AccelAes: Accelerating Diffusion Transformers for Training-Free Aesthetic-Enhanced Image Generation

Xuanhua Yin[0009−0006−4284−2400], Chuanzhi Xu[0009−0009−9076−7005], Haoxian Zhou[0009−0004−2640−350X], Boyu Wei[0009−0001−2730−9759], and Weidong Cai[0000−0003−3706−8896]⋆

School of Computer Science, The University of Sydney, NSW 2006, Australia
{xuanhua.yin,chuanzhi.xu,hzho0442,bwei0951,tom.cai}@sydney.edu.au

Abstract. Diffusion Transformers (DiTs) are a dominant backbone for high-fidelity text-to-image generation due to strong scalability and alignment at high resolutions. However, quadratic self-attention over dense spatial tokens leads to high inference latency and limits deployment. We observe that denoising is spatially non-uniform with respect to aesthetic descriptors in the prompt. Regions associated with aesthetic tokens receive concentrated cross-attention and show larger temporal variation, while low-affinity regions evolve smoothly with redundant computation. Based on this insight, we propose **AccelAes**, a training-free framework that accelerates DiTs through aesthetics-aware spatio-temporal reduction while improving perceptual aesthetics. AccelAes builds AesMask, a one-shot aesthetic focus mask derived from prompt semantics and cross-attention signals. When localized computation is feasible, SkipSparse reallocates computation and guidance to masked regions. We further reduce temporal redundancy using a lightweight step-level prediction cache that periodically replaces full Transformer evaluations. Experiments on representative DiT families show consistent acceleration and improved aesthetics-oriented quality. On Lumina-Next, AccelAes achieves a **2.11**× speedup and improves ImageReward by **+11.9%** over the dense baseline. Code is available at https://github.com/xuanhuayin/AccelAes.

Keywords: Diffusion Transformers · Training-Free Acceleration · Text-to-Image Generation · Aesthetic Enhancement

1 Introduction

Diffusion models have become dominant for high-fidelity text-to-image generation [13,36]. Recent research increasingly adopts Transformer backbones, namely the Diffusion Transformer (DiT) [5, 32, 33, 47], to achieve better scalability and improved visual quality. Compared with earlier U-Net based pipelines [36], DiTs provide higher capacity and stronger alignment at high resolutions [32], enabling the generation of more detailed, realistic, and semantically consistent images.

⋆ Corresponding author.

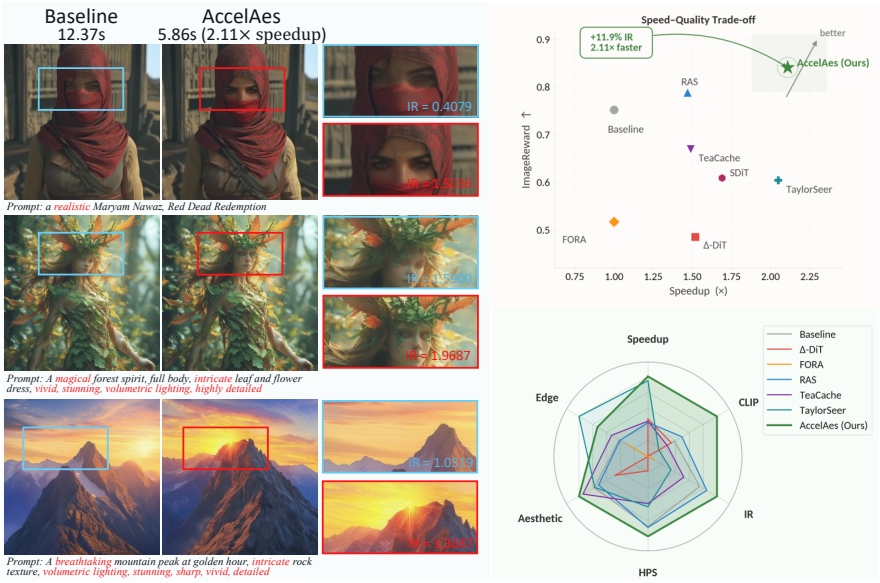

Fig. 1: AccelAes improves both efficiency and overall quality. Left: Qualitative comparison under identical prompts, with zoomed-in crops and per-sample ImageReward (IR). Right: Speed–quality summary, and a bottom-right multi-metric radar on **Lumina** benchmarking AccelAes against the dense baseline and representative training-free acceleration methods, demonstrating consistent gains across metrics.

However, the performance comes with a more immediate cost: inference becomes significantly slower. At each denoising step, a DiT must perform full self-attention over all spatial tokens, and the pairwise interactions among tokens make every forward pass computationally expensive. Moreover, high-resolution image generation typically requires dozens of denoising steps, leading to cumulative inference latency across iterations. In practice, model performance is often constrained not only by model size but also by repeated computation and runtime overhead during sampling. As attention operations and memory movement account for an increasing portion of total execution time, it has become a central challenge for DiTs to reduce redundant computation at inference while maintaining visual quality [10]. This motivates our focus on training-free acceleration that explicitly accounts for perceptual and aesthetics-oriented criteria.

To address this issue, many works aim to accelerate diffusion inference without retraining. Some research reduces the total number of denoising steps through improved samplers [6, 9, 21, 25, 48]. Others reduce per-step computation, for example, through token merging [2,3,12,42,46], structured sparse attention [46], or intermediate feature caching [6,7,9,12,21,25,34,38,45,48]. However, these methods share a common limitation: they apply uniform strategies across the spatial dimension, either merging or pruning all tokens in the same manner, or caching and skipping computation uniformly over the entire image. Such strategies are

often vulnerable in practice because image generation is not spatially uniform. Certain regions, such as main subjects, fine textures, and lighting boundaries, have a much greater impact on perceptual quality, while other regions, such as smooth backgrounds, contribute less to the final visual impression [28, 41]. If computation is aggressively reduced in perceptually important regions, image quality degrades significantly. If conservative settings are applied everywhere to protect quality, much of the potential speedup is lost. This leads to a natural question: can computation be allocated adaptively across space according to content importance?

From another perspective, text-to-image models based on DiTs are no longer solely optimized for strict semantic alignment with the prompt, but are increasingly designed to generate images that better reflect human preferences and aesthetic judgments [4, 15, 17, 18, 37]. With advances in alignment methods and human feedback optimization methods such as RLHF [8, 31], modern generative models place greater emphasis on aesthetic attributes, including "intricate", "volumetric-lighting", and "detailed" [39, 44]. These descriptors often appear as adjectives or stylistic tokens in prompts and influence image generation via cross-attention mechanisms [30, 36]. In other words, the model is not only concerned with what to generate, but also with how to generate it well.

