TexTailor: Inference-Time Textual Guidance Tailoring for Multimodal Diffusion Transformers

Binglei Li[1,2], Mengping Yang[1,3†], Zhiyu Tan[1,3], Junping Zhang[1*], and Hao Li[1,2,3*]

[1] Fudan University, Shanghai 200433, China
[2] Shanghai Innovation Institute, Shanghai 200030, China
[3] Shanghai Academy of AI for Science, Shanghai 200030, China
{blli24, zytan24}@m.fdu.edu.cn,
{mengpingyang, jpzhang, lihao_lh}@fdu.edu.cn

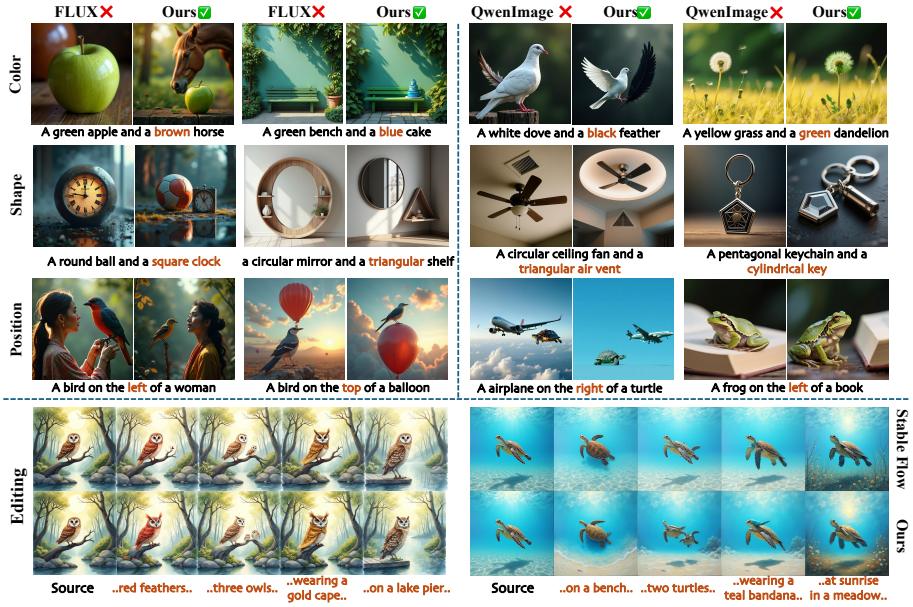


Fig. 1: Visual comparison of text-to-image and editing results, demonstrating that our TexTailor yields better text alignment across semantic attributes.

Abstract. Recent breakthroughs of transformer-based diffusion models, particularly with Multimodal Diffusion Transformers (MMDiT) driven models like FLUX and Qwen Image, have facilitated thrilling experiences in visual generation. However, these models rely only on the interactions between textual conditions and visual features to produce semantically aligned images. Once the interactions fail to reflect the nuanced compositional structure of the prompt, the generated images might be unsatisfactory. Thus, a comprehensive understanding of how different blocks

† Project Lead.
* H. Li and J. Zhang are the corresponding authors.

and their interactions with textual conditions is crucial for better understanding the intrinsic attributes and for enhancing their interactions accordingly to strengthen the prompts adherence. In this paper, we first develop a systematic pipeline to comprehensively investigate each block's functionality by *removing*, *disabling*, and *enhancing* textual hidden-states at corresponding blocks. Our analysis reveals that 1) semantic information appears in earlier blocks and finer details are rendered in later blocks, 2) removing specific blocks is usually less disruptive than disabling text conditions, and 3) enhancing textual conditions in selective blocks improves semantic attributes. Building on these observations, we propose TexTailor, a novel inference-time method for tailoring block-wise textual guidance. Our approach not only improves text-image alignment but also enables a range of downstream applications, including precise editing and inference acceleration. Extensive experiments demonstrated that our method outperforms various baselines and remains flexible across text-to-image generation, image editing, and inference acceleration. Our method improves T2I-Combench from 56.92% to 63.00% and GenEval from 66.42% to 71.63% on SD3.5, without sacrificing synthesis quality. These results advance understanding of MMDiT models and provide valuable insights to unlock new possibilities for further improvements. Our code is available at https://github.com/phil329/TexTailor.

Keywords: MMDiT · Block-wise Analysis · Inference-time Tailoring

1 Introduction

Diffusion models [14, 38, 39], especially diffusion transformers (DiT) [4, 32], have become the de-facto paradigm for real-world applications across various domains, including text-to-image [7, 34] and text-to-video generation [29, 47, 54], unlocking unprecedented experiences for AI generated content. In particular, the most recent approaches such as Stable Diffusion 3 [9], FLUX [19], and Qwen-Image [50] further advance the synthesis quality via incorporating the flow matching [22, 25] training objective and the top-performing multimodal diffusion transformer (MMDiT) architecture [9, 19]. Specifically, MMDiT concatenates vision and textual tokens and performs joint self-attention to facilitate a seamless textual-visual interaction between these modalities.

Despite their remarkable success, MMDiT-based models produce images based on interactions between textual representations and visual hidden states only. As shown in Fig. 2, if the textual-visual interactions fail to faithfully render semantic details of given instructions, the generated images might be unsatisfactory. Therefore, it is crucial to investigate the intrinsic mechanisms within MMDiT-based models, enabling better understanding on the internal interactions of MMDiT. Following this philosophy, Stable Flow [1] detected vital blocks by bypassing each block, and TACA [28] proposed a timestep-aware attention weighting mechanism to balance multimodal interactions. FreeFlux [48] and E-MMDiT [36] analyzed MMDiT's attention mechanism by shifting RoPE and decomposing attention metrics, respectively. Semantic Routing [20] and Hyper-

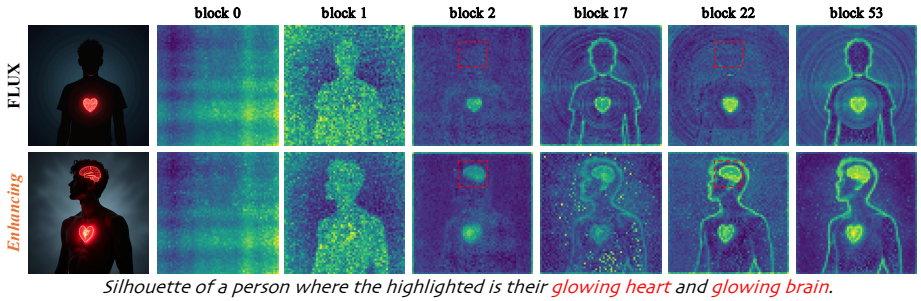


Fig. 2: Textual-Visual interaction attention maps across different blocks.

