The timesteps are sampled uniformly:

$$p_{\text{lin}}(t) = \frac{1}{T}. \tag{21}$$

Logit-Normal Sampling. Logit-normal sampling draws a normalized timestep $u \in (0,1)$ as

$$\xi \sim \mathcal{N}(\mu,\sigma^2), \quad u = \frac{1}{1+e^{-\xi}}, \quad t = 1+\lfloor(T-1)u\rfloor. \tag{22}$$

Its density is

$$p_{\text{LN}}(u) = \frac{1}{u(1-u)\sigma\sqrt{2\pi}}\exp\left[-\frac{\left(\log\frac{u}{1-u}-\mu\right)^2}{2\sigma^2}\right]. \tag{23}$$

This strategy emphasizes informative mid-SNR regions while retaining coverage of early and late denoising stages (Esser et al., 2024).

Mode Sampling with Heavy Tails. We introduce mode sampling with heavy tails (MSHT) in log-SNR space, where $\lambda = \log\text{SNR}(t)$. The density is defined as

$$p_{\text{MSHT}}(\lambda) = \eta\mathcal{N}\left(\lambda;\mu_m,\sigma_m^2\right)+(1-\eta)\tau_v\left(\lambda;\mu_m,s\right), \tag{24}$$

Table 1: Quantitative comparison of text-to-image generation models. The best result is shown in **bold**.

Res.	Model	Params(B)	Latency(s)	Speedup	GenEval↑	DPG↑	FID↓	CLIP↑
512^2	PixArt-α (Chen et al., 2024b)	0.6B	1.2	1.0×	0.48	79.58	6.14	27.55
	PixArt-Σ (Chen et al., 2024a)	0.6B	1.2	1.0×	0.52	84.86	6.34	27.62
	SANA 1.6B (Xie et al., 2025a)	1.6B	0.6	2.5×	0.64	**88.36**	**5.16**	28.19
	Ours	**0.8B**	**0.5**	**3.2**×	**0.65**	87.52	5.38	**29.10**
1024^2	SDXL (Podell et al., 2023)	2.6B	6.5	3.5×	0.55	88.00	7.95	27.57
	PixArt-Σ (Chen et al., 2024a)	0.6B	2.7	9.3×	0.54	85.78	6.34	28.26
	LUMINA-Next (Qin et al., 2025)	2B	9.1	2.8×	0.46	83.44	7.58	26.84
	PG 2.5 (Li et al., 2024)	2.6B	5.3	4.9×	0.56	88.50	6.09	29.13
	SD3 M (Esser et al., 2024)	2B	2.7	6.5×	0.63	**88.91**	11.92	27.83
	FLUX.1 [dev] (Black Forest Labs, 2024)	12B	23.0	1.0×	0.64	87.40	10.15	28.70
	Sana-1.6B (Xie et al., 2025a)	1.6B	**1.2**	**23.3**×	0.65	88.84	5.76	28.67
	FuseDiT (Tang et al., 2025)	2B	1.66	–	0.63	84.60	7.54	–
	Ours	**0.8B**	1.80	9.2×	**0.67**	88.89	**5.70**	**29.19**

Figure 3: Visual comparison with SOAT methods on prompts containing one, two, three, and four similar objects.

where τ_v is a Student-t distribution and η balances the central mode and heavy-tailed component. The timestep distribution is obtained by

$$p_{\text{MSHT}}(t) = p_{\text{MSHT}}(\lambda_t)\frac{dt}{d\lambda_t}. \tag{25}$$

MSHT focuses training on transition-critical regions while preserving exposure to both high-noise and low-noise regimes.

3.7 Complexity Analysis

The main additional cost comes from branch-specific attention projections and the affine modulation head. In the direct implementation, the parameter increase is approximately $\mathcal{O}(3d^2)$, for three semantic branches. This cost can be reduced by sharing key and value projections across branches while keeping branch-specific query projections. The low-rank relational refinement adds only $\mathcal{O}(dk), \quad k \ll d$. Thus, HSF–TAC improves semantic allocation without replacing the denoising transformer, enlarging the text encoder, or adding a second-stage refiner.

