FASTER AND BETTER REFERENCE-BASED GENERATION: EARLY TOKEN REDUCTION VIA DOWNSAMPLED SELF-PROBING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent generic reference-based image generation models natively support multiple conditions by concatenating all reference image and text tokens into a joint sequence processed by a single Diffusion Transformer (DiT). However, this versatile paradigm comes at an efficiency cost: accommodating multiple references leads to a token explosion, incurring massive $\mathcal{O}(N^2)$ computational complexity. Most acceleration methods operate inside the DiT, showing limited capability in reducing spatial redundancy and failing to alleviate peak VRAM. While some pre-model token reduction alternatives exist, they inevitably introduce external models. This raises a question: Is it possible to achieve acceleration and memory savings while preserving compatibility with in-model techniques and requiring no additional parameters? In this work, we realize this goal by proposing an early token reduction framework via downsampled self-probing. By executing a few denoising steps on spatially $2\times$ downsampled reference images, this cheap probing stage utilizes the DiT model itself to effectively and efficiently extract reference token importance. Using the probed importance, we apply a hierarchical reduction before the main denoising stage: we completely discard low-importance tokens, substitute mid-importance tokens with their downsampled counterparts to cheaply preserve global context, and keep high-importance tokens. Experiments across three base models and two benchmarks demonstrate that our early reduction accelerates inference significantly and mitigates peak GPU memory. Furthermore, we can improve generation quality by safely filtering out noisy, uninformative reference tokens, enabling the model to concentrate its limited attention capacity on genuinely useful tokens.

1 INTRODUCTION

Reference-based image generation and editing have fundamentally transformed visual synthesis, moving beyond pure text prompts to multimodal control. By allowing multiple reference inputs, these models achieve greater controllability, driving increasing demand in domains such as advertising and design. However, incorporating multiple reference images massively expands the token sequence length. Due to the $\mathcal{O}(N^2)$ nature of attention, this token explosion significantly increases both the **latency and peak VRAM**, limiting the number of references and resolutions that can be practically supported. The token explosion also degrades the generation **quality** due to: 1) Attention interference: some irrelevant reference tokens inevitably attract anomalous attention, interfering with the generation process; 2) Attention dilution: The explosion in the number



Figure 1: Compared to base model FLUX.2-klein, our method reduces inference latency while improving generation quality.

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

of tokens leads to a massive denominator in the attention Softmax, diluting weights of truly important tokens.

Despite growing works on efficient AI, a satisfying solution for generic reference-based generation is still missing. Most token reduction methods operate *inside* the DiT, pruning or merging tokens only after the full sequence has already entered the model. Therefore, they fail to reduce peak VRAM, which typically occurs at the earliest attention layers before reduction takes place. Some methods can reduce tokens before the main model, but they either degrade performance noticeably (Bolya & Hoffman, 2023) , limiting their practicality. Thus, to reduce peak memory and maintain compatibility with most DiT acceleration methods, we target an early reduction approach, meaning the token reduction is performed before the main denoising such that the entire main denoising process uses the reduced token sequence.

One straightforward early reduction method is to use an external model to identify unimportant regions (Zeng et al., 2026), which introduces extra parameters and lacks alignment with the generator's actual demands. Another method is to utilize internal features, such as attention maps, to estimate token importance. Yet, attention patterns for token importance remain unexplored in reference-based image generation, and insights from adjacent fields like ViTs/VLMs and text-to-image (T2I) are inapplicable to our task due to fundamental differences: Unlike VLMs, where visual tokens mainly provide high-level semantics, our task relies heavily on fine-grained, low-level reference details; T2I acceleration focuses on cross-timestep temporal redundancy rather than spatial token redundancy.

Thus, we firstly perform an in-depth analysis of attention, yielding two key observations: First, the total attention that generative and prompt tokens pay to a reference token, termed the *reference utilization score*, correlates well with its actual importance, thus serving as an effective proxy. Second, although some unimportant reference tokens inevitably exhibit anomalous attention scores, these abnormal tokens are stochastic and vary across layers. Furthermore, a single layer is usually biased because different layers may have different functions and focuses. Computing the mean utilization score across layers smooths out the bias and instantaneous noise observed in any individual layer, ensuring meaningless tokens are accurately identified, and yielding a robust importance estimation. This reveals a pathway to address degradation: reducing the number of uninformative tokens mitigates attention dilution and interference, and averaging across layers further mitigates interference by identifying unimportant tokens more accurately.

However, extracting attention maps during main denoising loop cannot prevent the VRAM spike and impairs compatibility with in-model acceleration methods, contradicting the goal of early reduction. Thus, we propose our **early token reduction via spatial-temporal downsampled self-probing**, which identifies unimportant tokens in a cheap probing denoising stage and prunes or replaces them prior to the main denoising. By executing a temporally downsampled, few-step denoising on spatially $2\times$ downsampled images, this probing stage cheaply approximates attention of full-resolution loop. Then, we adopt different policies for each 2×2 block of full-resolution reference tokens, guided by importance score of its corresponding downsampled reference token: High-importance blocks remain intact (preserving all 4 tokens), whereas low-importance blocks are completely discarded. To prevent quality drops from losing broader mid-importance context at high pruning rates, we introduce an intermediate level, replacing mid-importance blocks with the corresponding downsampled reference token instead of completely discarding them, cheaply capturing global context.

This early reduction yields three major benefits: (1) Because tokens are reduced before the entire main denoising stage, computational savings apply to every timestep and layer, slashing FLOPs significantly and mitigating peak VRAM. (2) It is inherently orthogonal, complementary with existing in-model acceleration techniques (e.g., feature caching and in-model token reduction). (3) The self-probing design utilizes the generator itself, naturally ensuring optimal alignment with the model's actual demands while avoiding external models.

More importantly, our method does not sacrifice performance for efficiency; rather, it comprehensively enhances generation quality. By discarding low-importance tokens, our method mitigates attention interference and dilution issues, allowing the model to better concentrate its limited attention capacity on prompt and generative tokens and genuinely informative reference tokens, ultimately improving overall generation quality.

Our key contributions are threefold: (1) We present the first in-depth analysis of attention in generic reference-based diffusion models. This analysis uncovers two key observations regarding reference

utilization, and layer-wise bias and noise, establishing a reliable proxy for reference token importance. (2) Guided by these insights, we propose a training-free, early token reduction framework, which is inherently compatible with the existing ecosystem of in-model acceleration. (3) Experiments across three base models and two benchmarks show that our method delivers up to $3\times$ mean speedups, cuts peak activation VRAM by up to 50%, and consistently improves generation quality.

2 Related Work

2.1 Reference-based Image Generation and Editing

Diffusion models have achieved success in text-to-image generation, motivating extensive work extending them to reference-conditioned generation or editing. These methods incorporate diverse visual conditions (e.g., masks, layouts, subject images, exemplar pairs, and pose maps) to achieve fine-grained controllability. Many methods (Lin et al., 2025; Sun et al., 2024; Pan et al., 2025; Wang et al., 2025a; Cha et al., 2025; Sun et al., 2026; Gong et al., 2025) integrated references via dedicated control encoders, adapters, or condition-selection mechanisms. While effective in their target tasks, they rely heavily on task-tailored architectures or input assumptions, limiting scalability.

More recently, general-purpose models have emerged to unify diverse conditioning, generation, and editing tasks within a shared DiT backbone (Black Forest Labs, 2025; Wu et al., 2025; Wang et al., 2025b; Super Intelligence Team, 2026; Zheng et al., 2026) by concatenating reference, text, and generative tokens into a single sequence. This paradigm supports arbitrary reference counts and resolutions, covering diverse downstream tasks. However, this generality comes at a steep cost: token counts and quadratic self-attention complexity explode with additional references. This work focuses on this emerging general-purpose paradigm, which is more versatile and representative of future directions.

2.2 Token Reduction for Vision Models

Token reduction, also named as token compression, has been explored to improve vision model efficiency (including ViTs and MLLMs) by shortening the token sequence (Shao et al., 2026a; Yao et al., 2026; Kong et al., 2025; Pathak & Han, 2026). One major category is *in-model* token reduction, which progressively discards or merges tokens during the forward pass of the main model (Li et al., 2026a; Choi et al., 2025). Another category is *pre-model* token reduction, which compresses tokens before entering the main model. Among these, training-free methods like similarity-based token merging (Bolya & Hoffman, 2023) achieve this goal without introducing extra trainable modules, though their performance is bounded. Some works (Yang et al., 2025a; Deng et al., 2026; Shang et al., 2025; Arif et al., 2025) exploit CLS-to-token attention from a vision encoder to prune tokens before feeding them to VLMs. Other works (Zeng et al., 2026; Lu et al., 2026) adopt more precise methods to predict region importance, but at the cost of introducing new external models.