Align [53] respectively employ linear fusion and hypernetwork-generated low-rank adaptation weights to modulate text-visual interactions within diffusion transformers. However, existing studies primarily focus on isolating or manipulating individual aspects, overlooking the synergistic effects that arise from the complex interactions across different blocks and modalities. That is, it remains unclear how different internal MMDiT blocks interact with textual representations and how they collaborate with each other to produce high-quality and coherent outputs. Consequently, a deeper and more detailed analysis of understanding how MMDiT blocks collectively contribute to rendering sophisticated outputs would not only enrich our understanding of MMDiT models but also open avenues for refining their synthesis quality and inference efficiency. For instance, by identifying which blocks control specific attributes (*e.g.,* color, shape, spatial relationships), we can revise the corresponding blocks accordingly (as illustrated from the results in Fig. 1).

To identify each block's detailed role and functionality, we first develop a systematic automatic pipeline to probe the internal cooperation of MMDiT blocks and their interactions with text conditions. Our pipeline involves two main stages: For the first stage, we construct a dedicated prompt set covering representative attributes including color, shape, and spatial relationship. These prompts are used to generate images from three popular MMDiT-based models (SD3.5, FLUX, Qwen Image). Then, we quantify the influence of block-wise interactions on the output images with perceptual (*i.e.,* DINOv2 [31]) and semantic scores (*i.e.,* CLIP Score [33]), as well as VQA results from general MLLMs (*i.e.,* Qwen-VL [2]) Specifically, we probe block-wise interactions and collaboration via: 1) *removing* specific blocks to assess their individual contributions; 2) *disabling* block-level textual conditions to test semantic understanding; and 3) *enhancing* textual hidden-states of different blocks to investigate their potential to refine the coherence and detail of synthesized outputs. Through these analysis, we reveal *several insightful findings*: First, semantic information appears in earlier blocks, while fine-grained details are rendered in later blocks. As illustrated by the enlarged details in Fig. 4, different attributes exhibit similar behaviors across various blocks, suggesting that a unified enhancement strategy

can be applied to all attributes. Second, removing certain blocks is less disruptive than disabling corresponding textual conditions, indicating MMDiT-based models' stronger reliance on conditional guidance and their robustness to removing individual blocks. Last but not least, enhancing textual representations of selective blocks could improve overall text alignment without compromising synthesis quality. These insights provide a better understanding towards the efficacy and interactions of different components within the MMDiT architecture, guiding further optimization and improvements in various applications.

Capitalizing on these observations, we propose TexTailor, a novel inference time textual guidance enhancement framework to improve the text alignment of text-to-image, facilitate image editing, and accelerate model inference within MMDiT-based models. Concretely, after identifying each block's contribution to the specific semantic attributes of output images, we can strategically enhance the text-visual interactions of these blocks to optimize the text alignment. Moreover, a token-level tailoring strategy is further designed to improve the interactions of key textual tokens of specific attributes. As shown in Fig. 2, the textual-visual interactions after enhancing present improved adherence of given prompts, leading to more satisfactory output. Regarding editing tasks, we can similarly attach higher importance to specific blocks that manage certain semantic attributes, such as "color" and "count", ensuring that edits are accurately and effectively performed. Additionally, we could accelerate the inference process by skipping blocks that are less critical for semantic understanding, thus streamlining computations while preserving synthesis quality. Together, our framework facilitates efficient, precise, and generalizable model performance across different tasks without requiring additional training. Extensive results demonstrate that our method significantly advances the model performance across various baselines (SD3.5, FLUX, and Qwen-Image), benchmarks (GenEval [11], T2I-Combench [17]) and metrics (CLIP-Score [33], Human-Preference Score [51], Aesthetic Score [35]), as well as different tasks (generation, editing, and acceleration), consistently showing the effectiveness and generalizability of our method.

We summarize our primary contributions as: 1) **A systematic framework to probe block-wise interactions of MMDiT-based models**. We systematically investigate the internal interactions across blocks and modalities within several MMDiT-based models, offering valuable and consistent insights to guide further improvements. 2) **A novel inference-time framework for tailoring textual guidance**. We propose TexTailor to enhance the interactions between textual conditions and visual features within different blocks, thereby improving the text alignment of generated images. Moreover, our method can be applied to other applications including precise editing and inference acceleration, fully unlocking the potential of baseline models. 3) **Extensive evaluations across multiple baseline models and diverse benchmarks** for various tasks consistently demonstrate the effectiveness and generalizability of our approach in advancing model performance.

2 Related Work

Diffusion Transformers. Diffusion Transformers [32] have become the dominant paradigm for high-fidelity image and video generation, which adopt transformer [45] architecture as the main backbone, demonstrating superior scalability and training efficiency compared to previous UNet-based [8, 14, 15] models. Recent variants, such as open-sourced SD3 [9], FLUX [19], Qwen Image [50], Hunyuan Image [44] and Hunyuan Video [43], and commercial models like Seedream series [10,12], Sora [30], Imagen3 [3], further advance text-to-image/video to an unprecedented level with the top-performing multimodal diffusion transformer (MMDiT) architecture. Besides scaling the MMDiT-based models, many efforts have also focused on accelerating the iterative denoising process [26,27,40], controlling the results [42, 52, 56], editing the outputs [1, 48], *etc.*

Understanding and Improving Diffusion Models. Numerous prior works proposed various techniques to analyze the roles of different components of UNet-based diffusion models. For instance, P2P [13] showed that cross-attention layers are essential for rendering the spatial layout, MasaCtrl [5] and [23] demonstrated that self-attention are more important for preserving the geometric and shape details. FreeU [37] and PBC [57] respectively analyzed the functionality of skip connections and position encoding mechanism in diffusion UNet. Further, Yi *et al*. [55] investigated the mechanism of text prompts and Williams *et al*. [49] developed a unified framework for analysing the UNet architectures. P+ [46] optimizes extended per-layer text embeddings to improve prompt fidelity. Attend-and-Excite [6] amplifies cross-attention maps for semantically relevant tokens at inference time. However, the understanding of different components within MMDiT-based models remains underexplored and it is crucial to gain a holistic understanding of these components to advance the field. Existing arts tried to identify the contribution of different layers [1], rotary position embeddings (RoPE) [48], and attention [36], yet they usually focus on specific applications like editing and lack a systematic analysis of the internal interplay of components within models. TACA [28] indicated an imbalanced issue in the cross-model attention and ameliorated it with a timestep-aware re-weighting scheme. Semantic Routing [20] explored the multi-layer LLM feature weighting for diffusion transformers, which trains a linear fusion module to modulate the text-visual interactions. HyperAlign [53] proposed a hypernetwork to generate LoRA weights to modulate the diffusion model's generation operators. Nevertheless, none of the approaches provides a holistic view of how these components jointly influence the model's performance.