This trend suggests an important insight: different spatial regions contribute unequally to aesthetic quality. Regions containing main subjects with fine textures, lighting transitions, and material details often dominate human perception and aesthetic judgment, while large smooth background areas typically have a smaller influence on aesthetic scores [28, 44]. Moreover, cross-attention explicitly models the correspondence between text tokens and image patches in modern text-to-image diffusion models [30, 36]. In particular, tokens associated with aesthetic descriptors tend to induce highly concentrated attention patterns over specific spatial regions. This suggests that during denoising, the model implicitly differentiates between aesthetically sensitive regions and less sensitive regions.

However, existing DiT acceleration methods rarely exploit this signal. They typically apply uniform token reduction based on numerical similarity or fixed structural rules, overlooking the spatial non-uniformity that is closely tied to human preference. In a generation paradigm that increasingly emphasizes aesthetic quality, such semantically unaware compression is suboptimal.

We find that spatial aesthetic non-uniformity and temporal redundancy are coupled. Regions related to aesthetic descriptors vary more across timesteps and are harder to extrapolate, while low-affinity regions evolve smoothly and are more reusable. This implies that the attention structure already indicates which regions require precise computation and which can be skipped. Leveraging both signals enables content-aware acceleration while preserving key aesthetic quality.

Based on the above observations, we propose **AccelAes (Accelerated Aesthetics)**, a training-free acceleration framework for DiTs. The core idea of AccelAes is to leverage aesthetics-driven spatial non-uniformity and temporal prediction redundancy during the denoising process, allocating computation preferentially to perceptually sensitive regions while compressing redundant com-

putation in less critical areas. Specifically, we use internal attention signals to construct AesMask, an aesthetics-aware semantic mask that identifies spatial regions associated with aesthetic descriptors, and refine them more carefully when the architecture permits localized updates. Meanwhile, we introduce SkipSparse, a unified acceleration module that combines spatially-adaptive sparse computation with an output-level step prediction reuse mechanism (StepCache) along the temporal dimension to reduce redundant computation across timesteps. These two mechanisms complement each other, enabling effective inference acceleration while preserving key perceptual quality. Fig. 1 previews this speed–quality tension and shows that AccelAes improves both efficiency and overall quality on representative prompts and metrics. Our contributions are summarized as follows:

- We propose **AccelAes**, a training-free acceleration framework for Diffusion Transformers, motivated by aesthetics-driven spatial non-uniformity and temporal redundancy during denoising.
- We introduce an aesthetics-aware acceleration design that couples **AesMask** and **SkipSparse** to localize perceptually sensitive regions from internal model signals and reallocate computation and guidance accordingly, while compressing less critical updates.
- We conduct extensive experiments on three representative DiT backbones, showing that AccelAes substantially reduces inference time while consistently improving aesthetics- and preference-oriented quality, yielding a markedly better speed–quality trade-off.

2 Related Work

2.1 Acceleration of DiT for Image Generation

Diffusion-based generators synthesize images via iterative denoising, making sampling computationally intensive and motivating training-free acceleration that improves efficiency without modifying backbone parameters [13, 29, 36]. Prior work mainly follows two directions: reducing sampling steps or changing the sampling trajectory with improved solvers and controlled exploration [24, 26, 40], or reducing per-step cost by exploiting temporal and spatial redundancy, e.g., feature caching/reuse [7, 9, 25, 34, 48], timestep forecasting/skipping [21, 38], and token-level compression (pruning/merging) or training-free sparse attention variants [2, 3, 12, 22, 42, 46].

Many methods decide reductions using numerical similarity, heuristics, or generic saliency proxies rather than perceptual objectives, so aggressive compression can remove compute from visually important regions and introduce localized artifacts. Region-adaptive methods such as RAS and SDiT allocate non-uniform budgets across spatial tokens [19, 23], but still rely on proxy signals (e.g., attention concentration, heuristic complexity, or semantic grouping) to define region priorities [19, 23]. Meanwhile, modern text-to-image systems increasingly optimize for human preference and aesthetic quality beyond semantic

correctness [8,18,31], yet such image-level objectives do not directly yield stable region priorities for inference-time budgeting. Moreover, trajectory-level acceleration (e.g., caching/forecasting) is typically applied uniformly over space, leaving open where computation should be preserved under limited budgets.

2.2 Aesthetic and Guidance Signals for Image Generation

Inference-time guidance offers a practical control interface for text-to-image diffusion models: CLIP supports text–image semantic consistency [35], and classifier-free guidance (CFG) strengthens prompt adherence by scaling the conditional–unconditional gap [14]. However, a single global guidance scale can be spatially inconsistent and induce local semantic/structural issues, motivating spatially varying guidance that applies different strengths to semantic regions without additional training [39]; related work further enables continuous modulation of attribute intensity [4]. These approaches keep dense computation and do not address how guidance should interact with inference-time compute budgeting, whereas we explicitly couple guidance with spatial compute allocation in a training-free acceleration setting.

Human preference and aesthetic assessment provides complementary signals for generative models. Preference datasets and reward models such as Pick-a-Pic and ImageReward align objectives with human feedback [16,44], and preference feedback can be used for post-training improvements such as step-by-step preference optimization [18]. For evaluation, preference and aesthetics are often measured by image-level scoring functions, including CLIP-based aesthetic predictors [37], benchmarks such as HPSv2 [43], and classic datasets like AVA [28]; aesthetic-related rewards have also been used for automatic prompt optimization [27]. In our work, preference-related models are used for evaluation rather than retraining, but their scalar image-level scores are difficult to translate into region-level priorities or token-level compute budgets. This leaves an interface problem for training-free acceleration: how to turn aesthetic objectives into executable spatial budgeting rules. We address this gap by extracting aesthetic-related cues from internal attention patterns conditioned on aesthetic descriptors and using them to guide spatial compute allocation while preserving global context.

