

Region-Aware Test-Time Scaling for Compositional Image Generation

Mingzhu Shen¹, Peng Ye², Xinyin Ma³, Gongfan Fang³,
Christos-Savvas Bouganis¹, Yiren Zhao¹, and Xinchao Wang^{3*}

¹ Imperial College London, the United Kingdom

² The Chinese University of Hong Kong, China

³ National University of Singapore, Singapore

m.shen23@imperial.ac.uk, xinchao@nus.edu.sg

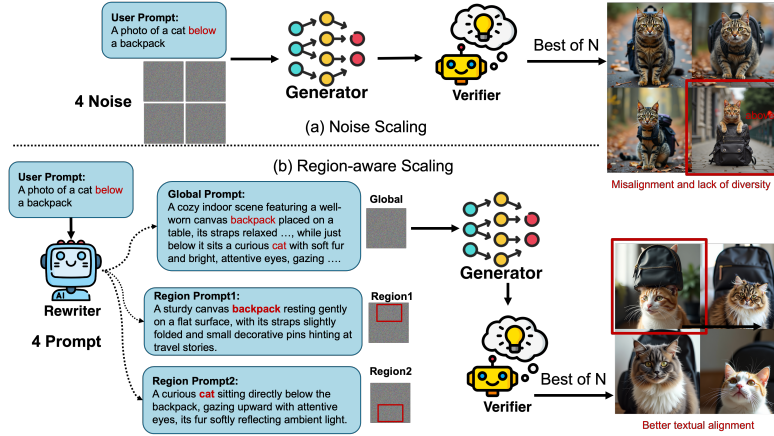


Fig. 1: Conceptual comparison between global noise scaling and our proposed region-aware scaling. (a) **Noise Scaling** performs test-time scaling along the noise axis by sampling N random seeds. However, in complex compositional scenarios, valid structural solutions are often sparse, requiring a large compute budget (e.g., $N = 32$) to achieve alignment. (b) Our proposed **Region-Aware Scaling (RAS)** introduces the **Region Axis** as a more efficient dimension for scaling. By decomposing the task into coordinated regional sub-tasks with localized guidance, RAS significantly densifies the distribution of valid candidates, achieving superior alignment with a much smaller sample budget (e.g., $N = 2$).

Abstract. Test-time scaling (TTS) has emerged as a promising paradigm for improving the performance of large-scale models. However, existing vision TTS methods largely rely on *global* exploration—e.g., resampling noise or rewriting prompts—and often struggle to efficiently search the vast compositional space. As a result, they can exhibit a “scaling plateau,” where additional computation

* Correspondence to: Xinchao Wang <xinchao@nus.edu.sg>.

yields diminishing returns in semantic alignment. In this paper, we propose **Region-Aware Scaling (RAS)**, a framework that bridges region-aware generation and test-time scaling. By treating *regional decomposition* as a powerful and previously overlooked **scaling axis**, RAS converts complex compositional prompts into coordinated regional sub-tasks, effectively increasing the density of valid candidates during inference. At its core, RAS builds on a **training-free Region-Aware Generation (RAG)**. Unlike many layout-based methods that incur substantial overhead, RAG injects regional guidance *only during early denoising*, enabling precise attribute binding while preserving global structural coherence. We evaluate RAS on the **GenEval** benchmark and observe consistent improvements in scaling efficiency across diverse compositional challenges. RAS achieves an overall score of **0.85 with only 2 samples**, matching a **32-sample** noise-scaling baseline; with 4 samples, it reaches **0.88**, surpassing the combination of 32-sample noise and prompt scaling. Overall, our results suggest that **structuring the search space via regional decomposition** provides a principled and computationally efficient direction for scaling compositional alignment.