3 Systematic Analysis of Block-wise Interactions

3.1 Preliminaries

Diffusion Models. Diffusion models (DMs) involve a forward process and a reverse generation process. During the forward process, random noises are grad-

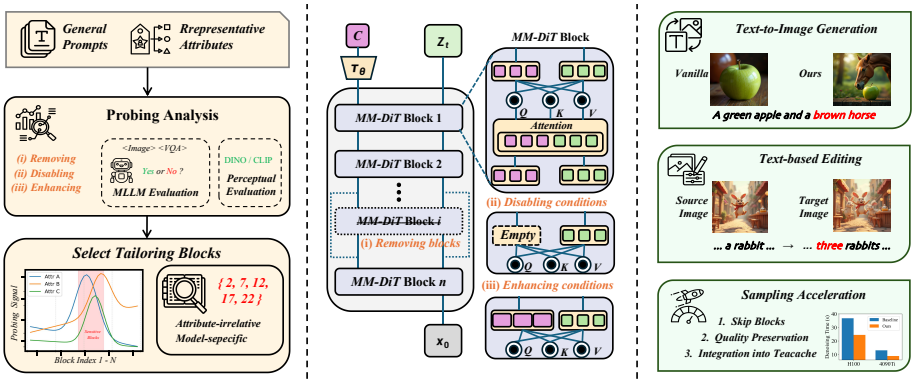


Fig. 3: Analysis and framework overview. (a) We propose TexTailor, a training-free framework to tailor the inference-time text-visual interactions. (b) Our probing analysis consists of three key operations: *removing*, *disabling*, and *enhancing*. The tailoring blocks are chosen from the high-signal region, which can be applied to various downstream tasks, including T2I generation, image editing and acceleration (c).

ually added to data ($\mathbf{x}_0 \sim q(x)$) across $t \sim (1...T)$ timesteps:

$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\epsilon_{t-1}. \tag{1}$$

In the reverse generation process, the model iteratively reconstruct the original data following a trajectory opposite to the forward process:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)), \tag{2}$$

where μ_θ and Σ_θ are learnable mean and covariance, respectively.

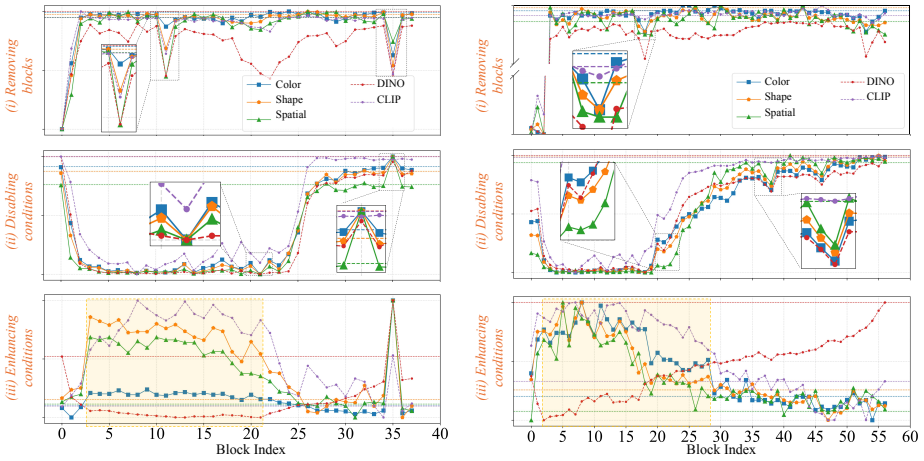
MMDiT-based Models. MMDiT-based models, pioneered in SD3 [9], leverage a joint multimodal architecture to process text embeddings $\boldsymbol{c} \in \mathbb{R}^{N_{\boldsymbol{c}} \times D}$ and visual features $\boldsymbol{x} \in \mathbb{R}^{N_{\boldsymbol{x}} \times D}$ in a unified attention operation by concatenating them as $h_{in} = [\boldsymbol{c}; \boldsymbol{x}] \in \mathbb{R}^{(N_{\boldsymbol{c}}+N_{\boldsymbol{x}}) \times D}$. This sequence is then processed by multiple MMDiT blocks with a joint self-attention layer:

$$\text{Attention}(Q, K, V) = softmax(QK^T/\sqrt{d_k})V, \tag{3}$$

where Q, K, V denote the concatenated query, key and value of text and image tokens. The MMDiT blocks serve as the core modules enabling effective and fine-grained integration of textual-visual information during denoising.

3.2 Understanding Block-wise Interactions of MMDiT

In this part, we develop a systematic framework to automatically investigate block-wise interactions and their influence on the representative semantic attributes (*i.e.,* color, shape, spatial relationships). As shown in Fig. 3 (b), our



(a) Probing analysis of SD3.5-Large. **(b)** Probing analysis of FLUX.1-Dev.

Fig. 4: Block-wise analysis results across various MMDiT-based models on representative attributes. We observe that different attributes exhibit similar block-wise behaviors, which indicates that the probing analysis is effective and generalizable. We select tailoring blocks from the common high-signal region (yellow box) for all attributes. Zoom in for details and see supplementary for the results of Qwen Image.

study involves three key operations: 1) *removing* specific blocks to probe each block's individual contribution; 2) *disabling* textual conditions of different blocks to evaluate their reliance on semantic guidance; 3) *enhancing* textual guidance of certain blocks to investigate their potential to refine coherence and detail. Specifically, we amplify the text condition hidden states by a factor of 2 as $\boldsymbol{c} \rightarrow 2\boldsymbol{c}$, to investigate each block's latent capacity for assimilating semantic information. Regarding the disabling operation, we mute the textual hidden states via attaching an empty tensor with $torch.zeros\_like(c)$.