3 Methodology

We propose **AccelAes**, a training-free acceleration framework for Diffusion Transformers (DiTs) based on aesthetics-driven spatio-temporal non-uniformity. Given a text prompt, AccelAes first constructs **AesMask**, an aesthetics-aware semantic mask over image tokens from cross-attention signals. It then applies **SkipSparse**, which performs spatially-adaptive sparse computation when localized updates are supported. SkipSparse also includes a universal step-level prediction cache that reuses noise predictions across timesteps. Together, AesMask decides *where* to spend compute, and SkipSparse decides *how* to reduce redundant computation in both space and time.

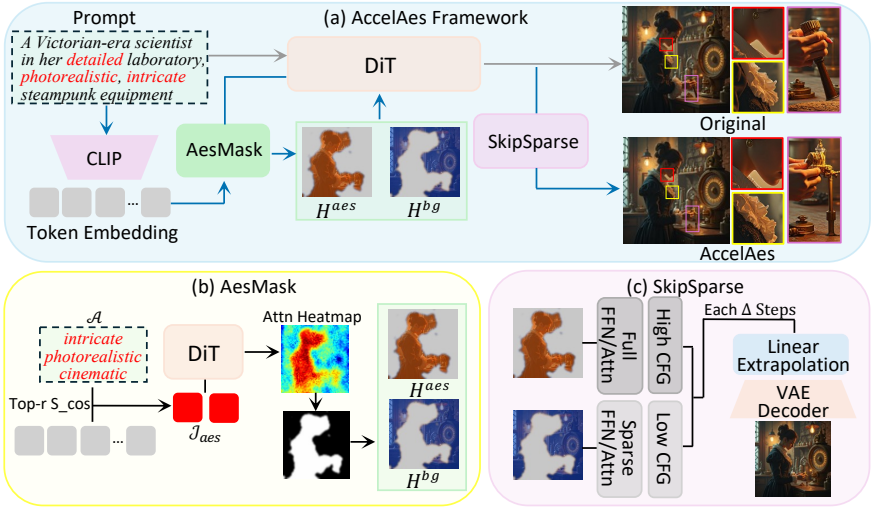

Fig. 2: The proposed AccelAes framework. (a) AccelAes builds AesMask from prompt semantics and cross-attention. It uses the mask for spatially adaptive computation and region-wise guidance. SkipSparse also includes a step-level cache with linear extrapolation to reduce redundant forwards across timesteps. (b) AesMask uses aesthetic anchors, CLIP similarity, and cross-attention aggregation, then applies percentile thresholding to obtain a binary mask. (c) SkipSparse updates attention and FFN mainly on aesthetic-focus tokens. It keeps global context with global keys/values and reuses predictions across steps via the cache.

3.1 Preliminaries: Aesthetically-Driven Non-Uniformity

We consider diffusion sampling in latent space. Let $\mathbf{z}_t \in \mathbb{R}^{d\times N}$ be latent tokens at timestep t, where N is the number of spatial tokens. A DiT parameterized by θ predicts noise (or velocity) as:

$$\hat{\boldsymbol{\epsilon}}_t = f_\theta(\mathbf{z}_t, \mathbf{c}, t)\,, \tag{1}$$

where $\mathbf{c}$ is the text condition and $\hat{\boldsymbol{\epsilon}}_t \in \mathbb{R}^{d\times N}$ is the per-token prediction.

DiT blocks use self-attention and cross-attention between image tokens and text tokens. Let $\mathbf{y} = [\mathbf{y}_1, \ldots, \mathbf{y}_M] \in \mathbb{R}^{d_c\times M}$ be M text token embeddings. For a cross-attention layer ℓ at timestep t, we denote its weights by:

$$\mathbf{A}_t^{(\ell)} \in [0,1]^{N\times M}, \qquad \sum_{j=1}^{M} \mathbf{A}_t^{(\ell)}[i,j] = 1 \quad \forall i, \tag{2}$$

where $\mathbf{A}_t^{(\ell)}[i,j]$ measures how much image token i attends to text token j.

We observe a consistent non-uniformity for aesthetics-oriented prompts. Attention mass concentrates on a subset of spatial tokens linked to aesthetic descriptors. These tokens often control textures, highlights, and boundaries that

dominate perceptual judgment. Other tokens are less sensitive and change more smoothly across nearby timesteps. AccelAes uses this property to reduce computation where it is less critical and to preserve computation where it matters.

3.2 AesMask: Aesthetics-Aware Semantic Mask Generation

We build a binary mask over image tokens for *aesthetic-focus regions*. This module is **AesMask**, as shown in Fig. 2 (b). In text-conditioned DiTs, the prompt affects image tokens through cross-attention. We use this internal signal to locate tokens related to aesthetic-relevant prompt semantics, and we mark them for later selective updates.

We define a set of aesthetic anchor words:

$$\mathcal{A} = \{a_1, \ldots, a_K\}, \tag{3}$$

where each a_k is a textual descriptor (e.g., "cinematic", "intricate"). The anchors provide a compact reference that helps us pick aesthetic-related prompt tokens without external supervision.

Aesthetic Prompt Tokens. Let $e(\cdot)$ be a frozen text encoder (CLIP text encoder). For each prompt token $\mathbf{y}_j$, we compute cosine similarity to each anchor a_k:

$$s_{j,k} = \frac{\langle e(\mathbf{y}_j), e(a_k)\rangle}{\|e(\mathbf{y}_j)\| \; \|e(a_k)\|}. \tag{4}$$

We assign each token an aesthetic score:

$$\alpha_j = \max_{k \in \{1,\ldots,K\}} s_{j,k}, \tag{5}$$

and select the top-r tokens by α_j. Their index set is $\mathcal{J}_{\text{aes}} \subseteq \{1, \ldots, M\}$. These tokens represent the aesthetic-related part of the prompt that we will use when aggregating attention.

Cross-attention Aggregation. At timestep t, each cross-attention layer ℓ outputs $\mathbf{A}_t^{(\ell)} \in \mathbb{R}^{N \times M}$. We aggregate weights for the selected aesthetic tokens:

$$m_t[i] = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \sum_{j \in \mathcal{J}_{\text{aes}}} \mathbf{A}_t^{(\ell)}[i, j], \tag{6}$$

where $m_t \in \mathbb{R}^N$ is an affinity map over image tokens and $\mathcal{L}$ is the set of layers used. A larger $m_t[i]$ means token i consistently attends to aesthetic-relevant text tokens. We treat it as an internal proxy for perceptual importance.