1 Introduction

Recent advances in text-to-image diffusion models [3, 7, 9, 16, 19, 27, 29–31, 42–44, 49] and flow-matching models [20, 23] have substantially improved generative quality, enabling efficient ODE-based sampling over traditional SDE formulations [12, 35]. These improvements have made high-quality synthesis increasingly accessible under practical inference budgets. In parallel, test-time scaling (TTS) [15, 26, 34, 36, 37, 46, 47] has proven highly effective in large language models by leveraging additional inference-time compute, suggesting a compelling paradigm: when training is fixed, allocate more compute at inference time to increase the probability of producing a correct or well-aligned output. Yet, TTS for vision still falls short on *compositional* prompts that require reliably combining multiple objects and constraints—including object presence, attributes (e.g., color/size), counting, and inter-object relations—within a single coherent image. In practice, these failures are not merely aesthetic; they often reflect violations of explicit user intent (e.g., missing objects, incorrect counts, or wrong relations), which are difficult to remediate without additional sampling. Meanwhile, post-training alignment has emerged as another complementary direction, including reinforcement-learning-based fine-tuning for generative models (e.g., Flow-GRPO) [22]. However, such approaches typically incur substantial training cost, motivating *training-free* alternatives that improve alignment purely at inference time.

Most existing vision TTS methods follow a Best-of- N paradigm, using *global* exploration such as noise scaling [1, 24, 28, 33, 48] or prompt rewriting [41, 50] to search the sampling space. While these global strategies improve average fidelity, they can become sample-inefficient when success requires satisfying *multiple coupled constraints* simultaneously: the valid set of samples is sparse, so increasing N yields diminishing returns under limited budgets. This issue is especially severe for relation-heavy or multi-object cases, where correct outputs may require the model to get both *what* to draw and *where* to place it right early in the generation process. We attribute the inefficiency of purely global scaling to a mismatch between *global conditioning* and *local*

constraint satisfaction. Noise scaling increases stochastic exploration without changing semantics; global prompt rewriting can refine descriptions but still provides weak handles for enforcing distinct object-level constraints; and reflection-style scaling [18, 50] can incorporate corrective feedback but typically requires additional training.

However, recent progress on *training-free* region-aware generation [2, 5, 45] offers a complementary perspective: by explicitly decomposing a compositional prompt into object-centric subregions, the model can be guided with localized constraints during generation (e.g., via region prompts and region layouts; see Fig. 1). Crucially, introducing region decomposition changes the picture: it expands the set of controllable test-time factors beyond global noise/prompt variations. Consequently, when early denoising commits to an incomplete or incorrect allocation of objects and attributes (e.g., wrong spatial arrangement or missing instances), global exploration often needs larger budgets, making global TTS inefficient in the low-budget regime. Motivated by this observation, we propose **Region-Aware Test-Time Scaling** (RAS), a simple and training-free extension of TTS that treats *regions* as a new scaling axis. RAS first performs chain-of-thought scene decomposition to identify key objects and assign spatial regions, then generates a descriptive global prompt and region-level sub-prompts, and finally injects these localized conditions *only during the early denoising steps*—when coarse object allocation and layout are formed—before reverting to standard global refinement. By turning a hard joint compositional requirement into easier region-level sub-tasks while preserving global coherence, this early-only design improves multi-object compositional controllability and avoids late-stage boundary artifacts, enabling region guidance to be composed with Best-of- N scaling under practical inference budgets.

Our contributions are summarized as follows:

- We identify a key limitation of existing global test-time scaling for compositional text-to-image generation (e.g., noise re-sampling or prompt rewriting): it struggles to efficiently explore the combinatorial compositional space, and we show that introducing region-level information provides a more effective search structure for improving semantic alignment.
- We propose Region-Aware Scaling (RAS), which enables a practical training-free Region-Aware Generation (RAG) via an efficient early-only regional injection mechanism, and we systematically explore regions as a new scaling axis at inference time, including multiple design choices for how region-aware scaling is performed.
- We demonstrate consistent performance gains and improved sample efficiency on compositional benchmarks under budget-matched settings, achieving stronger compositional accuracy with significantly fewer samples compared to global scaling baselines.