4 Experiments

Implementation Details. The model is trained with FP16 precision using AdamW, a learning rate of 1×10^{-4}, batch size 16, exponential moving average (EMA) updates, and 500K training iterations on the MJHQ-30K (Li et al., 2024) and MS-COCO datasets (Lin et al., 2014).

We evaluate performance using four metrics: FID (Heusel et al., 2017), CLIP score (Radford et al., 2021), GenEval (Ghosh et al., 2023), and DPG-Benc (Hu et al., 2024). These metrics measure image realism, text–image alignment, compositional accuracy, and prompt-following ability.

4.1 Main Results

Table 1 compares HSF-TAC with recent state-of-the-art text-to-image diffusion models on the MJHQ-30K dataset. The results show that HSF–TAC achieves a strong balance between generation quality, semantic alignment, and efficiency at both 512×512 and 1024×1024 resolutions. At 512×512, HSF-TAC obtains the highest GenEval score of 0.65 and the best CLIP score of 29.10, while using only 0.8B parameters and achieving the lowest latency of 0.5 s. Although SANA (Xie et al., 2025a) 1.6B reports slightly better DPG and FID scores, HSF-TAC remains highly competitive while being smaller and faster.

At 1024×1024, HSF–TAC further demonstrates strong overall performance. It achieves the best FID of 5.70 and the highest CLIP score of 29.19, while also reaching the best GenEval score of 0.67. Its DPG score of 88.89 is close to the best-performing model. These results indicate that HSF–TAC can match or outperform larger models while requiring fewer parameters and lower inference cost.

On MS-COCO, HSF-TAC also improves both image quality and text–image alignment. Compared with SDXL (Podell et al., 2023), it reduces FID from 7.38 to 6.40 and increases CLIPScore from 0.2913 to 0.2967. These gains suggest that timestep-aware conditioning helps the model learn coarse structure and fine details more effectively.

Visual comparisons. In Fig 3, shows visual comparisons further confirm the effectiveness of HSF–TAC. For prompts containing multiple similar subjects, competing models often miss objects, confuse attributes, or mix semantic identities. In contrast, HSF–TAC better preserves subject identity, counting accuracy, and relational consistency. These improvements mainly come from hierarchical semantic decomposition, timestep-aware conditioning,

and multi-branch fusion, which allow the model to select the most relevant semantic information at each denoising stage.

4.2 Stable High-Resolution Fine-Tuning

For high-resolution training, we first apply data filtering to remove explicit, low-quality, and duplicate samples. We also precompute autoencoder latents and text features to improve training efficiency. To stabilize BF16 fine-tuning, we apply

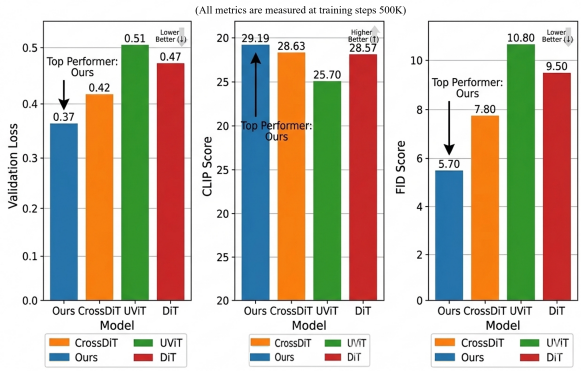

Figure 4: Comparison of transformer-based diffusion backbones.

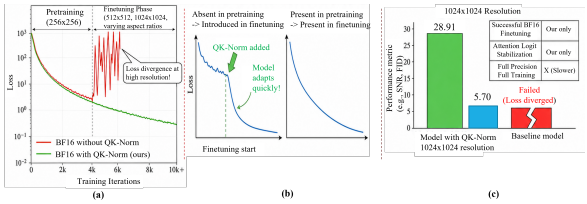

Figure 5: Training stabilization of high-resolution fine-tuning via QK-normalization. (Zoom in for a clearer view.