We expect that the probing analysis could generalize to various attributes and categories, so we construct a prompt dataset from the widely adopted HPDv2 [51] and COCO [21] datasets. It comprises 633 diverse prompts across three representative attributes: "color", "shape", and "spatial relationships". For each prompt, we perform *removing*, *disabling* and *enhancing* on SD3.5-Large [41], FLUX.1-Dev [19], and Qwen Image [50] models, operating on one block at a time. For the probing analysis, we evaluate generated images using Qwen2.5-VL-72B [2] via question-answering pairs on the prompts and the generated images. Further, we evaluate perceptual (DINOv2 [31]) and semantic similarities (CLIP Score [33]) between images from our modified and original models to quantify the effect of our block-wise manipulations. Notably, despite the limited number of prompts per attribute, repeated sampling with 5 fixed seeds produces consistent and reliable results (see supplementary for more details).

The analysis results on the representative attributes across SD3.5 (38 blocks), FLUX (57 blocks) are shown in Fig. 4. For each subfigure, we plot the DINOv2

and CLIP score curves, along with the quantitative curves of performing our analysis method, *i.e.,* removing (1st row), disabling (2nd row), enhancing (3rd row), on three representative attributes. Despite testing on different models, we could consistently observe several interesting findings from these results.

Findings 1: Removing less critical blocks tends not to significantly impact overall performance. Fig. 4a (top) and Fig. 4b (top) illustrate the impact of *removing* different blocks. We could observe that all models are sensitive to the removal of the earlier $(0-5)$ and late blocks, leading to a significant drop in synthesis performance. We attribute this sensitivity to the critical roles of early blocks in initializing inputs and of late blocks in refining details for the final output. By contrast, removing blocks from the middle layers generally results in a smaller impact, as also evidenced by both the synthesis performance and the DinoV2 and CLIP scores. Such observation indicates that these blocks might be less critical for maintaining the fidelity and coherence of the generated outputs. Accordingly, we could remove some blocks to optimize model efficiency without compromising the synthesis quality. Furthermore, we find that removing some blocks (the $18th$ block in FLUX, the $35th$ block in SD3.5) leads to a significant drop in performance, which may be the boundary between the dual-stream and single-stream blocks or the vital blocks for the semantic rendering.

Findings 2: Disabling textual conditions is more disruptive than removing specific blocks. Fig. 4a (middle) and Fig. 4b (middle) reveal that disabling textual conditions, especially in the earlier blocks $(0-20)$, causes a more pronounced degradation in the synthesis performance compared to merely removing specific blocks. That is, textual conditions play a crucial role in guiding the models' generative process. Moreover, we can see that the results of disabling conditions of late blocks are less detrimental to the overall performance, particularly the CLIP Score, suggesting that these blocks are specialized in refining details and the core semantics are rendered by the earlier blocks. Interestingly, we observe that *disabling* in the dual-stream part of FLUX causes larger degradation. In contrast, Qwen Image is more robust, whose performance is mainly affected by the first and last few layers, while disruptions in intermediate layers can be mitigated by later modules.

Findings 3: Enhancing textual conditions on certain blocks could improve the synthesis performance. Fig. 4a (bottom) and Fig. 4b (bottom) show the enhancing results. Though the simple $\times 2$ operation may not yield optimal results, enhancing textual conditions on certain blocks can improve the synthesis performance. Remarkably, for all three attributes, all models show performance improvements compared with the original baseline, despite Qwen Image showing less improvement due to its strong baseline. Interestingly, CLIP Score shows the same trend as the attributes, indicating that enhancing textual conditions on certain blocks could improve the text-visual alignment. To our knowledge, this observation has never been documented in existing literature. We highlight the blocks in the yellow region that all the representative attributes commonly exhibit significant performance improvements, which are the tailoring block ranges for each attribute in the latter sections. In return, one could ma-

nipulate specific attributes (e.g., "color", "shape", "count" and "position" in Fig. 1 and Fig. 9) by altering textual information at the selected tailoring blocks.

Overall, our analysis provides a comprehensive investigation of the block-wise capabilities and their interactions with textual conditions, yielding several interesting insights on how different blocks contribute to the output. These findings contribute to a better understanding of MMDiT-based models, offering valuable perspectives that could facilitate further enhancements and optimizations.

4 Methodology

4.1 Motivations

The diffusion models primarily rely on the text prompts to guide the generation process. Attention mechanisms in MMDiT models [9, 19, 50] are designed to capture the interactions between text and image features, which are crucial for the generation process. Fig. 2 presents attention maps from FLUX and our enhanced results, where our method strengthens attention to the relevant text phrase "glowing brain", improving text-visual alignment. Moreover, different blocks play different roles in the generation process, and our enhancing method can help the model to focus on the relevant text tokens.

Our analysis demonstrates that tailoring only a few specific blocks is sufficient to strengthen the text-image interactions. This tailoring strategy not only effectively modulates the attention mechanism, but also mitigates the problem of imbalanced visual and textual tokens [28]. Our work offers the community a fresh perspective and actionable findings that go beyond the traditional block-wise analysis. Although conducted on the three representative attributes, the tailoring blocks could generalize to other attributes.

4.2 Tailoring Textual Guidance for Denoising

We propose a straightforward, training-free method to tailor text-visual interactions within blocks by capitalizing on their pivotal roles, which is highlighted in the yellow boxes in Fig. 4(bottom panel). Specifically, we enhance the hidden states of textual conditions in these vital blocks $\mathcal{S} = \{s_1, s_2, \ldots, s_m\}$ by a factor of $\lambda(l)$ using the following equation:

$$\boldsymbol{c}_{enh}^{(l)} = (1 - M) \odot \boldsymbol{c}^{(l)} + \lambda(l) \cdot M \odot \boldsymbol{c}^{(l)}, \quad \forall l \in \mathcal{S}. \tag{4}$$

where $\boldsymbol{c}^{(l)}$ is the original textual hidden states of block l and $\odot$ is element-wise multiplication. $\lambda(l)$ can be a constant or a block-dependent function. M is a binary mask that denotes corresponding indices of enhanced textual tokens.