2 Method

2.1 Preliminary

Denoising Process. Given an initial noise latent $\mathbf{z}_0 \sim \mathcal{N}(0, I)$, a text prompt p , and a diffusion model f_θ , image generation proceeds via a denoising trajectory of T steps:

$$I = \mathcal{G}(f_\theta, \mathbf{z}_0, p, T), \quad (1)$$

where $\mathcal{G}$ denotes the denoising sampling scheduler. At each timestep $t \in \{1, \dots, T\}$, the model refines $\mathbf{z}_{t-1}$ according to the denoising prediction conditioned on the text embedding $\phi(p)$ generated by text encoder $\phi(\cdot)$:

$$\mathbf{z}_t = f_\theta(\mathbf{z}_{t-1}, \phi(p), t). \quad (2)$$

After the final step, the latent $\mathbf{z}_T$ is decoded into an image I using the image decoder $\mathcal{D}$:

$$I = \mathcal{D}(\mathbf{z}_T). \quad (3)$$

Best-of- N Test-Time Scaling. Test-time scaling typically follows a Best-of- N protocol. Given a prompt p , we sample N independent candidates by varying the initial noise seeds $\{\mathbf{z}_0^{(k)}\}_{k=1}^N$ (as in noise scaling) and feeding them into the denoising scheduler $\mathcal{G}$:

$$I_k = \mathcal{G}(f_\theta, \mathbf{z}_0^{(k)}, p, T), \quad k = 1, \dots, N. \quad (4)$$

A verifier then selects the final output from the candidate set. In our experiments, we use the same verifier $V(\cdot)$ [44] to score each candidate and choose the best one:

$$I^* = \arg \max_{k \in \{1, \dots, N\}} V(I_k, p). \quad (5)$$

We evaluate the selected sample I^* using standard benchmark protocols (e.g., GenEval [10]), so the reported scores reflect the performance of the entire Best-of- N pipeline.

2.2 Why Existing Training-free TTS Fails

Existing Problems. Most training-free test-time scaling (TTS) methods rely on *global* exploration, such as sampling multiple noise seeds (noise scaling) and generating prompt variants via rewriting (prompt scaling). Figure 2 shows that these global strategies are sample-inefficient for compositional prompts. While Single/Two saturate quickly (near 100% / 95–100%), Position improves from $< 20\%$ at $N=2$ to only $\sim 55\%$ at $N=32$. Prompt rewriting helps but still plateaus below 70%, and combining prompt+noise scaling yields only marginal additional gains. Moreover, the per-prompt success distribution exhibits a pronounced long tail: among 100 Position prompts, over 30% have less than 10% success rate even under 32-sample scaling (Figure 2(c)). These results suggest that many compositional configurations correspond to low-probability modes, where early coarse-layout mistakes are unlikely to be corrected by global resampling within practical budgets.

Task Decomposition. Inspired by recent progress in region-aware generation [2, 5, 45], we adopt a *task-decomposition* view for compositional prompts. A single global text embedding provides only coarse, whole-image control and cannot explicitly express the localized, object-centric constraints required by many compositional requirements; consequently, relation-constrained cases can be underrepresented in the model distribution and difficult to satisfy under global conditioning alone (Figure 2). In contrast, generating a single object within a specified spatial region is considerably easier and