QK-normalization before attention:

$$\tilde{Q}^{(l)} = \gamma_Q^{(l)} \frac{Q^{(l)}}{\sqrt{\mathrm{mean}((Q^{(l)})^2) + \delta}}, \quad \tilde{K}^{(l)} = \gamma_K^{(l)} \frac{K^{(l)}}{\sqrt{\mathrm{mean}((K^{(l)})^2) + \delta}}. \tag{26}$$

The local attention is then computed as

$$A_{\mathrm{loc}}^{(l)} = \mathrm{Softmax}\left(\frac{\tilde{Q}^{(l)} \tilde{K}^{(l)\top}}{\tau_l}\right) V^{(l)}. \tag{27}$$

Timestep Scaling. When image resolution increases, the denoising task can become easier because the effective SNR grows with the number of pixels. For an image with n pixels, the effective SNR is

$$\mathrm{SNR}_n = \frac{\alpha^2}{\sigma^2} n. \tag{28}$$

To maintain consistent difficulty when scaling from n to m pixels, we shift the log-SNR as

$$\lambda_m = \lambda_n - \log\left(\frac{m}{n}\right). \tag{29}$$

This adjustment introduces higher effective noise during high-resolution training and encourages the model to learn stronger structural and textural representations.

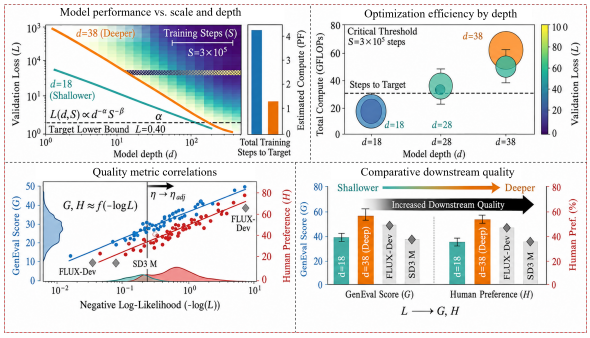

Figure 6: Models improving loss, efficiency, and quality.

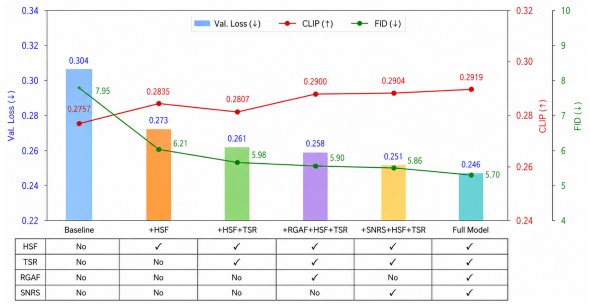

Figure 7: Ablation analysis of the proposed modules.

Enhancing Features. We extend HSF–TAC to high-resolution text-to-image synthesis by improving the backbone, autoencoder, prompt quality, and fine-tuning strategy in Fig **??**. We compare HSF–TAC with transformer-based diffusion backbones, including Cross-attn DiT, UViT, and Vanilla DiT. Across validation loss, CLIP score, and FID, HSF–TAC shows faster convergence, stronger text–image alignment, and better visual fidelity.

Impact of Model Size. Model performance follows a scaling trend where validation loss decreases with model depth d and training steps S:

$$L(d, S) \propto d^{-\alpha} S^{-\beta}. \tag{30}$$

Lower validation loss correlates with improved GenEval scores and human preference, indicating that stronger optimization leads to better semantic alignment and visual quality.

4.3 Ablation Analysis

The ablation study evaluates the contribution of each proposed module, as shown in Fig. 7 and Table **??**. Starting from the baseline (Podell et al., 2023), adding hierarchical semantic fusion (HSF) improves image quality and semantic alignment by reducing FID and increasing CLIP score. Timestep-aware semantic routing (TSR) further stabilizes training and lowers the validation loss. Residual-gated affine fusion (RGAF) and SNR-guided sampling (SNRS) provide additional improvements, indicating better feature aggregation and diffusion stability. The full model achieves the best overall performance, with the lowest validation loss of 0.246, the highest CLIP score of 0.2919, and the lowest FID of 5.70. These results confirm that all proposed modules are complementary and jointly improve generation quality , as shown in Fig. 9.