Full sentence tailoring. Formally, we set M to be a full one vector to tailor the entire textual condition. The equation becomes $\boldsymbol{c}_{enh}^{(l)} = \lambda(l) \odot \boldsymbol{c}^{(l)}$, which is suitable for all prompts. Full sentence tailoring can avoid the problem of difficulty locating the specific token indices, such as complex prompts containing multiple attributes, which cannot obtain an accurate mask M for token-level tailoring.

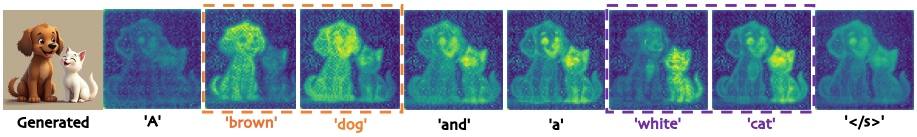


Fig. 5: Attention map of different tokens. The target phrases well match the attention map of the corresponding tokens, highlighted in dashed boxes. We can enhance specific tokens to boost their impact on the attention map.

Token-level tailoring. To further improve the semantic understanding capability of certain blocks on specific attributes, we introduce token-level enhancement to amplify key textual tokens. As shown in Fig. 5, we can determine the mask M by locating the token indices of the target phrases within the tokenized prompt. Such an operation ensures that critical semantic attributes receive greater emphasis. Specifically, some attributes, such as "color", "shape", and "count", are easy to locate the corresponding mask M, on which we conduct token-level tailoring experiments. Thus, this allows for a better understanding of textual conditions, emphasizing key semantic attributes within the model.

5 Experiments

5.1 Experiment Setups

Implementation Details. We apply our method on several state-of-the-art MMDiT-based models, namely SD3.5-Large [41], FLUX.1-Dev [19] and Qwen Image [50]. Based on the observations from our systematic analysis in Sec. 3, we tailor the textual conditions during denoising on the selected blocks according to the attribute in Tab. 1. The tailoring parameter $\lambda(l)$ in Eq. 4 is set to 1.5 unless otherwise specified. We compare our method on text-to-image generation with TACA [28], which attaches a timestep-aware importance on the textual conditions within the attention. We report the full sentence tailoring and token-level tailoring results on the T2I-CompBench and GenEval benchmarks. During inference, we sample 1024×1024 images using the EDM [18] sampler for 28 denoising steps, and use the official default settings for the other inference parameters.
Evaluation Metrics. We evaluate our method on the T2I-CompBench [16] and Geneval [11] benchmarks, which are widely adopted for text-to-image alignment. Additionally, we utilize Aesthetics score [35] and HPSv2 [51] to evaluate the overall image quality, ensuring that our method does not negatively impact the synthesis quality of the original models. We follow the official guidance for evaluation. All experiments are carried out on NVIDIA 4090 and H100 GPUs.
Selected Blocks for Tailoring. In the Sec. 3, we systematically investigate the block-wise interactions in MMDiT-based models, and identify similar patterns across three representative attributes. In the Fig. 4, the representative attributes exhibit similar block-wise behaviors, as highlighted in the enlarged details. This indicates that the selected tailoring blocks can be generalized to other attributes.

Table 1: Blocks for tailoring the text-visual interactions of different models.

Models	SD3.5-Large	FLUX.1-Dev	Qwen Image
Total Blocks	38	57	60
Tailoring Blocks	**{3,9,15,21}**	**{2,7,12,17,22}**	**{3,9,15,21,27}**

Table 2: Quantitative results on T2I-CompBench with token-level and full sentence tailoring results. TexTailor consistently improves the text alignment on three models across various attributes.

Model	Attribute Binding			Object Relationship			Count	Comp.	Image Quality	
	Color	Shape	Texture	2D Spa.	3D Spa.	NonSpa.			HPSv2	Aes.
SD3.5	0.7284	0.5592	0.7471	0.2866	0.3816	0.3118	0.5969	0.3727	29.2869	6.0978
TACA	0.7434	0.5784	0.7444	0.2947	0.3839	0.3114	0.6029	0.3820	**29.3225**	**6.2297**
Ours(token)	**0.8061**	**0.6789**	-	-	-	-	**0.6088**	-	29.0384	6.0346
Ours(full)	0.8052	0.6744	**0.8428**	**0.3647**	**0.3923**	**0.3169**	0.5987	**0.4047**	28.9501	5.9401
FLUX	0.7322	0.4908	0.6490	0.2935	0.3739	0.3044	0.5877	0.3597	29.1586	6.3563
TACA(r=64)	0.7535	0.5126	0.6522	0.3043	0.3814	0.3045	0.5855	0.3619	29.1525	6.3327
TACA(r=16)	0.7296	0.4898	0.6549	0.2991	0.3790	0.3034	0.5780	0.3585	29.1375	6.3205
Ours(token)	**0.7815**	**0.5500**	-	-	-	-	**0.6091**	-	**29.2329**	**6.4119**
Ours(full)	0.7804	0.5482	**0.6980**	**0.3280**	**0.3900**	**0.3054**	0.5860	**0.3691**	29.2267	6.4110
Qwen Image	0.8554	0.6358	0.7650	0.3973	0.4077	0.3110	0.7406	0.3983	28.8831	6.1925
Ours(token)	**0.8688**	**0.6369**	-	-	-	-	**0.7616**	-	29.0204	6.2307
Ours(full)	0.8677	0.6348	**0.7796**	**0.4560**	**0.4202**	**0.3123**	0.7359	**0.4104**	**29.0212**	**6.2378**

We uniformly select the proper number of blocks, 4 for SD3.5-Large and 5 for FLUX.1-Dev and Qwen Image, respectively, for tailoring in the yellow region of Fig. 4(bottom panel). Tab. 1 shows the selected tailoring blocks for different models, on which we conduct the tailoring experiments.

5.2 Improved Text Alignment of Text-to-Image Generation

Qualitative Results. Fig. 1 and Fig. 6 show the qualitative results of our TexTailor compared to three baseline models. TexTailor significantly improves the text alignment across various semantic attributes, including amount, colors, textures, and complex prompts, *etc*. Even though the tailoring blocks are selected based on the representative attributes, the other attributes can also benefit from our method. This observation demonstrates the generalizability of the tailoring blocks across different attributes.