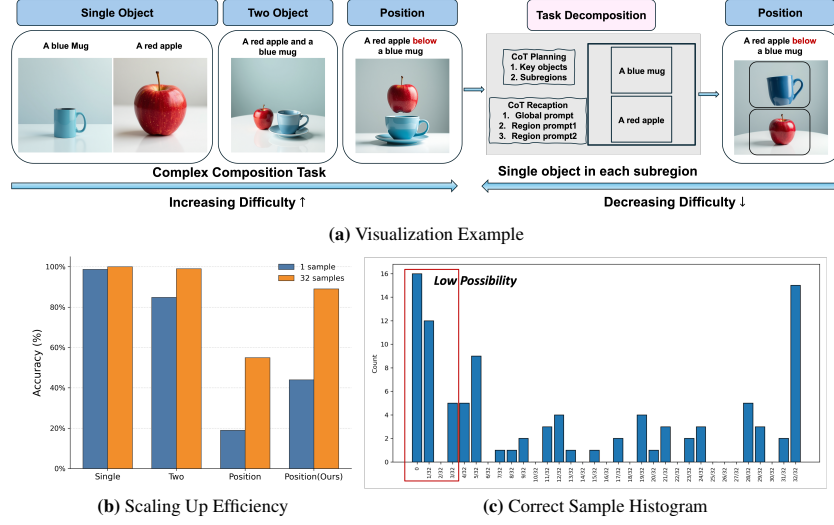


Fig. 2: (a) Visual examples showing an easy-to-hard progression: single-object attribute generation and two-object co-occurrence are relatively easy, whereas relation-constrained composition (**Position**, with the keyword *below* highlighted) is substantially harder. We decompose the compositional prompt into subregions via CoT planning/recaption, reducing the task to *single-object generation within each subregion* and thus lowering the effective difficulty. (b) Accuracy under Best-of- N scaling for $N=1$ and $N=32$. Single/Two saturate near 100% at $N=32$, while **Position** benefits much less from global resampling. (c) For $N=32$, we compute the per-prompt probability of obtaining correct sample. A large fraction of prompts have a success probability below 10%.

more stable, motivating us to decompose a complex prompt into simpler region-level sub-tasks that the diffusion model can satisfy more reliably.

From this perspective, chain-of-thought (CoT) planning provides an effective mechanism to (i) identify key objects, (ii) allocate spatial regions, and (iii) assign each region a localized sub-prompt capturing object identity, appearance, and relations. Such structured decomposition offers two advantages: it enables explicit verification and control of object-centric constraints, and it reduces ambiguity by narrowing each sub-task to a smaller semantic and spatial scope. Compared to random noise variations or global prompt perturbations, region-level reasoning provides a more direct and reliable way to guide the model toward correct compositional structures. **Building on this insight, we further explore how to incorporate region-aware generation into the test-time scaling framework, enabling more efficient scaling up under practical compute budgets.**

To this end, we design a region-aware generation procedure that modifies the sampling process to inject localized guidance only when it is most useful, while avoiding unnecessary complexity for simple layouts and preserving global coherence. We further treat *regions* and *region prompts* as additional scaling dimensions and study how to allocate inference-time compute across them effectively. Finally, we discuss both par-

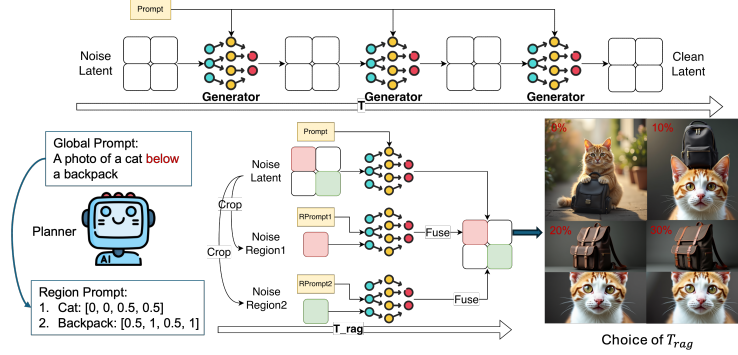


Fig. 3: Overview of the proposed Region-Aware Test-Time Scaling (RAS) framework. (a) Prompt Planning and Recaption. Given a user prompt (e.g., “a cat below a backpack”), the Rewriter performs chain-of-thought (CoT) reasoning to identify key objects and subregions, generating both a global prompt and region-specific prompts that capture object-centric constraints (identity, attributes, and relations). (b) Region-aware Generation. During diffusion sampling, regional latents are cropped from the global latent and guided by their respective region prompts for the first T_{rag} denoising steps before fusing back into the global latent. We visualize results under different T_{rag} settings, showing that early regional injection (e.g., 10%) achieves the best balance between compositional correctness and global coherence.

allel and sequential scaling paradigms and explore principled strategies for using these scaling dimensions to improve compositional success under practical compute budgets.

2.3 Region-aware Generation

Definition of Region-Aware Generation. Given a text prompt p , our goal is to generate an image I that aligns with p more faithfully than conventional global TTS. We first parse p into R semantic regions via CoT planning or a lightweight rule-based parser (e.g., “a cat below a backpack” $\rightarrow$ {region₁ : cat, region₂ : backpack}). Each region r_i is associated with a bounding-box mask $M_i \in [0, 1]^{H \times W}$ and a refined region-level prompt p'_i .