Percentile Thresholding. We binarize m_t with a percentile threshold:

$$\mathbf{M}_t[i] = \mathbb{I}(m_t[i] \geq \text{Perc}(m_t, \, p)), \tag{7}$$

where $\mathbf{M}_t \in \{0, 1\}^N$ and $p \in (0, 100)$ controls sparsity. We define $\mathcal{I}_t = \{i \mid \mathbf{M}_t[i] = 1\}$. The percentile form adapts to the scale of m_t across prompts, so it avoids tuning an absolute threshold. AesMask is training-free and uses cross-attention and a frozen text encoder. Its output $\mathbf{M}_t$ (or $\mathcal{I}_t$) is passed to SkipSparse.

3.3 Spatially-Adaptive Sparse Computation

When localized updates are supported, we use $\mathbf{M}_t$ to reduce computation. This module is **SkipSparse**, as shown in Fig. 2 (c). SkipSparse focuses updates on tokens in $\mathcal{I}_t$ and keeps global context through global keys/values. This is important because it allows focus tokens to read information from the full image.

Let $\mathbf{H}_t \in \mathbb{R}^{d\times N}$ be the hidden states entering a Transformer block at timestep t. We split tokens into focus and background:

$$\mathbf{H}_t = \left[\mathbf{H}_t^{\text{aes}},\ \mathbf{H}_t^{\text{bg}}\right], \tag{8}$$

where $\mathbf{H}_t^{\text{aes}}$ contains tokens in $\mathcal{I}_t$ and $\mathbf{H}_t^{\text{bg}}$ contains the rest. We scatter updated tokens back to the original order after each block so the block output matches the dense baseline shape.

Self-attention with Global Keys/Values. We update attention only for focus queries and keep keys/values global:

$$\text{Attn}\big(\mathbf{Q}_t^{\text{aes}}, \mathbf{K}_t^{\text{all}}, \mathbf{V}_t^{\text{all}}\big) = \text{Softmax}\left(\frac{(\mathbf{Q}_t^{\text{aes}})^\top \mathbf{K}_t^{\text{all}}}{\sqrt{d_h}}\right)\mathbf{V}_t^{\text{all}}, \tag{9}$$

where $\mathbf{Q}_t^{\text{aes}} = \mathbf{W}_Q\mathbf{H}_t^{\text{aes}}$, $\mathbf{K}_t^{\text{all}} = \mathbf{W}_K\mathbf{H}_t$, and $\mathbf{V}_t^{\text{all}} = \mathbf{W}_V\mathbf{H}_t$. This reduces the number of queries from N to $|\mathcal{I}_t|$ and keeps long-range interactions through global keys/values.

FFN on Aesthetic Tokens. We apply FFN only to focus tokens and reuse a cached background activation:

$$\tilde{\mathbf{H}}_t^{\text{aes}} = \text{FFN}(\mathbf{H}_t^{\text{aes}}), \qquad \tilde{\mathbf{H}}_t^{\text{bg}} = \mathbf{C}_t^{\text{bg}}, \tag{10}$$

where $\mathbf{C}_t^{\text{bg}}$ is taken from the most recent full update. This avoids recomputing FFN on background tokens and still allows refinement on aesthetic-focus tokens.

Spatial CFG. Let $\hat{\boldsymbol{\epsilon}}_t^{\text{cond}}$ and $\hat{\boldsymbol{\epsilon}}_t^{\text{uncond}}$ be conditional and unconditional predictions. We apply a spatially varying guidance scale:

$$\hat{\boldsymbol{\epsilon}}_t^{\text{cfg}}[i] = \hat{\boldsymbol{\epsilon}}_t^{\text{uncond}}[i] + g_t[i]\left(\hat{\boldsymbol{\epsilon}}_t^{\text{cond}}[i] - \hat{\boldsymbol{\epsilon}}_t^{\text{uncond}}[i]\right), \tag{11}$$

where $g_t[i] = g_{\text{bg}} + (g_{\text{aes}} - g_{\text{bg}})\mathbf{M}_t[i]$ and $g_{\text{aes}} \geq g_{\text{bg}} \geq 1$. This places stronger guidance on focus tokens and keeps background guidance milder.

3.4 Universal Step-Level Prediction Caching

SkipSparse also includes a universal step-level cache, as shown in Fig. 2 (c). It reuses the final output $\hat{\boldsymbol{\epsilon}}_t$ and works across DiT variants. It targets a simple fact that nearby timesteps often have correlated predictions.

Table 1: Main results on three DiT backbones. We report runtime and generation-quality metrics for AccelAes and competing methods on Lumina-Next, SD3-Medium, and FLUX.1-dev, showing consistent speedup with improved overall quality.

Models	Methods	Time↓	Speedup↑	CLIP↑	IR↑	HPS↑	Aesth↑	Edge↑
Lumina	Baseline	12.37	1.00×	0.2531	0.7518	0.2710	5.9407	0.5829
Lumina	**AccelAes**	**5.86**	**2.11**×	**0.2640**	**0.8410**	**0.2740**	**6.0414**	**0.6285**
SD3	Baseline	3.52	1.00×	0.2662	0.8788	0.2895	5.7330	0.9160
SD3	**AccelAes**	**2.34**	**1.50**×	**0.2744**	**0.9042**	**0.3014**	**5.9898**	**0.9437**
FLUX	Baseline	12.66	1.00×	0.2753	1.233	0.3200	6.243	0.560
FLUX	**AccelAes**	**7.31**	**1.73**×	**0.2772**	**1.317**	**0.3214**	**6.372**	**0.590**

We cache predictions at selected timesteps and reuse them with linear extrapolation. Let Δ be the stride between full forwards. Let τ be the latest timestep with a full forward. We store:

$$\mathbf{E}_\tau = \hat{\boldsymbol{\epsilon}}_\tau, \qquad \mathbf{E}_{\tau+\Delta} = \hat{\boldsymbol{\epsilon}}_{\tau+\Delta}. \tag{12}$$

For an intermediate timestep $t \in (\tau, \tau + \Delta)$, we approximate:

$$\hat{\boldsymbol{\epsilon}}_t \approx \mathbf{E}_\tau + \lambda(t)\left(\mathbf{E}_\tau - \mathbf{E}_{\tau+\Delta}\right), \tag{13}$$

where $\lambda(t) = \frac{t-\tau}{\Delta}$. We use a warmup of T_w initial steps with full inference. After warmup, we refresh the cache every Δ steps by running a full forward and updating Eq. 12. This periodic refresh limits drift and keeps stable.