2.3 Efficient Image and Video Generation

Diffusion acceleration has been investigated from multiple complementary perspectives. Cross-step caching methods (Ma et al., 2024b; Chen et al., 2024b; Zou et al., 2025; Liu et al., 2024a; Zhang et al., 2024b; Ma et al., 2024a; Zhao et al., 2025; Son et al., 2025; Shao et al., 2026b; Yuan et al., 2024; Zhu et al., 2026a) reuse features or attention maps across denoising steps to eliminate redundant computation. Distillation methods (Luo et al., 2023; Sauer et al., 2024) reduce sampling steps, model pruning (Castells et al., 2024; Fang et al., 2023; Zhu et al., 2024) removes structural weights, quantization (Li et al., 2023; Chen et al., 2025a) lowers arithmetic precision, and in-model token reduction methods (Chang et al., 2025; Fang et al., 2025; Zhu et al., 2026b; Cheng et al., 2025; Li et al., 2026b; Zhang et al., 2024a) compress tokens inside models.

Another direction is efficient attention (Zhou et al., 2026; Liu et al., 2024b; Xie et al., 2024; Liu et al., 2025b), which develop linear or log-linear attention variants. However, these methods only reduce computation inside attention, thus cannot save computation in Feed-Forward Networks.

2.4 Efficient Reference-Based Generation and Editing

Closer to our reference-based generation setting, several works accelerate specific conditional generation scenarios. RegionE and SpotEdit (Qin et al., 2026; Chen et al., 2025b) reduce computation on unchanged regions in image editing, but they are limited to layout-preserving single-image editing setting. OmniControl2 (Tan et al., 2025) removes near-zero tokens in Canny edge maps and requires

re-training. EVCtrl (Yang et al., 2025b) reduces computation within ControlNet, and ToPi (Lin et al., 2026) accelerates in-context generation by pruning tokens inside the main denoising stage.

To the best of our knowledge, an early-stage reference-based generation acceleration method that reduces reference tokens before the main denoising run is still absent. By shortening the input sequence prior to the DiT in main denoising loop, our approach is inherently orthogonal and complementary to in-model techniques (e.g., caching, quantization, efficient attention, KV sparsification, and in-model token reduction).

3 METHOD

3.1 PRELIMINARIES / BACKGROUND AND NOTATIONS

Generic reference-based diffusion models (e.g., Flux.2) unify visual conditions by concatenating target generative tokens G, text prompt tokens P (from a text encoder), and clean reference tokens R (from a frozen VAE) into a joint sequence processed by an L-layer Diffusion Transformer (DiT). Let $G^{t,l}$, $P^{t,l}$, and $R^{t,l}$ denote token states at timestep t and layer l. Denoising starts from pure noise $G^{T,0}$ down to $G^{0,L}$ for final decoding. At step t, reference tokens are first modulated as $R^{t,0} = \text{Modulate}(R, t)$. After passing through all L layers, the generative output $G^{t,L}$ is updated via a numerical solver as $G^{t-1,0} = \mathcal{F}_{\text{solver}}\big(G^{t,0}, G^{t,L}\big)$.

For the importance score I, omitting a superscript denotes averaging over the corresponding axis: omitting l denotes averaging over layers ($\frac{1}{L}\sum_l[\cdot]$), omitting t denotes averaging over timesteps ($\frac{1}{T}\sum_t[\cdot]$), and omitting both denotes averaging over both axes ($\frac{1}{LT}\sum_t\sum_l[\cdot]$). For all other variables (e.g., A, G, P, R), omitting t or l is strictly notational brevity and does not imply averaging.

3.2 INVESTIGATION OF ATTENTION IN REFERENCE-BASED GENERATION / ESTIMATION OF REFERENCE TOKEN IMPORTANCE

Attention scores are widely adopted as a proxy for token importance in ViTs and VLMs, where prior works observe the rule that visual information rapidly aggregates into a few "anchor" tokens during early layers, allowing most computationally silent vision tokens to be safely pruned (Yang et al., 2025a; Chen et al., 2024a). While visual tokens in VLMs primarily serve high-level semantic understanding, our task relies heavily on them to continuously provide fine-grained, low-level spatial details for generation. Thus, the rules observed in VLMs cannot be trivially transferred to our task. Furthermore, acceleration methods in text-to-image (T2I) generation primarily focus on cross-timestep temporal feature similarity. Given that the intermediate feature dynamics within generic reference-based generation models remain largely unexplored, we first perform an in-depth analysis of the attention patterns specific to this task.

3.2.1 REFERENCE UTILIZATION SCORE

We first explore how to quantify reference token importance using the model's intrinsic attention behaviour. Intuitively, a reference token's importance should be proportional to the attention it attracts during the denoising pass. Formally, let $A^{t,l}_{i,j} = \frac{1}{H}\sum_{h=1}^{H} q^{t,l}_{i,h}(k^{t,l}_{j,h})^\top$ denote the head-averaged attention score from query i to key j (for notational simplicity, we omit the normalization like Softmax here). For a given reference token r, we define the reference utilization score queried by a specific query set Ω as $I^{t,l}_r(\Omega) = \sum_{i\in\Omega} A^{t,l}_{i,r}$ and evaluate query combinations, including $\Omega \in \{G, P, G\cup P, G\cup P\cup R\}$: $\Omega = G$ (generative tokens only), $\Omega = P$ (prompt tokens only), $\Omega = G\cup P$, and $\Omega = G\cup P\cup R$. To establish a pseudo ground-truth, we define the oracle importance I^*_r as the gradient magnitude of the final decoded image pixels with respect to each reference token: $I^*_r = \big\|\nabla_{R_r}\big(\|\,\text{Decode}(G^{0,L})\|_2\big)\big\|_2$,

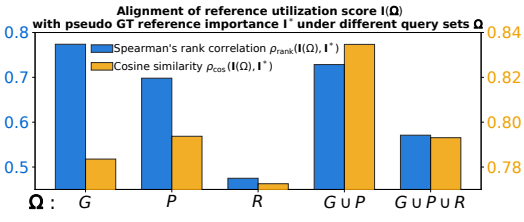


Figure 2: Alignment of reference utilization scores with pseudo ground-truth importance across different query sets (Ω). The set ($\Omega = G \cup P$) consistently yields the best overall alignment, demonstrating that querying reference tokens via both generative and prompt tokens optimally proxies the importance.

which reflects each reference token's impact on the generated result (Abnar & Zuidema, 2020). Then, we evaluate the alignment between this gold standard and our proposed attention-based proxy via Spearman's rank correlation $\rho_{\text{rank}}(\mathbf{I}(\Omega), \mathbf{I}^*)$, which evaluates relative token prioritization, and cosine similarity $\rho_{\cos}(\mathbf{I}(\Omega), \mathbf{I}^*)$.

We conduct this analysis on about 120 samples and take the average across all layers. As shown in Fig. 2, the combined queries ($\Omega = G \cup P$) consistently yield the highest alignment, meaning the value of a reference token can be effectively approximated by how intensely it is queried by the generative target and the prompt semantics. Intuitively, while $R \to R$ attention only captures task-agnostic reference structure, attention originating from generative tokens G and prompt P directly reflects actual conditioning demands for target generation. Consequently, we adopt the reference utilization score queried by G and P, formulated as $I_r = \frac{1}{LT} \sum_t \sum_l \sum_{i \in G \cup P} A_{i,r}^{t,l}$ (hereafter simply referred to as the Reference Utilization Score), as the proxy for the importance of a reference token r.

3.2.2 Layer-Wise Analysis of Utilization Score

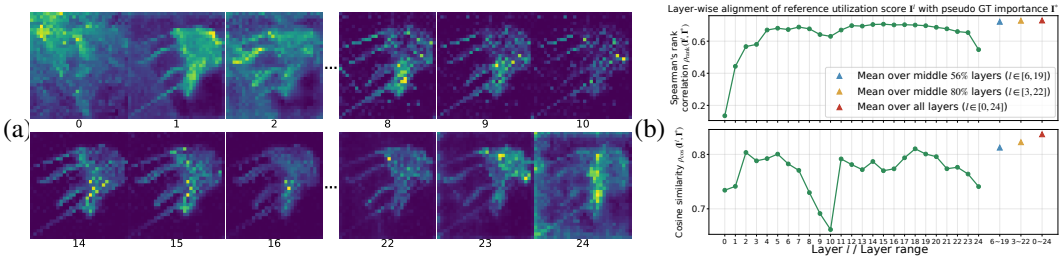


Figure 3: Analysis of layer-wise attention dynamics. (a) Layer-wise visualizations of the reference utilization score show stochastic anomalous values, and a single layer is usually biased because different layers may have different functions and focuses. (b) Computing the mean utilization score across layers smooths out the bias and instantaneous noise, achieving better alignment with the oracle token importance.

As the utilization score varies across layers, we analyze its layer-wise behavior to guide the design of the final token importance estimation. Visualizing the layer-wise attention maps of the reference utilization score $I_r^l = \frac{1}{T} \sum_t \sum_{i \in G \cup P} A_{i,r}^{t,l}$ reveals a three-stage dynamic (Fig. 3). The early layers exhibit extensive attention, likely acting as low-level feature extractors. The middle layers then transition into highly concentrated patterns that strongly correlate with the actual salient regions in the references requested by the text prompt. The final layers show dispersed attention, decoupling from the input conditions entirely, and shift their capacity toward refining the generative tokens G, largely ignoring the input references.