Unlike prior region-aware generation methods such as RPG [45], we do not apply regional conditioning throughout the entire sampling process, nor do we rely on complex mechanisms such as hard/soft binding used in [5]. Such designs often require intricate region definitions and planning, which can be difficult for LLM-based planners to generate reliably and may require multiple retries. Instead, we divide the sampling process into two stages. During the first T_{rag} denoising steps, we perform *region-aware generation* by injecting regional information as illustrated in Fig. 3. For the remaining steps, we revert to standard global denoising for refinement. We represent the generation state using a shared global latent $\mathbf{z}^{(g)}$ that captures holistic structure, together with local latents $\{\mathbf{z}^{(i)}\}_{i=1}^R$ that evolve under region prompts p'_i within their corresponding masks M_i . This formulation enables localized semantic updates while maintaining global coherence through the shared latent.

Regional Updates. For each semantic region r_i with refined text prompt p'_i and mask $M_i \in [0, 1]^{H \times W}$, we extract a *local latent* $\mathbf{z}_t^{(i)}$ from the global latent $\mathbf{z}_t^{(g)}$ at initialization for all R regions:

$$\mathbf{z}_0^{(i)} = M_i \odot \mathbf{z}_0^{(g)}, \quad i = 1, \dots, R \quad (6)$$

where $\odot$ denotes element-wise spatial cropping. Each local latent then evolves under its region-specific prompt in parallel with the global diffusion trajectory during the early denoising stage ($t < T_{\text{rag}}$):

$$\mathbf{z}_t^{(i)} = f_\theta(\mathbf{z}_{t-1}^{(i)}, \phi(p'_i), t), \quad i = 1, \dots, R, \quad t < T_{\text{rag}}. \quad (7)$$

$$\mathbf{z}_t^{(g)} = f_\theta(\mathbf{z}_{t-1}^{(g)}, \phi(p), t), \quad t = 1, \dots, T. \quad (8)$$

Weighted fusion. To integrate regional and global information, we normalize each mask as

$$\tilde{M}_i = \frac{M_i}{\sum_j M_j + \varepsilon}, \quad (9)$$

where $\varepsilon > 0$ ensures numerical stability. We then blend the global and local latents using a weighted ratio α_t :

$$\mathbf{z}_t = (1 - \alpha_t) \mathbf{z}_t^{(g)} + \alpha_t \sum_{i=1}^R \tilde{M}_i \odot \mathbf{z}_t^{(i)}, \quad t < T_{\text{rag}}. \quad (10)$$

After region-aware generation steps, the process continues with a single global trajectory:

$$\mathbf{z}_t = f_\theta(\mathbf{z}_{t-1}, \phi(p), t), \quad t \geq T_{\text{rag}}. \quad (11)$$

Choice of RAG steps. The choice of T_{rag} involves a critical trade-off between semantic accuracy and computational efficiency. As illustrated in Fig. 3, injecting regional guidance solely during the early denoising stage (e.g., the first 10% of steps) is sufficient to establish a correct coarse layout and prevent object leakage. Prolonged injection (e.g., beyond 30%) offers marginal alignment gains but can introduce noticeable boundary artifacts. In terms of computational cost, full-step regional conditioning nearly doubles the inference time. Conversely, our early-stage injection strategy incurs only a marginal $\sim 10\%$ overhead. Furthermore, our design ensures that this overhead remains nearly constant regardless of the number of regions. This presents a significant scalability advantage over existing methods like RPG [45] and other region-based approaches [5], whose costs scale linearly with the number of regions. Consequently, early-stage regional conditioning provides an optimal balance between compositional fidelity, visual coherence, and inference-time scalability.