4 Experiments and Results

4.1 Experimental Setup

Models and Baselines. We evaluate AccelAes on three Diffusion Transformer backbones: Lumina-Next-T2I [47], SD3-Medium [11], and FLUX.1-dev [1]. We follow the standard DiT formulation [32] and only modify inference-time computation. For training-free acceleration baselines, we include Δ-DiT [6], FORA [38], Region-Adaptive Sampling (RAS) [23], TeaCache [20], and TaylorSeer [21]. For each backbone, all methods are evaluated under the same backbone-specific sampler and guidance configuration, and we adhere to the same implementation-level rules when reproducing baselines.

Evaluation Protocol. We use a unified protocol with 10,000 prompts from Pick-a-Pic and three fixed seeds per prompt (seeds $\{1, 2, 3\}$; 30,000 images per configuration) for both main results and ablations. Within each table, all AccelAes variants and baselines share the same prompt subset and seeds to enable paired comparisons. All methods are implemented in a unified PyTorch pipeline without

Table 2: Lumina-Next-T2I: baseline comparisons. Comparison of AccelAes with representative acceleration baselines on Lumina-Next-T2I using the main metric set.

Methods	Time↓	Speedup↑	CLIP↑	IR↑	HPS↑	Aesth↑	Edge↑
Lumina-Next-T2I [47]	12.37	1.00×	0.2531	0.752	0.2710	5.941	0.583
Δ-DiT [6]	8.13	1.52×	0.2523	0.485	0.2540	5.873	0.522
FORA [38]	12.36	1.00×	0.2463	0.517	0.2496	5.766	0.564
RAS [23]	8.38	1.47×	0.2550	0.788	0.2713	5.927	0.581
SDiT [19]	7.30	1.69×	0.2502	0.6093	0.2538	5.822	0.4289
TeaCache [20]	8.28	1.49×	0.2512	0.670	0.2640	5.978	0.599
TaylorSeer [21]	6.02	2.05×	0.2491	0.604	0.2650	5.940	**0.668**
AccelAes (ours)	**5.86**	**2.11×**	**0.2640**	**0.841**	**0.2740**	**6.041**	0.629

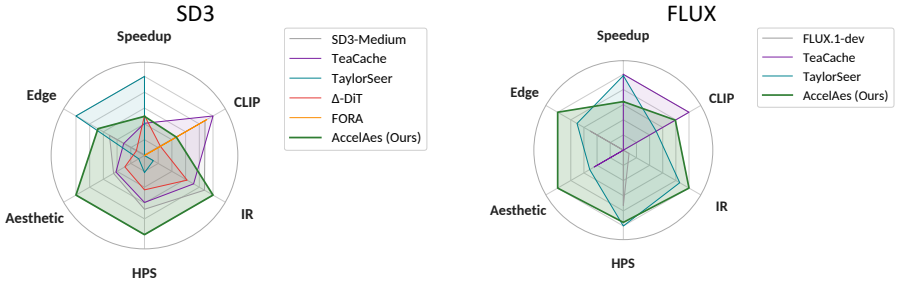

Fig. 3: Baseline comparisons on SD3 and FLUX. We compare AccelAes with the dense baseline and prior acceleration methods on SD3 and FLUX using the same prompts, seeds, and measurement protocol. AccelAes achieves a better speed quality trade off and improves aesthetics oriented scores while reducing sampling time.

architecture-specific engines, kernel-level optimizations, or external accelerators, and we follow backbone-default sampling settings. We evaluate all backbones at 1024×1024 resolution using 30 steps for Lumina-Next-T2I and 28 steps for SD3-Medium and FLUX.1-dev. Runtime is reported as end-to-end per-image latency from the start of sampling to the final decoded image, including VAE decoding and the two-pass CFG forward when enabled. Unless specified otherwise, AccelAes uses a cache refresh interval of $\Delta = 2$ with warmup $T_w = 5$, a one-shot AesMask at `mask_step=5`, and the same fixed 34-anchor vocabulary across paired comparisons.

Metrics and Measurement. We report runtime and six automatic metrics: Time (seconds per image) and Speedup defined as $\text{Time}_{\text{baseline}}/\text{Time}_{\text{method}}$, CLIP score (CLIP) [35] for text–image alignment, ImageReward (IR) [44] and HPSv2 (HPS) [43] as preference-oriented proxies, LAION Aesthetic Score (Aesth) computed with the LAION aesthetic predictor [37], and Edge Density (Edge) as a lightweight sharpness/detail proxy. Time is averaged over the full evaluation set and includes the complete sampling trajectory, including conditional/unconditional

Table 3: Blind pairwise human preference study. Win rates exclude ties. Each comparison uses randomized left/right ordering and four questions: overall preference, prompt alignment, visual appeal, and fewer artifacts.

Comparison	Overall	Alignment	Appeal	Fewer artifacts
AccelAes vs. Dense	67%	59%	68%	64%
AccelAes vs. FORA	**89%**	**90%**	**87%**	**94%**
AccelAes vs. w/o Spatial-CFG	66%	69%	66%	68%

passes for CFG. Unless otherwise specified, AccelAes constructs a one-shot semantic mask at `mask_step=5` and reuses it for the entire sampling trajectory and classifier-free guidance follows the standard formulation [14].

4.2 Main Results

Overall Results across Backbones. Table 1 shows that AccelAes improves the speed–quality trade-off across three DiT backbones. The key pattern is that acceleration does not come from uniformly weakening denoising. Instead, AccelAes reduces compute where the denoising trajectory is redundant and preserves updates where aesthetic cues are concentrated. This leads to simultaneous runtime reduction and stronger preference-oriented metrics. On Lumina-Next, AccelAes reaches the highest speedup and also yields a clear improvement on ImageReward together with gains on CLIP, HPS, Aesth, and Edge. On SD3-Medium, AccelAes again improves ImageReward and HPS under a 1.50× speedup. On FLUX.1-dev, AccelAes maintains a strong balance, which suggests the same mechanism transfers beyond a single architecture and sampler. Overall, these results support that aesthetics-driven spatio-temporal non-uniformity is a reliable signal for compute reduction under matched prompts, seeds, and backbones.