We also show quantitative statistics in Fig. 3(b), which plots the alignment between the layer-wise reference utilization score and the oracle importance from layer 0 to 24, measured by both rank correlation (top) and cosine similarity (bottom). Consistent with our visual observations, the middle layers exhibit better correlation compared to the early and late layers. However, relying on any single layer remains highly susceptible to localized and anomalous token spikes.

Fortunately, as we can see in Fig. 3(a), these abnormal attention spikes are stochastic and vary across different layers. Thus, they can effectively cancel each other out when aggregated. As shown on the right side of Fig. 3(b), averaging the scores across multiple layers consistently outperforms any single layer. While averaging the middle 56% or 80% of layers and averaging all 100% of layers both yield excellent alignment, we simply average across all layers for better generalization across different base models. By averaging the utilization score across all layers $I_r = \frac{1}{L} \sum_l I_r^l = \frac{1}{LT} \sum_l \sum_t \sum_{i \in G \cup P} A_{i,r}^{t,l}$, we apply an implicit low-pass filter that smooths these fluctuations and combines the different focuses of individual layers, yielding a relatively unbiased, comprehensive, and robust importance estimation.

3.3 Early Token Reduction via Downsampled Self-Probing

Although the full-resolution attention score is an accurate proxy for token importance, extracting it during the main denoising loop violates our early reduction goal. Thus, we propose using a sepa-

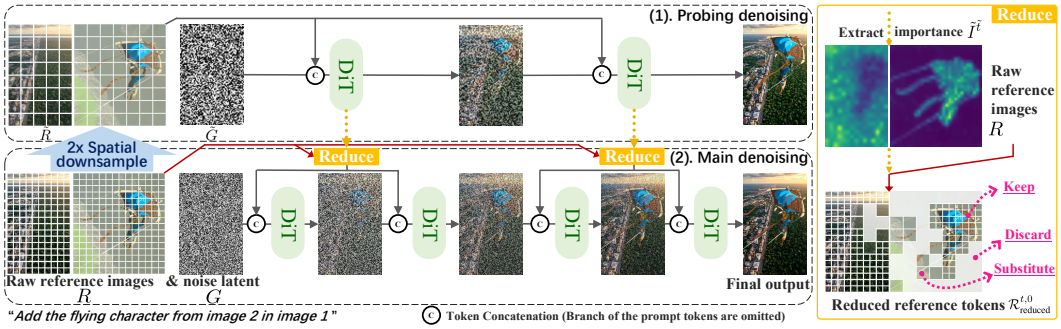


Figure 4: The illustration of our token reduction by a spatial-temporal downsampled self-probing. By executing a downsampled time step schedule denoising loop on spatially $2\times$ downsampled input images, this probing stage cheaply extracts the reference importance. The extracted importance scores then guide the reduction strategy outside the main denoising run. The visual example is generated using FLUX.2-Klein-4B (4-step).

rate lightweight loop, termed the *probing stage*, prior to the formal denoising stage. Unimportant reference tokens are identified and reduced before they enter the DiT in the main denoising loop.

3.3.1 Extracting Importance via Spatial-Temporal Downsampled Self-Probing

To cheaply identify redundant tokens, the probing stage evaluates a temporally sparse DiT denoising loop on spatially $2\times$ downsampled inputs. Let $\tilde{\cdot}$ designate the downsampled probing stage, meaning tilde-accented variables are evaluated on downsampled inputs, whereas unaccented variables refer to the standard full-resolution run. We extract the reference utilization score in the probing stage $\tilde{I}^{\tilde{t}}_{\tilde{r}} = \frac{1}{L}\sum_l \sum_{i\in\tilde{G}\cup P} \tilde{A}^{\tilde{t},l}_{i,\tilde{r}}$, where vision tokens operate at a reduced spatial scale: $\tilde{R}$ (whose individual tokens are denoted by $\tilde{r}$) is obtained by VAE-encoding the downsampled RGB references, and $\tilde{G}$ represents the generative noisy tokens sampled at the reduced latent resolution. $\tilde{t}$ is a timestep in the sparse temporal schedule, meaning the probe denoising loop executes an accelerated trajectory over a coarse time grid of only T/D steps, where $D = \max(2, \lfloor T/4 \rfloor)$ is the temporal reduction factor.

This probe stage shares completely the same configuration as the formal denoising stage, except that it operates on the spatially-downsampled image tokens and undergoes a temporally-downsampled DiT denoising loop (as defined by $\tilde{t}$). Thus, the extracted $\tilde{I}$ effectively approximates the reference utilization score I of the main denoising.

3.3.2 Reference Token Reduction

To better illustrate the spatial correspondence, we introduce (w, h, n) as an alternative notation for the token index r, representing the token at position (w, h) in the n-th reference image. Due to the $2\times$ spatial downsampling, each downsampled token covers a 2×2 block of full-resolution tokens. Based on this, we apply different policies (preserve, discard, or substitute) for each 2×2 block of full-resolution reference tokens $R^{t,0}_{(w,h,n)}$, guided by the importance score of its corresponding downsampled reference token $\tilde{I}^{\tilde{t}}_{(\tilde{w},\tilde{h},n)}$: (1) **Keep High-importance blocks (Full-Res Tier)** keep all 4 full-resolution tokens intact ($4 \to 4$), preserving fine-grained details. (2) **Discard Low-importance blocks (Discard Tier)** are completely discarded ($4 \to 0$), slashing the computational overhead of irrelevant background noise. (3) **Substitute Mid-importance blocks (Low-Res Tier)**: Since a simple binary strategy (keeping high-importance blocks and discarding all others) only preserves a few highly salient regions and strips away excessive global context, we introduce an intermediate level. We replace mid-importance blocks with their corresponding downsampled tokens instead of fully discarding them. This captures global context at just $1/4$ the cost.

The resulting reduced sequence $\mathcal{R}^{t,0}_{\text{reduced}}$ is then fed into the DiT for the formal denoising stage. This tri-tier reduction can be formulated as:

$$\mathcal{R}^{t,0}_{\text{reduced}} = \underbrace{\left\{R^{t,0}_{(w,h,n)} \mid \tilde{I}^{\tilde{t}}_{(\tilde{w},\tilde{h},n)} > \tau_1\right\}}_{\text{Preserve high-importance tokens}} \cup \underbrace{\left\{\tilde{R}^{t,0}_{(\tilde{w},\tilde{h},n)} \mid \tau_2 \le \tilde{I}^{\tilde{t}}_{(\tilde{w},\tilde{h},n)} \le \tau_1\right\}}_{\text{Substitute mid-importance tokens}}, \tag{1}$$

where $\begin{cases} \tilde{w} = \lfloor w/2 \rfloor,\ \tilde{h} = \lfloor h/2 \rfloor \\ \tilde{t} = \lfloor t/D \rfloor \cdot D \end{cases}$, $\tilde{I}^{\tilde{t}}_{(\tilde{w},\tilde{h},n)} = \frac{1}{L} \sum_l \sum_{i \in \tilde{G} \cup P} \tilde{A}^{\tilde{t},l}_{i,(\tilde{w},\tilde{h},n)}$.

The core rationale behind the probing stage is to approximate the main denoising behavior with minimal overhead: Limiting the spatial downsampling to only $2\times$ ensures the probing denoising closely approximates the behaviour of main denoising. Meanwhile, this $2\times$ downsampling cuts the 2D token count to $1/4$. Considering attention only, the $\mathcal{O}(N^2)$ complexity reduces its cost by $16\times$; with at least $2\times$ temporal downsampling, the probe is about $\frac{1}{32}$ of the main denoising, making its overhead negligible relative to the acceleration (rigorous derivation in Appendix A.1).

Unlike methods that introduce external models, the self-probing design naturally ensures optimal alignment with the generator's actual demands while avoiding the introduction of any extra parameters. The spatial-temporal downsampling ensures that the probe introduces only minor overhead, which is significantly outweighed by the acceleration achieved through token reduction. This early reduction slashes FLOPs and peak VRAM while remaining orthogonal and compatible with existing in-model acceleration techniques, such as feature caching.

4 Experiment

4.1 Setup

We evaluate on three base models: Flux.2-Klein-4B, Qwen-Image-Edit-2511, and HiDream-O1-Image. See details in the Appendix B.1. We evaluate our method on two datasets: MultiBanana (Oshima et al., 2026) and Macro (Chen et al., 2026), which both score the generation quality from multiple dimensions. For efficiency, we report end-to-end latency, peak activation VRAM, and speedup over the corresponding base model. Details, including the full names of these dimensions and subtasks, are in the appendix B.3.

4.2 Main results

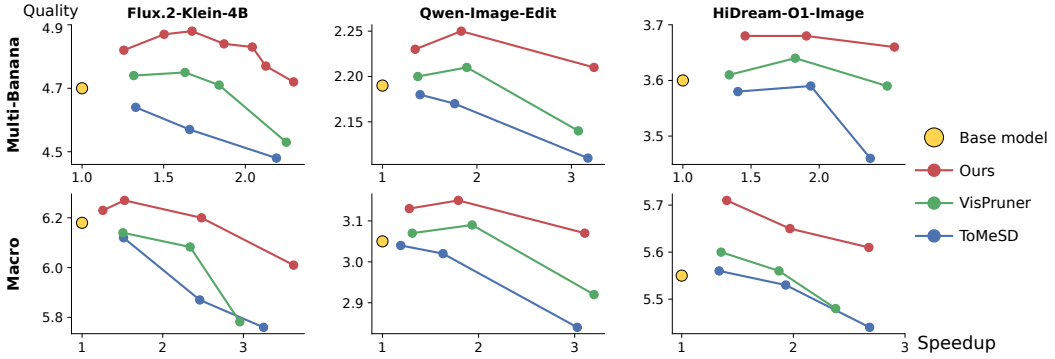


Figure 5: Average quality versus speedup across three base models and two benchmarks.