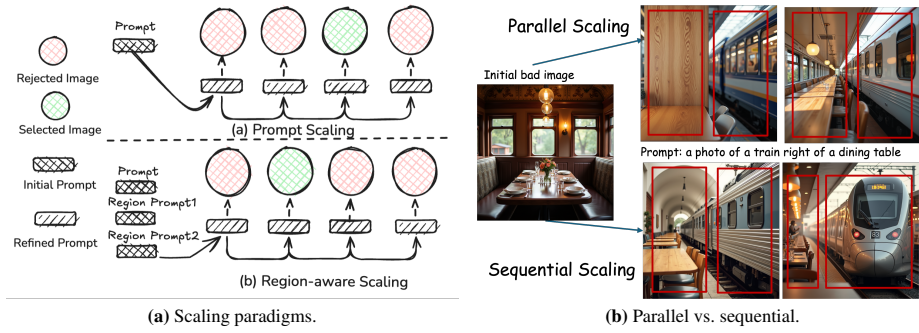


Fig. 4: Scaling paradigms and scheduling. **Left:** Best-of- N prompt scaling vs. region-aware scaling (green/red: selected/rejected; hatched boxes: prompts). **Right:** Sequential scaling updates region layouts using information from earlier generations, enabling corrections that parallel scaling (fixed boxes) cannot achieve.

2.4 From RAG to RAS

Scaling Dimensions Enabled by Region Decomposition. Classical Best-of- N scaling for text-to-image generation typically resamples the *stochasticity*—i.e., the noise seed—under a fixed global prompt, or performs prompt scaling by rewriting the prompt multiple times (Fig. 4a). Once we introduce region decomposition, the *resampling space* expands substantially: beyond varying $\mathbf{z}_0$, we may also resample (i) the global prompt recaption $\tilde{p}$, (ii) region-level prompts $\{p'_i\}_{i=1}^R$, and (iii) region layouts $\{M_i\}_{i=1}^R$ (boxes/masks) as shown in Fig. 4a. This enlarged configuration space is a key distinction between region-aware and prompt-only scaling, enabling richer test-time search over controllable factors. Notably, classical noise or prompt scaling arise as special cases of our pipeline by simply holding the region decomposition fixed (or omitting it) while varying only $\mathbf{z}_0$ and/or $\tilde{p}$.

RAS: Best-of- N over RAG configurations. We define **Region-Aware Test-Time Scaling (RAS)** as multiple image generation built on top of the single **RAG** sample. Instead of viewing a candidate as a function of noise only, we bundle all factors into a *configuration*

$$\xi = \left(\mathbf{z}_0, \tilde{p}, \{(M_i, p'_i)\}_{i=1}^R, T_{\text{rag}} \right), \quad (12)$$

and write a single RAG run as

$$I = \text{RAG}(f_\theta; \xi) \equiv \text{RAG}(f_\theta, \mathbf{z}_0, \tilde{p}, T; \{(M_i, p'_i)\}_{i=1}^R, T_{\text{rag}}). \quad (13)$$

Given a sampling budget N , RAS draws N independent configurations $\{\xi_k\}_{k=1}^N$ from a user-specified distribution $q(\xi)$ (e.g., noise-only, prompt-only, layout-only, or their combinations), generates candidates, and selects the best by any scoring rule $s(\cdot)$:

$$I_k = \text{RAG}(f_\theta; \xi_k), \quad \xi_k \sim q(\xi), \quad k = 1, \dots, N, \quad (14)$$

$$I^* = \arg \max_{k \in \{1, \dots, N\}} s(I_k). \quad (15)$$

This formulation makes the relation explicit: **RAG** specifies *how to generate one candidate* given a region plan, while **RAS** specifies *how to allocate a test-time budget* over a (potentially large) space of region-aware configurations and select the best.

Parallel vs. Sequential scaling. With an expanded configuration space, test-time scaling can be organized in two ways. *Parallel* scaling samples ξ_k independently from $q(\xi)$ and runs all candidates in parallel; this is simple, efficient, and compatible with standard Best-of- N pipelines. *Sequential* scaling performs iterative refinement: after producing an image I_t , the system analyzes it, updates the configuration (e.g., adjusts $\{M_i\}$ or rewrites $\{p'_i\}$ and/or $\tilde{p}$), and resamples to obtain I_{t+1} . Recent reflection-based methods [18, 50] instantiate this paradigm by using a VLM to critique the current output and propose corrections for the next round, often coupled with a reflection-aware generation pipeline. In contrast, our setting remains fully training-free: the feedback loop (when used) operates only on the *conditioning variables*—region layouts $\{M_i\}$, region prompts $\{p'_i\}$, and/or the global prompt $\tilde{p}$ —without modifying the underlying generator.