4.3 Human and Robustness Validation

Table 3 adds a direct human validation of the automatic metric gains. The study uses 15 annotators, 60 pairs, and four questions per pair. AccelAes is preferred over the dense baseline, FORA, and the variant without Spatial CFG across all criteria. Binomial tests on non-tie votes reject the 50% chance null for the main preference comparisons ($p < 0.001$, except alignment against the dense baseline). We further checked whether these gains depend on prompt wording or simple scene bias. On descriptor-free prompts, AccelAes improves ImageReward from 1.109 to 1.144 (+3.1%); with aesthetic descriptors, it improves the stronger dense baseline from 1.168 to 1.224 (+4.8%). Across six hard-scene categories, the ImageReward gain remains positive for single-subject (+1.8%), multi-object (+7.2%), dense-crowd (+5.0%), landscape (+3.3%), indoor-clutter (+6.9%), and abstract-art prompts (+13.7%). These results indicate that AesMask is not a center-subject shortcut, but follows prompt-conditioned aesthetic

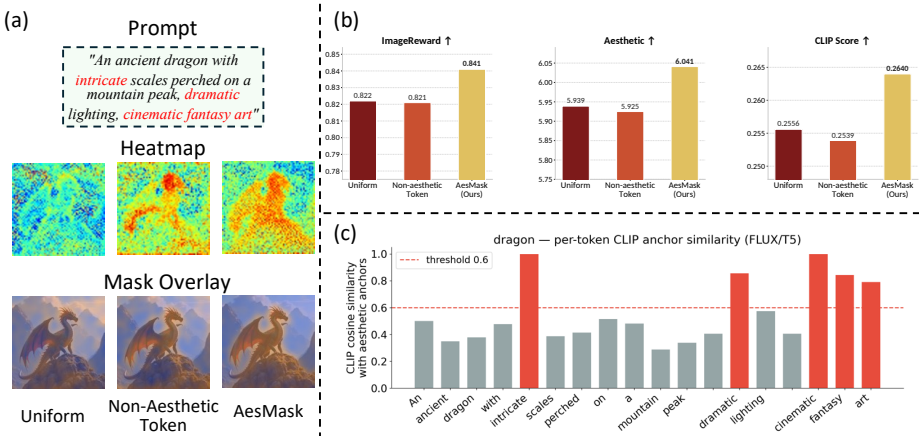

Fig. 4: AesMask analysis and ablation. We study how AesMask is constructed and how it affects acceleration and quality. We ablate key design choices such as anchor selection, cross attention aggregation, and mask thresholding, and we report their impact on both runtime and aesthetic metrics.

affinity even when the foreground/background structure is weak. The additional runtime diagnostics also support the practical cost profile: on Lumina-Next, full AccelAes reduces latency from 12.37 s to 5.86 s (2.11×), while AesMask construction and Spatial-CFG splitting add only 0.18 s over dense inference. On the 4-step FLUX.1-schnell distilled DiT, AccelAes improves IR from 1.090 to 1.132 without degrading HPSv2 or CLIP, showing compatibility even when temporal redundancy has already been compressed by distillation. Finally, dynamic mask refresh does not improve the default static AesMask on hard non-centric prompts; frequent refresh introduces region jitter and lowers both quality and speed. We also stress tested the main failure cases by mining the worst AccelAes-to-dense ImageReward drops from dense-crowd, indoor-clutter, abstract-art, and landscape prompts. These cases reveal the expected boundary of early region selection: when attention is diffuse, an overly narrow active region can freeze small details outside the mask. The late-mask and entropy fallback rules recover 92.7–95.4% of dense ImageReward without per-prompt tuning. We therefore keep AesMask static by default and activate corrective expansion only when mask confidence becomes unreliable.

Lumina-Next-T2I Comparisons. We provide a strict comparison on the same metric set and protocol in Table 2. At a similar speedup range, several training-free baselines show large drops on ImageReward, which indicates that uniform temporal skipping or generic token reduction can be brittle when preference-oriented quality matters. For example, TaylorSeer operates in a comparable speed regime but incurs a substantial ImageReward degradation. Δ-DiT and SDiT further reduce ImageReward despite non-trivial speed gains. In contrast, AccelAes achieves the best ImageReward while also reaching the largest speedup

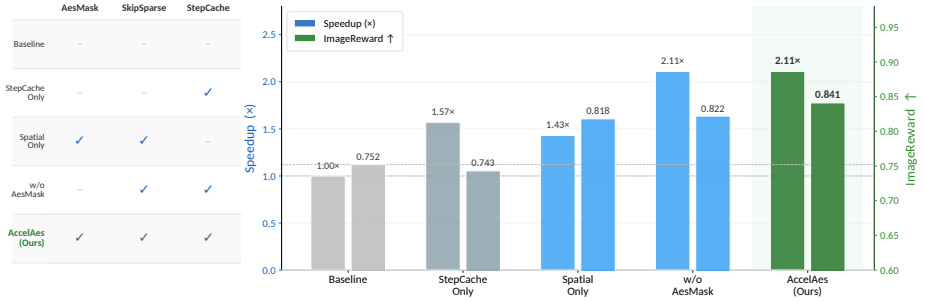

Fig. 5: SkipSparse ablation. We evaluate variants that enable spatial allocation, step-level reuse, or both, and report speedup and ImageReward with and without AesMask.

among listed methods. This gap suggests that allocating compute and guidance according to aesthetic token affinity is more robust than applying uniform reuse across all regions and steps.

SD3 and FLUX Comparisons. Fig. 3 evaluates the same design on SD3 and FLUX and highlights two practical observations. First, the relative ranking between methods can change across backbones, which suggests that aggressive temporal reuse is sensitive to the architecture and sampling configuration. Second, AccelAes remains more stable in preference proxies at practical speedups, which is consistent with its design that preserves updates in aesthetics-relevant regions instead of uniformly replacing Transformer evaluations. This supports that the benefit of aesthetics-guided spatial allocation is not restricted to Lumina-Next.

4.4 Ablation Results

AesMask Ablation. Fig. 4 compares Uniform masking, a Non-Aesthetic-Token variant, and the proposed AesMask. In Fig. 4 (a), Uniform and Non-Aesthetic-Token masks produce less targeted focus regions, while AesMask yields a compact region that aligns with aesthetic descriptors in the prompt. Fig. 4 (b) shows that this alignment translates into a more reliable quality profile on ImageReward, Aesthetic score, and CLIP score. Fig. 4 (c) further indicates that words with clear aesthetic meaning receive higher importance under anchor matching, which supports our token selection step. Together, these results suggest that AesMask provides an effective region prior for compute budgeting and region-wise guidance across prompts with different aesthetic emphasis.