Quantitative Results. Tab. 2 and Tab. 3 show the token-level and full tailoring results on T2I-CompBench and GenEval benchmarks. We could observe that our proposed method consistently obtains performance gains across various attributes on all three models, demonstrating the superiority and flexibility of our method. Remarkably, we achieve substantial improvement of 12% on Shape, 10% on Texture, and 8% on Color, in a totally training-free manner. Although due to the inherent limitations of the baseline models in the attribute of quantity, our token-level tailoring performs better than default sentence tailoring, outperforming the baseline performance. Additionally, the quantitative results

Table 3: Quantitative results on GenEval. * denotes token-level tailoring.

Model	Overall	Single	Two	Count*	Color	Pos.	ColorAttr	HPSv2	Aes.
SD3.5	0.6642	0.9438	0.8939	0.6344	0.8059	0.2325	0.4750	**29.5759**	**5.8871**
Ours(full)	**0.7163**	**0.9781**	**0.9672**	**0.6375**	**0.8650**	**0.3925**	**0.4825**	29.3729	5.7902
FLUX	0.6538	**0.9904**	0.8258	0.6375	0.7713	0.2575	0.4400	29.8115	6.3650
Ours(full)	**0.6826**	0.9688	**0.8914**	**0.6438**	**0.7739**	**0.3475**	**0.4700**	**29.8207**	**6.4043**
Qwen Image	0.8551	**0.9906**	0.9520	0.8562	0.8617	0.7375	0.7325	30.4510	**6.2327**
Ours(full)	**0.8777**	**0.9906**	**0.9722**	**0.8594**	**0.8989**	**0.7475**	**0.7975**	**30.6851**	6.2113

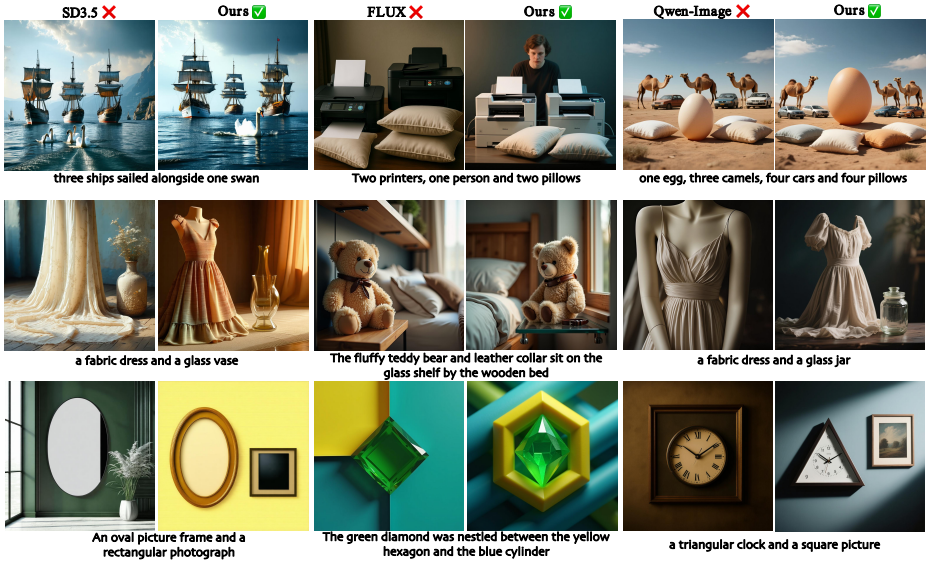


Fig. 6: Qualitative comparisons between baselines and our method. Our method significantly improves the text alignment across various semantic attributes, including amount, colors, textures, and complex prompts, *etc.* Zoom in for details.

of HPSv2 and Aesthetics scores demonstrate that our method improves the text alignment while maintaining the high aesthetic quality. Together with the qualitative results in Fig. 1 and Fig. 6, the quantitative results further show that the proposed TexTailor consistently improves semantic understanding in image generation tasks and generalizes well across various attributes.

5.3 Ablation Analysis

Analysis on $\lambda(l)$. This part investigates the sensitivity of the scale $\lambda(l)$. Specifically, we evaluate the performance of different attributes on FLUX with $\lambda(l)$ ranging from 1.2 to 2.0. As shown in Fig 7 (a), our method achieves significantly better results than the baseline despite some fluctuations, showing the effectiveness of our method. Additionally, we also evaluate the performance of

Table 4: Ablation analysis on smaller λ and block selections.

Configs	Color	Shape	2D Spatial
$\lambda = 0.7$	0.3161	0.2653	0.1100
$\lambda = 0.9$	0.6891	0.4395	0.2611
Enhancing Random 5 blocks	0.7624	0.5072	0.3119
Enhancing *All* blocks	0.2360	0.2736	0.0495
Ours	**0.7804**	**0.5482**	**0.3280**

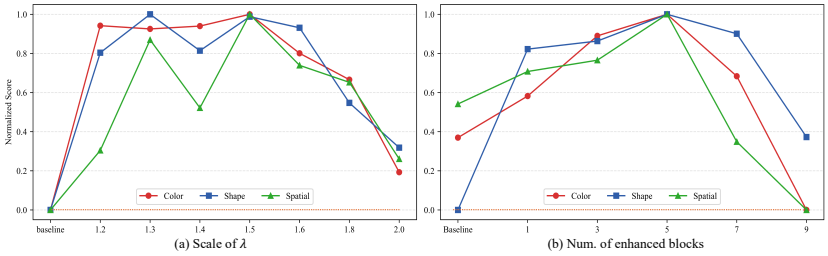


Fig. 7: Ablation analysis on enhancing scale (a) and block selections (b).

weakening the textual conditions in Tab. 4. It turns out that the weakening operation significantly decreases the model's performance, further demonstrating the importance of these vital blocks and validating the soundness of our method.

Analysis on the selection of enhanced blocks. To evaluate the effectiveness of our analysis in selecting the proper number of blocks for enhancement, we apply our enhancement to varying block counts $N \in \{1, 3, 5, 7, 9\}$. The results in Fig. 7 (b) show that increasing N initially improves performance, yet manipulating more than 9 blocks may degrade it due to distribution shift. We further tailor random FLUX blocks (5 and *all*) instead of our selected ones, and the results are shown in Tab. 4. The table shows that enhancing random or all blocks performs worse than enhancing the dedicated blocks identified by our analysis, highlighting the effectiveness of our approach. What's more, this observation also reflects that different blocks do not contribute equally to different attributes, consistent with our findings in Sec. 3.