Method	MultiBanana									MACRO							
	Quality↑						Efficiency			Quality↑					Efficiency		
	Inst. Align	Ref. Cons.	B-S Match.	Phys. Real.	Visu. Qual.	**Avg**	Latency (s)↓	Speed↑	Act. VRAM (GB)↓	Cust.	Illus.	Spat.	Temp.	**Avg**	Latency (s)↓	Speed↑	Act. VRAM (GB)↓
Baseline	4.57	4.86	3.96	4.40	5.72	4.70	5.74	1.00×	15.3	7.49	6.17	3.83	7.23	6.18	7.93	1.00×	24.7
Fast																	
ToMeSD	4.45	4.51	3.72	4.17	5.55	4.48	2.62	2.19×	11.2	7.10	5.82	3.45	7.10	5.87	3.23	2.46×	17.4
VisPruner	4.54	4.53	3.72	4.24	5.62	4.53	2.55	2.25×	8.7	7.27	5.99	3.87	7.20	6.08	3.39	2.34×	13.7
Ours	**4.63**	**4.85**	**3.97**	**4.38**	**5.77**	**4.72**	2.50	2.30×	7.3	**7.50**	**6.15**	**3.88**	**7.27**	**6.20**	3.20	2.48×	11.5
Slow																	
ToMeSD	4.48	4.53	3.82	4.38	5.62	4.57	3.46	1.66×	10.7	7.35	6.18	3.70	7.25	6.12	5.24	1.51×	17.4
VisPruner	4.68	4.85	3.90	4.44	5.88	4.75	3.52	1.63×	10.2	7.37	6.03	3.91	7.25	6.14	5.27	1.51×	17.2
Ours	**4.86**	**5.06**	**4.03**	**4.54**	**5.92**	**4.88**	3.43	1.67×	9.8	**7.55**	**6.20**	**4.01**	**7.32**	**6.27**	5.20	1.53×	16.4

Table 1: Comparison of efficiency and quality on MultiBanana and MACRO (Flux.2-Klein-4B).

Fig. 5 presents the main comparison across three base models and two benchmarks, against the original base model (**Baseline**), ToMeSD (Bolya & Hoffman, 2023), and VisPruner (Zhang et al., 2025).

Table 1 reports the fine-grained scores of this comparison on Flux.2-Klein-4B. **Ours(fast)** accelerates inference and lowers peak VRAM, while matching or surpassing baseline quality; **Ours(slow)** further improves all quality metrics, at a smaller speedup. VisPruner can improve quality over the baseline at some speedup, while its quality stays below ours because it identifies unimportant tokens less accurately. Fig. 6 shows visual results.

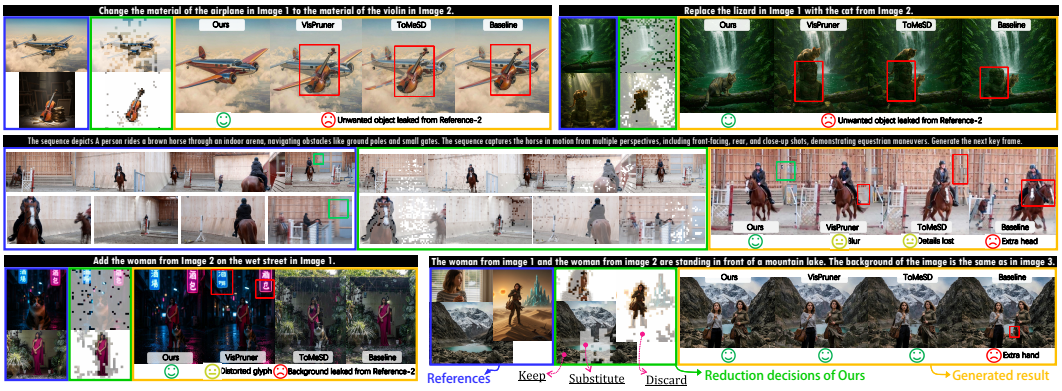


Figure 6: Visual comparison of generation quality across diverse tasks from MultiBanana and Macro datasets (all methods are under the Slow level in Table 1 with comparable speedups). Best viewed when zoomed in.

4.3 Orthogonality with in-model methods

We compare with in-model acceleration methods including TaylorSeer (Liu et al., 2025a), RepreShift (Choi et al., 2025), and TransPrune (Li et al., 2026a) in Table 2. As demonstrated in Table 2, integrating our method with existing in-model methods yields compounded speedups. Detailed discussion of our orthogonality with other existing methods can be found in Appendix A.2.

Method	MultiBanana									MACRO							
	Quality↑						Efficiency			Quality↑					Efficiency		
	Inst. Align	Ref. Cons.	B-S Match.	Phys. Real.	Visu. Qual.	**Avg**	Latency (s)↓	Speed↑	Act. VRAM (GB)↓	Cust.	Illus.	Spat.	Temp.	**Avg**	Latency (s)↓	Speed↑	Act. VRAM (GB)↓
Baseline	4.57	4.86	3.96	4.40	5.72	4.70	5.7	1.00×	15.3	7.49	6.17	3.83	7.23	6.18	7.9	1.00×	24.7
Ours	4.63	4.85	3.97	4.38	5.77	4.72	2.50	2.30×	7.3	7.50	6.15	3.88	7.27	6.20	3.20	2.48×	11.5
RepreShift	4.51	4.69	3.93	4.44	5.78	4.67	3.5	1.63×	11.3	7.45	6.12	3.83	7.11	6.13	4.1	1.93×	22.1
RepreShift + Ours	4.61	4.66	3.96	4.40	5.72	4.67	2.4	2.38×	8.3	7.44	6.10	3.85	7.11	6.12	3.0	2.65×	15.2
Taylorseer	4.52	4.69	3.85	4.41	5.64	4.62	4.8	1.19×	47.5	7.13	5.63	3.16	6.82	5.69	6.5	1.21×	62.0
Taylorseer + Ours	4.55	4.72	3.90	4.44	5.70	4.66	3.1	1.85×	34.3	7.21	5.47	3.26	6.86	5.70	4.8	1.65×	45.4
TransPrune	4.56	4.85	3.92	4.43	5.81	4.71	3.2	1.77×	14.9	7.52	6.15	3.87	7.17	6.18	3.3	2.40×	23.3
TransPrune + Ours	4.58	4.82	3.94	4.38	5.77	4.70	2.1	2.79×	8.4	7.48	6.14	3.88	7.13	6.16	2.5	3.12×	15.6

Table 2: Compatibility of Ours(fast) with existing in-model acceleration methods on FLUX.2-Klein-4B, yielding compounded inference speedups.

4.4 Ablation study

Table 3 validates our designs. "All Low-Res Tier" operates on the 2× downsampled references, suffering a severe quality drop. Removing the intermediate tier ("w/o Low-Res Tier") causes a noticeable quality decline, confirming the necessity of capturing global context via Low-Res Tier tokens. "Random Reduction" also degrades quality, indicating the extracted reference utilization score is necessary for token reduction decisions.

Method	Instruction Alignment↑	Reference Consistency↑	BG-Subject Match↑	Physical Realism↑	Visual Quality↑	**Overall↑**
Only Low-Res Tier (i.e., infer baseline on 2× downsampled images)	4.25	4.10	3.53	4.12	5.63	4.33
No Low-Res Tier	4.56	4.67	3.88	4.37	5.72	4.64
Random Reduction	4.25	4.08	3.62	4.22	5.60	4.35
Ours	**4.63**	**4.85**	**3.97**	**4.38**	**5.77**	**4.72**

Table 3: Ablation study on Ours(fast) on MultiBanana (controlled to the same reduction ratio). Base model: Flux.2-Klein-4B.

4.5 Analysis of Quality Improvement

4.5.1 Effect of Token Reduction Ratio

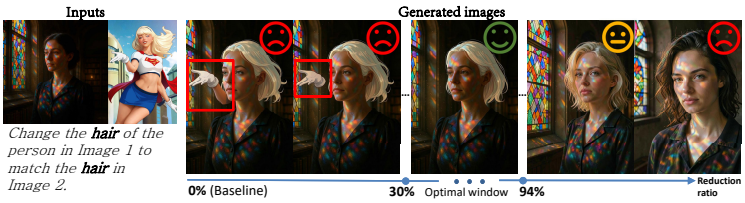


Figure 7: Qualitative evolution of generated images. At 0% reduction (unpruned baseline), the model mistakenly retains visual distractors (e.g., an extra floating glove). Within a broad window (30% to 94%), our method successfully filters out these distractors, boosting quality. At extreme reduction levels ($> 95\%$), the irreversible loss of informative reference tokens causes quality degradation.