SkipSparse Ablation. Fig. 5 shows distinct roles of the two SkipSparse paths. Step-level reuse achieves a larger speedup (1.57×) but slightly reduces ImageReward (0.743 vs. 0.752), which indicates that uniform temporal replacement can be fragile for preference-oriented quality. Spatial allocation improves ImageReward (0.818) with a moderate speedup (1.43×), suggesting that focusing updates on

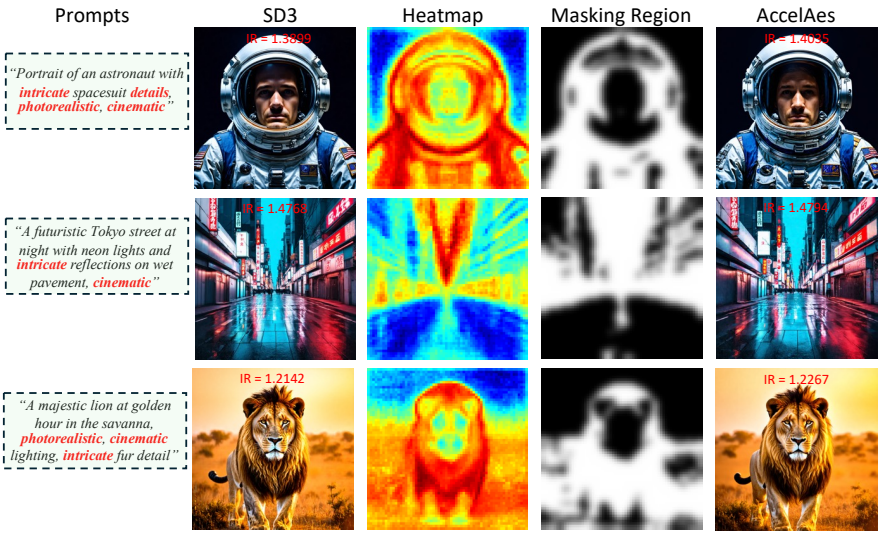

Fig. 6: Aesthetics-induced spatial non-uniformity and its effect. For representative prompts on SD3, we visualize cross-attention heatmaps, the derived aesthetic-focus mask, the selected regions, and the final AccelAes outputs. The localized reallocation of compute and guidance enhances perceptually important details, consistent with improvements in ImageReward (IR) annotated for each example.

critical regions is important. At the same speedup (2.11×), removing AesMask lowers ImageReward (0.822 vs. 0.841), showing that AesMask provides a stronger region prior that makes SkipSparse more selective. Together, these results explain why combining the two paths with AesMask yields the best trade-off.

4.5 Visualizations

Aesthetic Spatial Non-uniformity. Fig. 6 shows cross-attention maps induced by aesthetic descriptors, the resulting AesMask, selected regions, and SD3 outputs. The attention is spatially concentrated on perceptually decisive structures (e.g., faces and high-frequency details) rather than uniform. Thresholding these responses yields a compact mask that suppresses low-impact background regions. AccelAes uses AesMask to reallocate both computation and CFG toward masked tokens, and the annotated IR gains indicate a consistent preference improvement under identical prompts and sampling settings.

Qualitative Comparison across Acceleration Methods. Fig. 7 compares AccelAes with representative baselines under identical prompts and reports the corresponding speedups. Uniform skipping at higher aggressiveness often blurs fine textures or introduces local artifacts, whereas AccelAes better preserves perceptually decisive details while achieving the largest acceleration in the shown setting. This qualitative evidence is consistent with the quantitative results. Taken

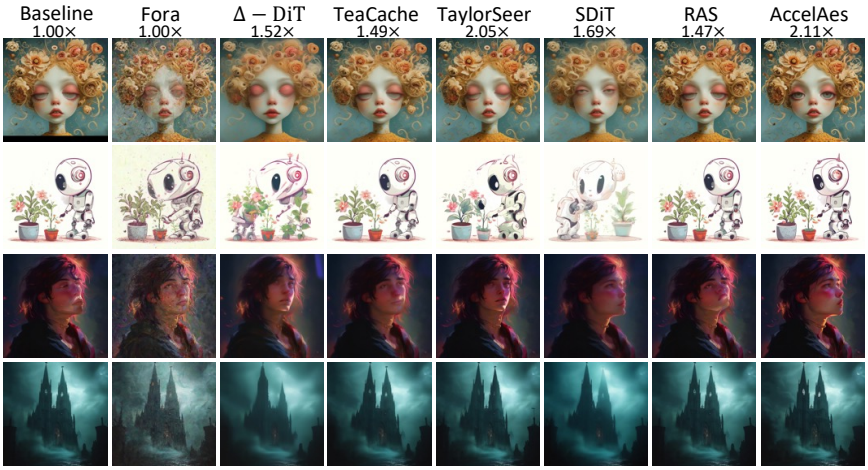

Fig. 7: Qualitative comparison across training-free acceleration methods. We compare AccelAes with the dense baseline and representative training-free acceleration baselines under identical prompts. The speedup (relative to the baseline) is shown above each column. AccelAes preserves fine textures and locally salient details more reliably at high speedup, consistent with its superior multi-metric performance in Table 2.

together, the metric comparisons, human preference study, ablations, and visual evidence indicate that AccelAes improves the practical speed–quality frontier across diverse prompt and backbone conditions by protecting aesthetics-relevant regions rather than uniformly reducing computation.

5 Conclusion

We presented **AccelAes**, a training-free acceleration framework for Diffusion Transformers that leverages aesthetics-driven spatial non-uniformity and temporal prediction redundancy in denoising. By combining **AesMask**, **SkipSparse**, and output-level step prediction reuse, AccelAes protects perceptually sensitive tokens while compressing low-impact computation across regions and timesteps. Across three representative DiT backbones, AccelAes consistently improves the speed–quality trade-off, and ablations, human preferences, and robustness checks confirm the complementarity of spatial adaptivity and temporal reuse.

The broader implication is that training-free DiT acceleration need not be purely temporal caching. Preserving computation in prompt-relevant regions connects efficiency to aesthetic preference, while the main limitation remains dependence on reliable region estimates when salient regions shift. This makes AccelAes suitable when preference quality and latency must be optimized together, especially in deployment settings with strict latency budgets. Future work may study confidence tests and finer-grained refresh schedules.