Robustness across blocks, samplers, timesteps and resolution. The tailoring blocks are identified once per model and remain fixed across all inference configurations. Tab. 5 validates robustness across different samplers (EDM, dpm++), denoising steps ($\mathcal{S}$), and image resolutions ($\mathcal{R}$). TexTailor yields consistent gains in all settings, confirming the robustness of block selection while inference hyperparameters change.

Varying CFG Values. To verify that our improvements are not attributable to CFG upscaling effects, we evaluate our method on FLUX across a range of CFG values. As shown in Fig. 8, our method consistently outperforms the baseline on GenEval, T2I-CompBench, and HPSv2 across all CFG settings, demonstrating that the gains originate from our targeted text hidden state enhancement rather than from guidance scaling.

Table 5: T2I-CompBench Score across inference configurations on FLUX.

	EDM	dpm++	$\mathcal{S}$=8	$\mathcal{S}$=15	$\mathcal{S}$=28	$\mathcal{S}$=50	$\mathcal{R}$=512	$\mathcal{R}$=1024
Baseline	47.39	47.42	36.72	43.57	47.39	47.80	47.24	47.39
Ours	**50.46**	**50.59**	**40.32**	**47.58**	**50.46**	**50.86**	**50.41**	**50.46**

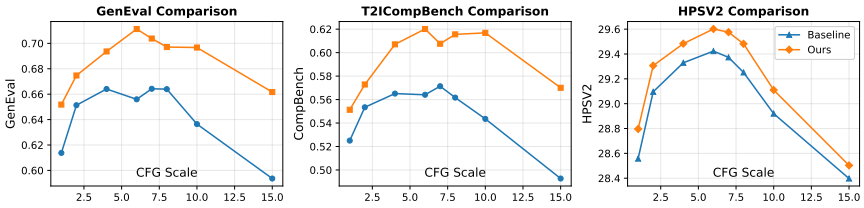


Fig. 8: Robustness to varying CFG values. TexTailor consistently outperforms the FLUX baseline across all CFG settings on GenEval, T2I-CompBench, and HPSv2.

6 Applications

Enabling Precise Text-based Editing. We incorporate our enhancement into editing tasks, facilitating precise textual editing with the target text instructions. Following the Stable Flow [1], we perform image editing via parallel generation, injecting self-attention features from the source image into the target image to preserve visual content. Differently, our empirical findings motivate us to enhance target text embeddings across critical blocks to improve the editing accuracy. The results in Fig. 1, Fig. 9 and Tab. 6 show the effectiveness of our method. Our method outperforms Stable Flow on CLIP_{txt} score (↑0.94), showing more accurate editing towards textural instructions. Meanwhile, the CLIP_{img} similarity remains nearly unchanged (↓0.008), suggesting that our method effectively enables more precise editing in line with the given instructions while preserving the visual integrity and coherence of the images. Furthermore, the result of human preference further reflects the effectiveness of our method.

Inference Acceleration. Recall that our analysis indicates that removing some blocks causes a smaller impact on the output, suggesting their role in rendering fine-grained details instead of vital semantics. Accordingly, we accelerate inference with a training-free mechanism by skipping specific blocks identified as less critical from our probing analysis. Formally, we skip the corresponding block or its CFG operation during inference. Notably, our method is complementary to existing acceleration techniques, such as Teacache [24], to achieve further acceleration. Tab. 7 reports inference acceleration results by skipping less critical blocks, showing averaged inference time over 400 prompts on NVIDIA 4090 and H100 GPUs. Our method substantially reduces inference time and can be seamlessly combined with TeaCache for further acceleration. The minor drops of HPSv2 (↓ 0.19) and Aesthetic (↓ 0.01) scores are justified by the acceptable visual quality and faster inference time.

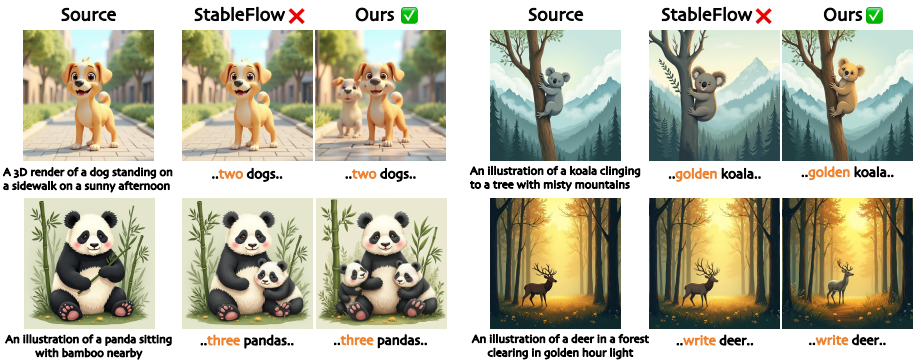


Fig. 9: Qualitative comparisons of editing results between Stable Flow and our method. Our method enables more precise editing on specific attributes by changing the color, amount, *etc.*

Table 6: Editing results.

Method	CLIP_{img}	CLIP_{txt}	HumanPref
Stable Flow	**0.9642**	35.2584	40.98%
+ Ours	0.9637	**36.1988**	**59.02%**

Table 7: Acceleration results.

Method	T(s,4090)	T(s,H100)	HPSv2	Aes.
FLUX	36.79	13.09	**29.0533**	**6.1903**
+ TeaCache	26.62	9.61	28.8951	6.2067
+ Ours	**24.53**	**8.88**	28.8647	6.1801

7 Conclusions

In this work, we systematically analyze block-wise contributions and their interactions with text conditions, offering a better understanding of the internal mechanisms within MMDiT-based generative models. Meanwhile, our analysis reveals several valuable findings that unlock new possibilities for improving the synthesis quality. Based on these findings, we introduce novel training-free techniques to facilitate text alignment of T2I tasks, precise semantic manipulation of editing tasks, and accelerated inference in a training-free manner. Extensive results demonstrate the effectiveness of our method.

Despite substantial performance gains, TexTailor has limitations. It relies on automatic block-wise analysis and cannot perfectly synthesize highly complex prompts due to pretraining constraints. Future work could incorporate trainable modules and token-level dynamic routing to further improve performance.

Acknowledgements

This work was supported in part by AI for Science Program, Shanghai Municipal Commission of Economy and Informatization (Grant No. 2025-GZL-RGZN-BTBX-02017) and the National Natural Science Foundation of China (Grant No.62576103).