Fig. 7 shows how quality evolves with the reduction ratio, driven by two competing factors: (1) mitigating attention interference and dilution versus (2) losing informative tokens at extreme sparsity. At 0% reduction, the baseline leaks a floating glove from a salient but irrelevant region in ref 2; Moving into a broad optimal window, Factor 1 dominates, and the model successfully filters out these visual distractors. As the reduction ratio escalates to extreme levels (>95%), Factor 2 eventually takes over. Crucial features are irreversibly pruned, resulting in degraded reference consistency. Note that this **optimal window** forms a **wide** plateau rather than a narrow range, as demonstrated by the case in Fig. 7 and quantitative curve in Fig. 5.

4.5.2 Effective Attention Ratio

Our method can effectively mitigate attention dilution and attention interference, thus making the model allocate its limited attention capacity to genuinely useful reference tokens. Here, we explicitly validate whether the model's attention is more focused on meaningful reference regions, by evaluating the Effective Attention Ratio (EAR).

Method	Effective Attention Ratio↑	Instruction Alignment↑	Reference Consistency↑	BG-Subject Match↑	Physical Realism↑	Visual Quality↑	Overall↑
Baseline (No Reduction)	43.6	4.57	4.86	3.96	4.40	5.72	4.70
Top-1 layer	51.7	4.81	5.08	3.92	4.42	5.85	4.82
Top 20% of layers	**52.9**	**4.89**	**5.13**	4.01	4.47	**5.93**	**4.89**
All layers	52.7	4.86	5.06	**4.03**	**4.54**	5.92	4.88

Table 4: Effective Attention Ratio (EAR) on all samples from the object composition task of MultiBanana. Baseline: Flux.2-Klein-4B. All variants use the same reduction ratio.

To calculate EAR, an objective ground-truth (GT) importance map is required. While general reference-based tasks lack such GT importance, the object composition task in MultiBanana (Oshima et al., 2026) inherently aims to composite the salient object in each reference image. Thus, the salient object region (extracted via BiRefNet (Zheng et al., 2024)) naturally serves as the GT important area. A higher ratio indicates that the model allocates more attention to genuinely useful regions. We evaluate four variants: (1) *Baseline (No Reduction)*, the unpruned reference sequence; (2) *Top-1 layer*, using the utilization score from the best-aligned single layer; (3) *Top 20% of layers*, using the mean over the top 20% best-correlated layers; and (4) *All layers*, using the mean across all layers. Details are in Appendix B.5.

Table 4 shows two gains. Token reduction with *Top-1 layer* markedly raises EAR, validating that our token reduction makes the model allocate more attention to genuinely useful reference regions. *Top 20% of layers* and *All layers* further slightly raise EAR over *Top-1 layer*, validating that averaging across layers further mitigates interference by identifying unimportant tokens more accurately.

All layers is comparable to *Top 20% of layers*, but does not require model-specific layer selection. We therefore adopt averaging across all layers as the default for generalization across base models.

5 Conclusions

In this work, we propose an early token reduction framework via downsampled self-probing to address the token explosion in reference-based diffusion models. By utilizing the model itself to cheaply extract token importance, we perform a tri-tier reduction before the main denoising stage. Experiments show that our training-free method reduces latency, mitigate peak VRAM, seamlessly compounds with in-model acceleration techniques, and improves generation quality.

AI USE STATEMENT

In this work, we used generative AI tools for research execution, including assistance with implementation and with checking the complexity derivation. Additionally, we used generative AI tools to aid and polish writing, and for retrieval and discovery of related work. We have reviewed all AI-assisted work: suggested literature was checked against the sources, and AI-assisted text, code, and derivations were verified by the authors before use. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Implementation and evaluation protocols are in Sec. B. We will release the full code upon acceptance.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4190–4197, 2020.

Kazi Hasan Ibn Arif, JinYi Yoon, Dimitrios S Nikolopoulos, Hans Vandierendonck, Deepu John, and Bo Ji. Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models. 39(2):1773–1781, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. In *IEEE Conference on Computer Vision and Pattern Recognition – Workshops (CVPRW)*, 2023. URL https://arxiv.org/abs/2303.17604.

Qi Cai, Jingwen Chen, Chengmin Gao, Zijian Gong, Yehao Li, Yingwei Pan, Yi Peng, Zhaofan Qiu, Kai Yu, Yiheng Zhang, Hao Ai, Siying Bai, Yang Chen, Zhihui Chen, Fengbin Gao, Ying Guo, Dong Li, Zhen Shen, Leilei Shi, Jing Wang, Siyu Wang, Yimeng Wang, Rui Zheng, Ting Yao, and Tao Mei. Hidream-o1-image: A natively unified image generative foundation model with pixel-level unified transformer. arXiv preprint arXiv:2605.11061, 2026. URL https://arxiv.org/abs/2605.11061.

Thibault Castells, Hyoung-Kyu Song, Bo-Kyeong Kim, and Shinkook Choi. Ld-pruner: Efficient pruning of latent diffusion models using task-agnostic insights. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 821–830, 2024.

Hyunsoo Cha, Byungjun Kim, and Hanbyul Joo. Durian: Dual reference image-guided portrait animation with attribute transfer. arXiv preprint arXiv:2509.04434, 2025. URL https://arxiv.org/abs/2509.04434.

Shuning Chang, Pichao Wang, Jiasheng Tang, Fan Wang, and Yi Yang. Sparsedit: Token sparsification for efficient diffusion transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2412.06028.

Lei Chen, Yuan Meng, Chen Tang, Xinzhu Ma, Jingyan Jiang, Xin Wang, Zhi Wang, and Wenwu Zhu. Q-dit: Accurate post-training quantization for diffusion transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a. URL https://arxiv.org/abs/2406.17343.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024a.

Pengtao Chen, Mingzhu Shen, Peng Ye, Jianjian Cao, Chongjun Tu, Christos-Savvas Bouganis, Yiren Zhao, and Tao Chen. Δ-dit: A training-free acceleration method tailored for diffusion transformers. *arXiv preprint arXiv:2406.01125*, 2024b.

Pengtao Chen, Xianfang Zeng, Maosen Zhao, Mingzhu Shen, Peng Ye, Bangyin Xiang, Zhibo Wang, Wei Cheng, Gang Yu, and Tao Chen. Regione: Adaptive region-aware generation for efficient image editing. *arXiv preprint arXiv:2510.25590*, 2025b.

Zhekai Chen, Yuqing Wang, Manyuan Zhang, and Xihui Liu. Macro: Advancing multi-reference image generation with structured long-context data. *arXiv preprint arXiv:2603.25319*, 2026. URL https://arxiv.org/abs/2603.25319.

Xinle Cheng, Zhuoming Chen, and Zhihao Jia. Cat pruning: Cluster-aware token pruning for text-to-image diffusion models. arXiv preprint arXiv:2502.00433, 2025. URL https://arxiv.org/abs/2502.00433.

Joonmyung Choi, Sanghyeok Lee, Byungoh Ko, Eunseo Kim, Jihyung Kil, and Hyunwoo J. Kim. Representation shift: Unifying token compression with flashattention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

LMDeploy Contributors. Lmdeploy: A toolkit for compressing, deploying, and serving llm. https://github.com/InternLM/lmdeploy, 2023.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022.

Jinhong Deng, Wen Li, Joey Tianyi Zhou, and Yang He. Scope: Saliency-coverage oriented token pruning for efficient multimodel llms. *Advances in Neural Information Processing Systems*, 38: 161527–161552, 2026.

Gongfan Fang, Xinyin Ma, and Xinchao Wang. Structural pruning for diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Haipeng Fang, Sheng Tang, Juan Cao, Enshuo Zhang, Fan Tang, and Tong-Yee Lee. Attend to not attended: Structure-then-detail token merging for post-training dit acceleration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2505.11707.

Yan Gong, Yiren Song, Yicheng Li, Chenglin Li, and Yin Zhang. Relationadapter: Learning and transferring visual relation with diffusion transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2506.02528.

Zhenglun Kong, Yize Li, Fanhu Zeng, Lei Xin, Shvat Messica, Xue Lin, Pu Zhao, Manolis Kellis, Hao Tang, and Marinka Zitnik. Token reduction should go beyond efficiency in generative models – from vision, language to multimodality. *arXiv preprint arXiv:2505.18227*, 2025.

Ao Li, Yuxiang Duan, Jinghui Zhang, Congbo Ma, Yutong Xie, Gustavo Carneiro, Mohammad Yaqub, and Hu Wang. Transprune: Token transition pruning for efficient large vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026a.

Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://arxiv.org/abs/2302.04304.

Zhuojin Li, Hsin-Pai Cheng, Hong Cai, Shizhong Han, and Fatih Porikli. Coredit: Spatial coherence-guided token pruning and reconstruction for efficient diffusion transformers. arXiv preprint arXiv:2605.14191, 2026b. URL https://arxiv.org/abs/2605.14191.

Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2404.09967.

Junqing Lin, Xingyu Zheng, Pei Cheng, Bin Fu, Jingwei Sun, and Guangzhong Sun. Token pruning for in-context generation in diffusion transformers. *arXiv preprint arXiv:2602.01609*, 2026.

Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model. *arXiv preprint arXiv:2411.19108*, 2024a.

Jiacheng Liu, Chang Zou, Yuanhuiyi Lyu, Junjie Chen, and Linfeng Zhang. From reusing to forecasting: Accelerating diffusion models with taylorseers. pp. 15853–15863, 2025a.

Songhua Liu, Weihao Yu, Zhenxiong Tan, and Xinchao Wang. Linfusion: 1 gpu, 1 minute, 16k image. 2024b.

Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Clear: Conv-like linearization revs pre-trained diffusion transformers up. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b. URL https://arxiv.org/abs/2412.16112.

Yu Lu, Tianyi Wang, Fengyun Rao, Yi Yang, Linchao Zhu, et al. Flexselect: Flexible token selection for efficient long video understanding. *Advances in Neural Information Processing Systems*, 38: 102751–102777, 2026.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023.

Xinyin Ma, Gongfan Fang, Michael Bi Mi, and Xinchao Wang. Learning-to-cache: Accelerating diffusion transformer via layer caching. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a. URL https://arxiv.org/abs/2406.01733.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b. URL https://arxiv.org/abs/2312.00858.

Yuta Oshima, Daiki Miyake, Kohsei Matsutani, Yusuke Iwasawa, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. Multibanana: A challenging benchmark for multi-reference text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 448–460, June 2026.

Yanjie Pan, Qingdong He, Zhengkai Jiang, Pengcheng Xu, Chaoyi Wang, Jinlong Peng, Haoxuan Wang, Yun Cao, Zhenye Gan, Mingmin Chi, et al. Pixelponder: Dynamic patch adaptation for enhanced multi-conditional text-to-image generation. *arXiv preprint arXiv:2503.06684*, 2025.

Surendra Pathak and Bo Han. Towards efficient large vision-language models: A comprehensive survey on inference strategies. arXiv preprint arXiv:2603.27960, 2026. URL https://arxiv.org/abs/2603.27960.

Zhibin Qin, Zhenxiong Tan, Zeqing Wang, Songhua Liu, and Xinchao Wang. Spotedit: Selective region editing in diffusion transformers. In *CVPR*, 2026.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision (ECCV)*, 2024. URL https://arxiv.org/abs/2311.17042.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22857–22867. IEEE, 2025.

Kele Shao, Keda Tao, Kejia Zhang, Sicheng Feng, Mu Cai, Yuzhang Shang, Haoxuan You, Can Qin, Yang Sui, and Huan Wang. A survey of token compression for efficient multimodal large language models. *arXiv preprint arXiv:2507.20198*, 2026a.

Tong Shao, Yusen Fu, Guoying Sun, Jingde Kong, Zhuotao Tian, and Jingyong Su. Soda: Sensitivity-oriented dynamic acceleration for diffusion transformer. In *CVPR*, 2026b.

Yonglak Son, Suhyeok Kim, Seungryong Kim, and Young Geun Kim. Corgi: Contribution-guided block-wise interval caching for training-free acceleration of diffusion transformers. *arXiv preprint arXiv:2512.24195*, 2025.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Caiyi Sun, Yujing Sun, Xiangyu Li, Yuhang Zheng, Yiming Ren, Jiamin Wang, Yuexin Ma, and Siu-Ming Yiu. Unibiotransfer: A unified framework for multiple biometrics transfer. arXiv preprint arXiv:2603.19637, 2026. URL https://arxiv.org/abs/2603.19637.

Yanan Sun, Yanchen Liu, Yinhao Tang, Wenjie Pei, and Kai Chen. Anycontrol: create your artwork with versatile control on text-to-image generation. In *European Conference on Computer Vision*, pp. 92–109. Springer, 2024.

Super Intelligence Team. Firered-image-edit-1.0 technical report. 2026. URL https://arxiv.org/abs/2602.13344.

Zhenxiong Tan, Qiaochu Xue, Xingyi Yang, Songhua Liu, and Xinchao Wang. Ominicontrol2: Efficient conditioning for diffusion transformers. arXiv preprint arXiv:2503.08280, 2025. URL https://arxiv.org/abs/2503.08280.

Haoxuan Wang, Jinlong Peng, Qingdong He, Hao Yang, Ying Jin, Jiafu Wu, Xiaobin Hu, Yanjie Pan, Zhenye Gan, Mingmin Chi, et al. Unicombine: Unified multi-conditional combination with diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18325–18334, 2025a.

Peiyu Wang, Yi Peng, Yimeng Gan, Liang Hu, Tianyidan Xie, Xiaokun Wang, Yichen Wei, Chuanxin Tang, Bo Zhu, Changshi Li, et al. Skywork unipic: Unified autoregressive modeling for visual understanding and generation. *arXiv preprint arXiv:2508.03320*, 2025b.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report. arXiv preprint arXiv:2508.02324, 2025. URL https://arxiv.org/abs/2508.02324.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformer, 2024. URL https://arxiv.org/abs/2410.10629.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802. IEEE, 2025a.

Zixiang Yang, Yue Ma, Yinhan Zhang, Shanhui Mo, Dongrui Liu, and Linfeng Zhang. Evctrl: Efficient control adapter for visual generation. *arXiv preprint arXiv:2508.10963*, 2025b.

Linli Yao, Long Xing, Yang Shi, Sida Li, Yuanxin Liu, Yuhao Dong, Yi-Fan Zhang, Lei Li, Qingxiu Dong, Xiaoyi Dong, et al. Towards efficient multimodal large language models: A survey on token compression. 2026.

Zhihang Yuan, Hanling Zhang, Pu Lu, Xuefei Ning, Linfeng Zhang, Tianchen Zhao, Shengen Yan, Guohao Dai, and Yu Wang. Ditfastattn: Attention compression for diffusion transformer models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 1196–1219. Curran Associates, Inc., 2024. doi: 10.52202/079017-0037. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0267925e3c276e79189251585b4100bf-Paper-Conference.pdf.

Quan-Sheng Zeng, Yunheng Li, Qilong Wang, Peng-Tao Jiang, Zuxuan Wu, Ming-Ming Cheng, and Qibin Hou. A glimpse to compress: Dynamic visual token pruning for large vision-language models. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Evelyn Zhang, Bang Xiao, Jiayi Tang, Qianli Ma, Chang Zou, Xuefei Ning, Xuming Hu, and Linfeng Zhang. Token pruning for caching better: 9 times acceleration on stable diffusion for free. arXiv preprint arXiv:2501.00375, 2024a. URL https://arxiv.org/abs/2501.00375.

Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 20857–20867. IEEE, 2025. doi: 10.1109/ICCV51701.2025.01939. URL https://doi.org/10.1109/ICCV51701.2025.01939.

Wentian Zhang, Haozhe Liu, Jinheng Xie, Francesco Faccio, Mike Zheng Shou, and Jürgen Schmidhuber. Cross-attention makes inference cumbersome in text-to-image diffusion models. *arXiv preprint arXiv:2404.02747v1*, 2024b.

Xuanlei Zhao, Xiaolong Jin, Kai Wang, and Yang You. Real-time video generation with pyramid attention broadcast. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 3296–3319, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/092c2d45005ea2db40fc24c470663416-Paper-Conference.pdf.

Huayu Zheng, Guangzhao Li, Baixuan Zhao, Siqi Luo, Hantao Jiang, Guangtao Zhai, and Xiaohong Liu. A^2-edit: Precise reference-guided image editing of arbitrary objects and ambiguous masks. *arXiv preprint arXiv:2603.10685*, 2026.

Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI Artificial Intelligence Research*, 3:9150038, 2024.

Yifan Zhou, Zeqi Xiao, Tianyi Wei, Shuai Yang, and Xingang Pan. Trainable log-linear sparse attention for efficient diffusion transformers. In *CVPR*, pp. 9424–9433, 2026.

Haowei Zhu, Dehua Tang, Ji Liu, Mingjie Lu, Jintu Zheng, Jinzhang Peng, Dong Li, Yu Wang, Fan Jiang, Lu Tian, et al. Dip-go: A diffusion pruner via few-step gradient optimization. *Advances in Neural Information Processing Systems*, 37:92581–92604, 2024.

Haowei Zhu, Tingxuan Huang, Xing Wang, Tianyu Zhao, Jiexi Wang, Weifeng Chen, Xurui Peng, Fangmin Chen, Junhai Yong, and Bin Wang. Tap: A token-adaptive predictor framework for training-free diffusion acceleration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36082–36091, June 2026a.

Haowei Zhu, Ji Liu, Ziqiong Liu, Dong Li, Jun-Hai Yong, Bin Wang, and Emad Barsoum. Diffsparse: Accelerating diffusion transformers with learned token sparsity. In *International Conference on Learning Representations*, volume 2026, pp. 57438–57456, 2026b.

Chang Zou, Xuyang Liu, Ting Liu, Siteng Huang, and Linfeng Zhang. Accelerating diffusion transformers with token-wise feature caching. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 75467–75488, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/bbe024e0517fe12ac3a8d388b19ff9fe-Paper-Conference.pdf.