References

1. Black Forest Labs: FLUX.1 [dev] (2024), https://huggingface.co/black-forest-labs/FLUX.1-dev, accessed 28 June 2026
2. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. In: International Conference on Learning Representations (ICLR) (2023)
3. Bolya, D., Hoffman, J.: Token merging for fast stable diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 4599–4603 (2023)
4. Chen, D., Duan, Z., Li, Z., Chen, C., Chen, D., Li, Y., Chen, Y.: Attrictrl: A generalizable framework for controlling semantic attribute intensity in diffusion models. In: International Conference on Learning Representations (ICLR) (2026)
5. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In: International Conference on Learning Representations (ICLR) (2024)
6. Chen, P., Shen, M., Ye, P., Cao, J., Tu, C., Bouganis, C.S., Zhao, Y., Chen, T.: δ-dit: A training-free acceleration method tailored for diffusion transformers. arXiv preprint arXiv:2406.01125 (2024)
7. Chen, Z., Li, K., Jia, Y., Ye, L., Ma, Y.: Accelerating diffusion transformer via increment-calibrated caching with channel-aware singular value decomposition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18011–18020 (2025)
8. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
9. Chu, H., Wu, W., Feng, G., Zhang, Y.: Omnicache: A trajectory-oriented global perspective on training-free cache reuse for diffusion transformer models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16302–16312 (2025)
10. Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 16344–16359 (2022)
11. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis. In: Proceedings of the 41st International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 235, pp. 12606–12633. PMLR (2024)
12. Guo, B., Tang, S., Zeng, C., Shen, Z.: Mosaicdiff: Training-free structural pruning for diffusion model acceleration reflecting pretraining dynamics. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1655–1664 (2025)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 6840–6851 (2020)
14. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
15. Kang, H., Moon, S.D.: A review of artistic image generation and evaluation: Models, metrics, and applications. KSII Transactions on Internet and Information Systems **20**(1), 23–37 (2026)

16. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 36652–36663 (2023)
17. Li, D., Kamko, A., Akhgari, E., Sabet, A., Xu, L., Doshi, S.: Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. arXiv preprint arXiv:2402.17245 (2024)
18. Liang, Z., Yuan, Y., Gu, S., Chen, B., Hang, T., Cheng, M., Li, J., Zheng, L.: Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13199–13208 (2025)
19. Lin, B., Ye, F., Liu, Y., Guo, Z., Zhang, B., Zheng, W., Xu, Y., Xing, T., Wang, Y., Zhang, C.: Sdit: Semantic region-adaptive for diffusion transformers. arXiv preprint arXiv:2601.12283 (2026)
20. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7353–7363 (2025)
21. Liu, J., Zou, C., Lyu, Y., Chen, J., Zhang, L.: From reusing to forecasting: Accelerating diffusion models with taylorseers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15853–15863 (2025)
22. Liu, X., Li, Z., Gu, Q.: Cachequant: Comprehensively accelerated diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23269–23280 (2025)
23. Liu, Z., Yang, Y., Zhang, C., Zhang, Y., Qiu, L., You, Y., Yang, Y.: Region-adaptive sampling for diffusion transformers. arXiv preprint arXiv:2502.10389 (2025)
24. 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 (NeurIPS). pp. 5775–5787 (2022)
25. Ma, X., Fang, G., Wang, X.: Deepcache: Accelerating diffusion models for free. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15762–15772 (2024)
26. Mao, S., Guo, W., Zhang, C., Long, J., Xie, K., Cai, W.: Ctrl-z sampling: Scaling diffusion sampling with controlled random zigzag explorations. arXiv preprint arXiv:2506.20294 (2025)
27. Mo, W., Zhang, T., Bai, Y., Su, B., Wen, J.R., Yang, Q.: Dynamic prompt optimizing for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26627–26636 (2024)
28. Murray, N., Marchesotti, L., Perronnin, F.: Ava: A large-scale database for aesthetic visual analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
29. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 139, pp. 8162–8171. PMLR (2021)
30. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 162, pp. 16784–16804. PMLR (2022)

31. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 27730–27744 (2022)
32. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4195–4205 (2023)
33. Qin, Q., Zhuo, L., Xin, Y., Du, R., Li, Z., Fu, B., Lu, Y., Li, X., Liu, D., Zhu, X., et al.: Lumina-image 2.0: A unified and efficient image generative framework. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20031–20042 (2025)
34. Qiu, J., Liu, L., Wang, S., Lu, J., Chen, K., Hao, Y.: Accelerating diffusion transformer via gradient-optimized cache. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 17608–17617 (2025)
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021)
36. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022)
37. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 25278–25294 (2022)
38. Selvaraju, P., Ding, T., Chen, T., Zharkov, I., Liang, L.: Fora: Fast-forward caching in diffusion transformer acceleration. arXiv preprint arXiv:2407.01425 (2024)
39. Shen, D., Song, G., Xue, Z., Wang, F.Y., Liu, Y.: Rethinking the spatial inconsistency in classifier-free diffusion guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9370–9379 (2024)
40. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (ICLR) (2021)
41. Talebi, H., Milanfar, P.: Nima: Neural image assessment. IEEE Transactions on Image Processing (2018)
42. Wu, H., Xu, J., Le, H., Samaras, D.: Importance-based token merging for efficient image and video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4983–4995 (2025)
43. 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)
44. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 15903–15935 (2023)
45. Zhang, H., Gao, T., Shao, J., Wu, Z.: Blockdance: Reuse structurally similar spatio-temporal features to accelerate diffusion transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12891–12900 (2025)

46. Zhang, J., Xiang, C., Huang, H., Wei, J., Xi, H., Zhu, J., Chen, J.: Spargeattention: Accurate and training-free sparse attention accelerating any model inference. In: Proceedings of the 42nd International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 267, pp. 76397–76413 (2025)
47. Zhuo, L., Du, R., Xiao, H., Li, Y., Liu, D., Huang, R., Liu, W., Zhu, X., Wang, F.Y., Ma, Z., Luo, X., Wang, Z., Zhang, K., Zhao, L., Liu, S., Yue, X., Ouyang, W., Qiao, Y., Li, H., Gao, P.: Lumina-next: Making lumina-t2x stronger and faster with next-dit. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
48. Zou, C., Liu, X., Liu, T., Huang, S., Zhang, L.: Accelerating diffusion transformers with token-wise feature caching. In: International Conference on Learning Representations (ICLR) (2025)