References

1. Avrahami, O., Patashnik, O., Fried, O., Nemchinov, E., Aberman, K., Lischinski, D., Cohen-Or, D.: Stable flow: Vital layers for training-free image editing. In: CVPR. pp. 7877–7888 (2025)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Baldridge, J., Bauer, J., Bhutani, M., Brichtova, N., Bunner, A., et al.: Imagen 3. arXiv preprint arXiv:2408.07009 (2024)
4. Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., Zhu, J.: All are worth words: A vit backbone for diffusion models. In: CVPR. pp. 22669–22679 (2023)
5. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: ICCV. pp. 22560–22570 (2023)
6. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM transactions on Graphics (TOG) **42**(4), 1–10 (2023)
7. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In: ICLR (2024)
8. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: Advances in Neural Information Processing Systems. vol. 34, pp. 8780–8794 (2021)
9. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)
10. Gao, Y., Gong, L., Guo, Q., Hou, X., Lai, Z., Li, F., Li, L., Lian, X., Liao, C., Liu, L., et al.: Seedream 3.0 technical report. arXiv preprint arXiv:2504.11346 (2025)
11. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. In: Advances in Neural Information Processing Systems. vol. 36, pp. 52132–52152 (2023)
12. Gong, L., Hou, X., Li, F., Li, L., Lian, X., Liu, F., Liu, L., Liu, W., Lu, W., Shi, Y., et al.: Seedream 2.0: A native chinese-english bilingual image generation foundation model. arXiv preprint arXiv:2503.07703 (2025)
13. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. pp. 6840–6851 (2020)
15. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
16. Huang, K., Duan, C., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
17. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In: Advances in Neural Information Processing Systems. vol. 36, pp. 78723–78747 (2023)
18. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Advances in Neural Information Processing Systems **35**, 26565–26577 (2022)

19. Labs, B.F.: Flux. github.com/black-forest-labs/flux (2024)
20. Li, B., Guan, Y., Li, H., Zeng, B., Ji, Y., Ding, Y., Wan, P., Gai, K., Zhang, Y., Zhang, W.: Semantic routing: Exploring multi-layer llm feature weighting for diffusion transformers. arXiv preprint arXiv:2602.03510 (2026)
21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755. Springer (2014)
22. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: ICLR (2023)
23. Liu, B., Wang, C., Cao, T., Jia, K., Huang, J.: Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In: CVPR. pp. 7817–7826 (2024)
24. Liu, F., Zhang, S., Wang, X., Wei, Y., Qiu, H., Zhao, Y., Zhang, Y., Ye, Q., Wan, F.: Timestep embedding tells: It's time to cache for video diffusion model. In: CVPR (2025)
25. Liu, X., Gong, C., et al.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: ICLR (2023)
26. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In: Advances in Neural Information Processing Systems. vol. 35, pp. 5775–5787 (2022)
27. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint arXiv:2310.04378 (2023)
28. Lv, Z., Pan, T., Si, C., Chen, Z., Zuo, W., Liu, Z., Wong, K.Y.K.: Rethinking cross-modal interaction in multimodal diffusion transformers. In: ICCV (2025)
29. Ma, X., Wang, Y., Jia, G., Chen, X., Liu, Z., Li, Y.F., Chen, C., Qiao, Y.: Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048 (2024)
30. OpenAI: Video generation models as world simulators (2024), technical Report, OpenAI
31. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2023)
32. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: CVPR. pp. 4195–4205 (2023)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
35. Schuhmann, C.: Laion aesthetics. github.com/LAION-AI/aesthetic-predictor (2022)
36. Shin, J., Hwang, A., Kim, Y., Kim, D., Park, J.: Exploring multimodal diffusion transformers for enhanced prompt-based image editing. In: ICCV (2025)
37. Si, C., Huang, Z., Jiang, Y., Liu, Z.: Freeu: Free lunch in diffusion u-net. In: CVPR. pp. 4733–4743 (2024)
38. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML. pp. 2256–2265. pmlr (2015)

39. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
40. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. In: ICML. pp. 32211–32252. PMLR (2023)
41. Stability-AI: Stable diffusion 3.5. github.com/Stability-AI/sd3.5 (2024)
42. Tan, Z., Liu, S., Yang, X., Xue, Q., Wang, X.: Ominicontrol: Minimal and universal control for diffusion transformer. arXiv preprint arXiv:2411.15098 (2024)
43. Team, H.F.M.: Hunyuanvideo: A systematic framework for large video generative models. github.com/Tencent-Hunyuan/HunyuanVideo (2024)
44. Team, H.F.M.: Hunyuanimage-2.1: An efficient diffusion model for high-resolution (2k) text-to-image generation. github.com/Tencent-Hunyuan/HunyuanImage-2.1 (2025)
45. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (2017)
46. Voynov, A., Chu, Q., Cohen-Or, D., Aberman, K.: P+: Extended textual conditioning in text-to-image generation. arXiv preprint arXiv:2303.09522 (2023)
47. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
48. Wei, T., Zhou, Y., Chen, D., Pan, X.: Freeflux: Understanding and exploiting layer-specific roles in rope-based mmdit for versatile image editing. In: ICCV (2025)
49. Williams, C., Falck, F., Deligiannidis, G., Holmes, C.C., Doucet, A., Syed, S.: A unified framework for u-net design and analysis. In: Advances in Neural Information Processing Systems. vol. 36, pp. 27745–27782 (2023)
50. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
51. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
52. Xiang, Q., Sun, S., Li, B., Song, D., Li, H., Chen, N., Tang, X., Hu, Y., Zhang, J.: Instanceassemble: Layout-aware image generation via instance assembling attention. In: Advances in Neural Information Processing Systems (2025)
53. Xie, X., Guo, J., Gong, D.: Hyperalign: Hypernetwork for efficient test-time alignment of diffusion models. arXiv preprint arXiv:2601.15968 (2026)
54. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: ICLR (2025)
55. Yi, M., Li, A., Xin, Y., Li, Z.: Towards understanding the working mechanism of text-to-image diffusion model. In: Advances in Neural Information Processing Systems. pp. 55342–55369 (2024)
56. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV. pp. 3836–3847 (2023)
57. Zhou, F., Cao, P., Ma, Y., Yang, L., Yin, J.: Exploring position encoding in diffusion u-net for training-free high-resolution image generation. arXiv preprint arXiv:2503.09830 (2025)