The appendix is organized as follows:

- Section A presents additional analysis:
  - Sec. A.1: Theoretical FLOPs analysis.
  - Sec. A.2: Orthogonality with existing acceleration techniques.
  - Sec. A.3: Results under varying numbers of reference images.
- Section B provides implementation details:
  - Sec. B.1: Models and inference.

- Sec. B.2: Reduction decision policies.
- Sec. B.3: Datasets and evaluation.
- Sec. B.4: Details of compared methods.
- Sec. B.5: Effective Attention Ratio (EAR).

• Section C discusses the limitations of our work.

A Additional Analysis

A.1 Theoretical Complexity Analysis

We analyze the theoretical FLOPs of the probing stage and the formal denoising stage, and derive the break-even reduction ratio at which the end-to-end cost matches that of the base model (i.e., theoretical speedup $= 1$). Since the attention and FFN modules dominate the computational complexity, we only analyze their contributions to the computational cost here.

Per-layer cost. Let N be the joint sequence length and d the embedding dimension. We assume the FFN intermediate width is $4d$ (a standard dense-Transformer setting). A common accounting then gives

$$C(N) = \underbrace{4\,N^2 d}_{\text{attention}} + \underbrace{24\,N d^2}_{\text{QKVO + FFN}} . \tag{2}$$

Here we count FLOPs as floating-point operations, including both multiplications and additions: each matrix product of size $(N\times d)\times(d\times N)$ costs $N^2 d$ multiplications and as many additions, giving the attention term $2\,N^2 d$ (for $QK^\top$) $+\,2\,N^2 d$ (for AV) $=\ 4\,N^2 d$; the Q/K/V/O projections each cost $2Nd^2$ ($8Nd^2$ in total), and the FFN's two linear layers ($d \to 4d$ and $4d \to d$) each cost $2Nd{\cdot}4d = 8Nd^2$, i.e., $16Nd^2$ in total. With T denoising steps and L layers, the baseline cost is $TL\,C(N)$.

Sequence lengths. Let N_P, N_G, and N_R denote the numbers of prompt, generative, and reference tokens, so $N = N_P + N_G + N_R$. In the probing stage, images are spatially downsampled by $2\times$ (vision tokens become $1/4$) and the temporal schedule is downsampled by $D = 4$. Thus

$$N' = N_P + \tfrac{1}{4}N_G + \tfrac{1}{4}N_R.$$

Let reduction_ratio $\in [0, 1]$ denote the fraction of reference tokens removed after token reduction. The formal-stage length is

$$N_f = N_P + N_G + (1 - \text{reduction_ratio})\,N_R. \tag{3}$$

The total cost of our method is therefore

$$\frac{T}{4}\,L\,C(N') + T\,L\,C(N_f). \tag{4}$$

Break-even condition. Matching the baseline requires

$$\frac{1}{4}\,C(N') + C(N_f) = C(N). \tag{5}$$

Numerical instantiation. We instantiate a representative two-reference setting with $N_G = N_{R_1} = N_{R_2} = 4096$ (so $N_R = 8192$), $N_P = 100$, and $d = 3072$. Then $N = 12388$, $N' = 3172$, and solving equation 5 yields a break-even ratio of approximately 4.9%. This break-even threshold is quite low that it is almost always satisfied in practice, meaning the savings from token reduction easily outweigh the probing overhead, yielding net end-to-end speedup.

A.2 Orthogonality with Most Existing Acceleration Techniques

Our method is naturally orthogonal and complementary to existing acceleration methods from two perspectives:

- **Stage-level Compatibility:** Most existing acceleration methods operate as *in-model token reduction*, which prunes or merges tokens inside the model during the main denoising loop. In contrast, our method operates outside the main denoising loop and reduces reference tokens before the main denoising stage begins. It directly provides a standard, shortened token sequence to the base model without altering internal model structures or operations, allowing existing in-model methods to be directly applied on top of our reduced sequence.
- **Redundancy-level Orthogonality:** Our method specifically addresses the *spatial token redundancy* in reference images. This is naturally orthogonal to methods focusing on *temporal redundancy* (such as cross-step feature caching and distillation) or methods that prune or sparsify generative tokens (like V-V sparse attention in video generation).

Consequently, our early token reduction can be seamlessly combined with these existing acceleration techniques to achieve compounded efficiency improvement in diffusion-based visual generation tasks.

Clarification on Terminology: In literature, methods compressing tokens before the main network backbone are broadly termed **"pre-model"** token reduction. In this paper, we primarily refer to our method as **"early"** token reduction to avoid ambiguity. Since our framework executes a lightweight downsampled self-probing stage using the DiT model itself to extract importance scores, the DiT model has already been executed before the token reduction takes place. Therefore, strictly calling it "pre-model" might be misleading. We instead adopt the term "early" token reduction to precisely denote that reference tokens are reduced at an early stage (prior to the main denoising stage).

In addition to the default timestep-mapped reduction (Sec 3.3.2), our framework seamlessly supports reducing reference tokens just once for the entire main denoising loop. By averaging the reference utilization scores across all probing steps, tokens can be filtered once before the main loop begins. Note that both modes share the same fundamental principle: reference tokens are always pruned or substituted outside the DiT in the main denoising, requiring zero structural or operational changes to the DiT in the main loop. We default (all experiments in the paper) to timestep-mapped reduction simply because matching the reference retention to the evolving demands across different denoising stages offers a slight performance gain.

A.3 Results under varying numbers of reference images

Figure 8 reports the score and latency of Ours (slow) on MACRO with 2–10 reference images.

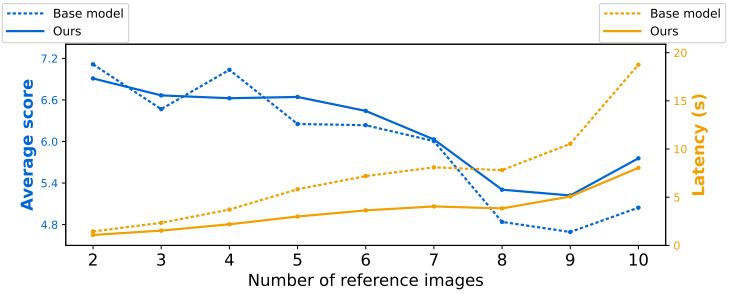


Figure 8: Average score and latency of Ours (slow) (Flux.2-Klein) on MACRO with 2–10 reference images.

B Implementation Details

Our **code** will be made publicly available upon acceptance, containing all implementation details.

B.1 Models and Inference

We evaluate on three base models: the latent-space Diffusion Transformers (DiTs) Flux.2-Klein-4B and Qwen-Image-Edit-2511, and HiDream-O1-Image, a Pixel-level Unified Transformer (UiT) that directly processes raw pixels without a VAE. Evaluating on these two architectures confirms our method is effective for diverse generic reference-based generation models. For all base models, we use their official code and default configurations. Specifically, we use 4, 40, and 50 inference steps for Flux.2-Klein-4B (Black Forest Labs, 2025), Qwen-Image-Edit-2511 (Wu et al., 2025), and HiDream-O1-Image (Cai et al., 2026), respectively. $\texttt{true_cfg_scale} = 4.0$ for Qwen-Image-Edit-2511; $\texttt{guidance_scale} = 1.0$ for Flux.2-Klein-4B; $\texttt{guidance_scale} = 5.0$ for HiDream-O1-Image. FlashAttention (Dao et al., 2022) is used as the attention backend. Since the attention matrix is not materialized under FlashAttention, we use a specialized Triton kernel to extract the reference utilization score. VAE slicing and tiling are enabled when inferring on samples that encounter out of memory during the VAE encoding stage. For each model, we report a quality-oriented **Ours(slow)** and a speed-oriented **Ours(fast)** by adjusting τ_1 and τ_2. All experiments run on a single NVIDIA H800 (80GB) GPU.

For MultiBanana (Oshima et al., 2026), which does not prescribe the input reference or output target size, we use each reference at its native resolution, except that any image whose area exceeds $\mathrm{MAX_AREA} = 4096 \times 4096$ is resized to this area budget; the last reference's size is used as the target generation size. For Macro (Chen et al., 2026), we follow its official settings. All runs use seed 42.

Handling of Rotary Positional Embeddings (RoPE) Since the image generation models (Black Forest Labs, 2025; Wu et al., 2025; Cai et al., 2026) encode spatial layouts via RoPE (Su et al., 2024), we maintain geometric consistency for the pruned reference sequence. The preserved full-resolution tokens retain their original 2D RoPE coordinates. For the replaced low-resolution tokens representing mid-importance blocks, we assign their RoPE coordinates precisely to the spatial center of the corresponding 2×2 block.

B.2 Reduction Decision Policies

B.2.1 Ratio-based Decision (Default).

In our experiments, we adopt a ratio-based policy. Under this setting, retention proportions p_1 and p_2 are explicitly designated for the Full-Res Tier and Low-Res Tier, respectively. The cutoffs τ_1 and τ_2 in Eq. (1) are dynamically determined on a per-sample basis: after sorting all downsampled reference token scores $\{\tilde{I}^{\tilde{t}}_{\tilde{r}}\}$ at probe timestep $\tilde{t}$ of the current sample in descending order, τ_1 is set to the score at the top-p_1 percentile rank, and τ_2 is set to the score at the top-$(p_1 + p_2)$ percentile rank. Consequently, the top-p_1 fraction enters the Full-Res Tier, the subsequent p_2 fraction is substituted into the Low-Res Tier, and the remaining tokens are discarded. The primary advantage of this policy is budget predictability, providing deterministic sequence lengths.