Figure 8: Failure case analysis.

4.4 Failure Case Analysis

As shown in Fig. 8, HSF–TAC may still fail when semantic cues are not properly balanced across denoising stages. Some outputs preserve global structure but lose fine details, causing distorted faces, unclear object boundaries, or inconsistent textures. Other cases show weak relational coherence, especially in complex street or architecture scenes. These results suggest that the model can overemphasize global semantics while underusing local and relational cues during later refinement.

5 Conclusion and Future Work

We proposed HSF–TAC, a latent diffusion transformer framework that improves text-to-image generation through adaptive semantic conditioning. The method decomposes prompts into global, local, and relational branches and fuses them using timestep-aware routing, residual-gated fusion, affine modulation, and SNR-aware sampling.

Experiments on MJHQ-30K and MS-COCO show that HSF–TAC improves prompt alignment, relational understanding, and generation stability without relying only on larger model size. Future work will explore finer semantic branches, better SNR-routing optimization, and extensions to multilingual, LLM-based, and video diffusion models.

References

Black Forest Labs. 2024. FLUX.1 [dev]. Official project page.

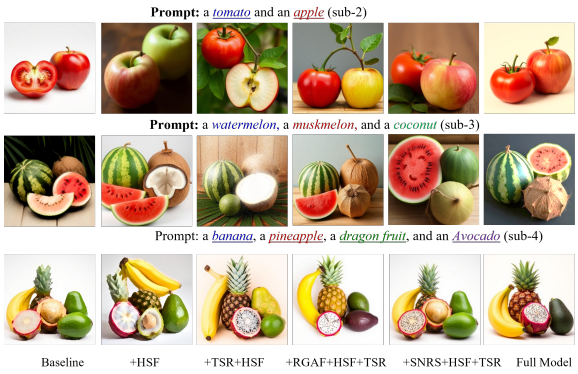

Figure 9: Qualitative ablation on prompts with two to four subjects.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024a. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Available online at Hugging Face Papers and GitHub.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024b. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederik Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *Computer Vision and Pattern Recognition (CVPR)*.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. GitHub repository.

Oren Greenberg. 2025. Demystifying flux architecture. *Medium AI Research Blog*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. 2024. Ella: Equip diffusion models with llm for enhanced semantic alignment. Project page and technical report.

Daiqing Li, Arash Kamko, Elham Akhgari, Ali Sabet, Lian Xu, and Sushant Doshi. 2024. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. Playground AI technical report.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer.

T. Han Salih Meral, Erkut Simsar, Federico Tombari, and Pinar Yanardag. 2024. Conform: Contrast is all you need for high-fidelity text-to-image diffusion models. Project page.

S. N. Pearl and M. S. Arefin. 2025. Developing a transformer-based autoencoder model for sentence embedding. In *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE.

Dustin Podell, Zachary English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *SIGGRAPH Asia 2023 Conference Papers*. ACM.

Qinghao Qin, Liang Zhuo, Yutong Xin, Rui Du, Zhe Li, Bowen Fu, and 1 others. 2025. Lumina-image 2.0: A unified and efficient image generative framework. GitHub repository and technical report.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Bin Tang, Boyang Zheng, Xinyi Pan, Sayak Paul, and Saining Xie. 2025. Exploring the deep fusion of large language models and diffusion transformers for text-to-image synthesis. *Transactions on Machine Learning Research*.

Takuma Wada, Yuki Hirakawa, Ryo Shimizu, Takuya Kawashima, and Yutaro Saito. 2025. Static word embeddings for sentence semantic representation. *Natural Language Engineering*.

Enze Xie, Junsong Chen, Jun Chen, Haotian Cai, Hang Tang, Yuxiang Lin, and 1 others. 2025a. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE.

Enze Xie, Junsong Chen, Yutong Zhao, Jiahui Yu, Lei Zhu, Conghui Wu, and 1 others. 2025b. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *arXiv preprint*.