In Table 1, we configure two operational modes for this setting: **Ours(fast)** and **Ours(slow)**. For **Ours(fast)**, we set the Full-Res Tier ratio to $p_1 = 15\%$ and the Low-Res Tier ratio to $p_2 = 2 \times p_1 = 30\%$. For **Ours(slow)**, both proportions are doubled, setting $p_1 = 30\%$ and $p_2 = 60\%$. These proportions represent global budgets evaluated across all reference images for a given sample. To prevent severe inter-reference imbalance—such as a visually dominant reference image exhausting all available quotas and starving other references—we enforce a minimum retention guarantee on each individual reference image. Specifically, every reference image is constrained to retain at least $\frac{2}{3}$ of the global Full-Res ratio within its own token grid, yielding a per-reference lower bound of $p_{\min} = \frac{2}{3}p_1$ (i.e., at least 10% for Ours(fast) and 20% for Ours(slow)).

Timestep-Aware Ratio Shift. Because DiTs prioritize global layout planning in early steps and fine-grained detail refinement in late steps, high-resolution details are less critical early on, whereas a wider spatial context is more beneficial. We therefore dynamically modulate the retention proportions across denoising timesteps: during the earliest steps, the effective Full-Res ratio is scaled to $0.8p_1$, while the Low-Res ratio is expanded to $1.2p_2$. As generation proceeds toward the final refinement steps, these scaling factors linearly transition back to 1.0 (restoring the base ratios p_1

and p_2). This dynamic adjustment slightly improves generation quality, so we adopt it in our default configuration.

B.2.2 THRESHOLD-BASED DECISION.

We also support a threshold-based decision policy for sample-difficulty awareness, where fixed absolute thresholds τ_1 and τ_2 are applied directly to the normalized reference utilization score $\tilde{I}_{\text{norm}}$. Before thresholding, we normalize the raw utilization score by the total number of downsampled reference tokens of the current sample, $\tilde{I}_{\text{norm}} = \tilde{I} \cdot N_{\tilde{R}}$, making the score agnostic to varying reference counts and resolutions across samples. The primary benefit of this strategy is sample-difficulty awareness: for complex reference scenes requiring rich context, the model naturally allocates high attention over broader regions, automatically preserving a larger token budget; simpler references with clean backgrounds are pruned more aggressively.

B.3 DATASETS AND EVALUATION

MultiBanana (Oshima et al., 2026) is a challenging benchmark for multi-reference image generation, consisting of 3K+ samples with up to 8 reference images. For the two-reference setting, it covers 11 editing tasks including subject addition, replacement, background change, color/material/-pose/hairstyle/makeup modification, tone transformation, style transfer, and text correction. For the ≥ 3 references setting, it includes four tasks: object composition, background composition, local attribute transfer, and global tone edit. The benchmark incorporates four difficulty axes: cross-domain diversity (e.g., mixing photos, anime, and sketches), scale and viewpoint differences, rare concept references (e.g., a red banana), and multilingual references (English, Chinese, and Japanese), comprehensively evaluating the model's ability on multi-reference generation tasks. MultiBanana uses a VLM to score generation quality from five dimensions: Text-Instruction Alignment, Reference Consistency, Background-Subject Match, Physical Realism, and Visual Quality, each on a 10-point scale. To align with our focus on multi-reference efficiency setting, our main experiment targets all multi-reference tasks (excluding the pure instruction-based editing task in MultiBanana). We report the mean across tasks for each quality dimension.

To reduce decoding randomness, each sample is scored twice under different decoding configurations and the two scores are averaged: (1) greedy decoding (`temperature` $= 0$ and `top_k` $= 1$); (2) Top-p sampling with `temperature` $= 0.8$, `top_p` $= 0.8$, and `random_seed` $= 0$.

Following the recommendation in Issue #2 of the official MultiBanana GitHub repository, we exclude the local attribute transfer samples with IDs ≤ 200, since the maintainers confirmed a prompt-reference mismatch bug on these samples and advised excluding them from evaluation.

Macro (Chen et al., 2026) is also a reference-based image generation dataset covering four tasks: customization (composing multiple subjects into coherent scenes), illustration (generating narrative images based on interleaved multimodal context), spatial synthesis (novel view synthesis), and temporal synthesis (predicting a future keyframe from history keyframes), each comprising 1000 samples with 1–10 reference images, in total 4000 samples. Macro evaluates each sample across multiple dimensions to compute an overall score. We report the average overall score across all samples.

Qwen3-VL (Bai et al., 2025), deployed via the LMDeploy (Contributors, 2023) framework on an NVIDIA H800 (80GB), is used as the VLM-as-judge for the two benchmarks.

For efficiency, we report end-to-end latency (including the VAE and probing denoising stage, excluding disk I/O such as reference image reading and output image dumping), peak activation VRAM, and speedup over the corresponding base model.

B.4 DETAILS OF COMPARED METHODS

ToMeSD. For ToMeSD, we directly apply the official method to the reference image tokens.

VisPruner. VisPruner (Zhang et al., 2025) first extracts an importance score for each visual token from a vision encoder's self-attention. If the encoder has a CLS token, importance is the CLS-to-patch attention; if not (e.g., SigLIP), it averages attention over the query axis. It then keeps

a budget of tokens by selecting the top-scoring important subset and filling the rest via iterative cosine-similarity deduplication for diversity. For HiDream-O1-Image, we reuse the model's built-in Qwen3-VL vision encoder to score reference tokens. For FLUX.2 Klein, the model has no vision encoder, so we follow VisPruner's default and use CLIP (ViT-L/14-336) for importance. Because CLIP patches do not match Flux's VAE-packed reference tokens, we compute CLS-to-patch attention at CLIP's native resolution and bilinearly map the score map onto the Flux reference grid, yielding a spatially aligned importance map for subsequent stages. Important and diverse token selection then proceeds as in VisPruner, with cosine deduplication on the packed VAE latents that enter the DiT. For Qwen-Image-Edit, we reuse the built-in Qwen2.5-VL vision encoder to score tokens. Importance is computed on the pipeline's 384×384 condition views and bilinearly mapped onto the VAE-packed reference grid; VisPruner then selects important and diverse tokens with cosine deduplication on those packed VAE latents.

B.5 Effective Attention Ratio (EAR)

The Effective Attention Ratio (EAR) in Sec. 4.5.2 is only possible when the ground-truth important region is known. While general reference-based tasks do not mark it, the object composition task in MultiBanana (Oshima et al., 2026) does Since the object composition task aims to composite the salient object in each reference image, the ground-truth important region is the salient object region in each reference image. We obtain this salient region with BiRefNet (Zheng et al., 2024), which returns a soft saliency map $s \in [0, 1]$. The map is resized to the latent grid of the reference, and (w, h) indexes a reference token on that grid.

With the region fixed, we measure the attention that generative tokens G and prompt tokens P place on each reference token inside the main denoising loop. At timestep t and layer l, their total attention to the token at (w, h) is $\sum_{i \in G \cup P} A^{t,l}_{i,(w,h)}$, where $A^{t,l}_{i,(w,h)}$ is the head-averaged attention from query i. Summing over timesteps and layers, the Effective Attention Ratio (EAR) is the share of this attention that lands on the salient region,

$$\text{Effective Attention Ratio (EAR)} = \frac{\sum_{w,h} \left(\sum_t \sum_l \sum_{i \in G \cup P} A^{t,l}_{i,(w,h)} \right) s_{(w,h)}}{\sum_{w,h} \sum_t \sum_l \sum_{i \in G \cup P} A^{t,l}_{i,(w,h)}}. \tag{6}$$

Table 4 compares the unpruned base model and three ways of building the reference utilization score that drives token reduction. The reduction budget is the same in all four: the same ratio-based rule fixes how many tokens are preserved and how many are substituted. (1) ***Baseline (No Reduction)*** keeps the full reference sequence, so the main denoising sees every reference token. (2) ***Top-1 layer*** reduces tokens with the score of the single layer best aligned with the oracle importance I^*_r. On Flux.2-Klein-4B (Black Forest Labs, 2025) this is layer l=15, shown in Fig. 3(b), with $I^l_r = \frac{1}{T} \sum_t \sum_{i \in G \cup P} A^{t,l}_{i,r}$. (3) ***Top 20% of layers*** uses the mean of the top 20% best-correlated middle layers. On Flux.2-Klein-4B these are $\mathcal{L} = \{14, 15, 16, 17, 18\}$, and $I_r = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} I^l_r$. (4) ***All layers*** uses the mean over every transformer block, $I_r = \frac{1}{L} \sum_{l=1}^{L} I^l_r = \frac{1}{LT} \sum_{l=1}^{L} \sum_t \sum_{i \in G \cup P} A^{t,l}_{i,r}$.

C Limitations

This paper focuses on the efficiency drop caused by token growth when multiple reference images are used. In the single-image setting, namely pure instruction-based editing, the acceleration gain is relatively lower than in the multi-reference case.