As illustrated in Figure 4b, sequential scaling can correct fine-grained compositional errors by using information revealed only after generation. However, each refinement round typically requires an additional VLM inference (and an extra sampling pass), which can incur substantial computational overhead. In this work, we find that **parallel** region-aware scaling already delivers strong compositional gains under practical budgets, and thus adopt it as the default in all experiments.

Pipeline Overview. RAS comprises (i) a **region specification** module that produces a region decomposition and text guidance, and (ii) an **efficient single-sample RAG** primitive that generates one candidate conditioned on that specification. Concretely, given an input prompt p , we use CoT planning and optional recaption to obtain layouts $\{M_i\}_{i=1}^R$ (boxes/masks) and region prompts $\{p'_i\}_{i=1}^R$, and optionally an enriched global prompt $\tilde{p}$. RAS then runs the RAG procedure inside a Best-of- N outer-loop, where each candidate corresponds to a configuration $\xi = (\mathbf{z}_0, \tilde{p}, \{(M_i, p'_i)\}_{i=1}^R, T_{\text{rag}})$ drawn from a user-specified distribution $q(\xi)$. By choosing $q(\xi)$, the outer-loop can instantiate different scaling dimensions—e.g., resampling only $\mathbf{z}_0$ (noise scaling), only $\tilde{p}$ (prompt scaling), or region-level components such as $\{M_i\}$ and/or $\{p'_i\}$ (region-aware scaling), as well as their combinations.

3 Experiments

Evaluation Metrics. We report **GenEval** scores [10] following the standard protocol, including the overall score and category-wise scores (e.g., multi-object, relations, and attribute binding) when applicable. To further assess compositional generation, we evaluate on **T2I-CompBench** [13], which covers attribute binding (color, shape, texture), object relationships (spatial, non-spatial), and complex compositions. Following recent practice, we report per-category scores using GPT-4o [14] as a normalized judge (higher is better), providing a finer-grained measure of attribute and relational fidelity than object-centric detection alone. Finally, to capture overall image quality and human preference, we report **HPSv3** [25], a learned human-preference reward model that

Table 1: Quantitative comparisons of our framework against standard text-to-image models and various inference-time scaling approaches on the GenEval [10] benchmark. Our proposed method consistently outperforms all baselines across the 1, 2, 4, 8 and 32-sample settings, demonstrating superior efficiency and alignment under limited sampling budgets.

Methods	Samples	Overall	Single	Two	Counting	Colors	Position	Attribution
<i>Text-to-Image Models</i>								
SANA-1.5 4.8B [44]	1	0.72	0.99	0.85	0.77	0.87	0.34	0.54
Lumina-Image 2.0 [38]	1	0.73	0.99	0.87	0.67	0.88	0.34	0.62
SD3 [7]	1	0.74	0.99	0.94	0.72	0.89	0.33	0.60
Playground v3 [21]	1	0.76	0.99	0.95	0.72	0.82	0.50	0.54
Janus-Pro-7B [4]	1	0.80	0.99	0.89	0.59	0.90	0.79	0.66
<i>Inference Time Scaling</i>								
SANA-1.0-1.6B [43]	1	0.66	0.99	0.77	0.62	0.88	0.21	0.47
+ Noise Scaling [24]	20	0.80	1.00	0.93	0.79	0.91	0.55	0.62
+ Reflect-DiT [18]	20	0.81	0.98	0.96	0.80	0.88	0.66	0.60
+ RAS	8	0.81	0.99	0.93	0.80	0.88	0.60	0.65
FLUX.1-dev [16]	1	0.67	0.99	0.81	0.75	0.80	0.21	0.48
+ Reprompt [41]	1	0.72	0.99	0.79	0.68	0.84	0.51	0.52
+ Reprompt+GRPO [41]	1	0.76	0.99	0.87	0.77	0.85	0.62	0.49
+RAS	1	0.77	0.99	0.84	0.76	0.84	0.67	0.54
+Noise Scaling [24]	32	0.85	1.00	0.96	0.91	0.91	0.52	0.78
+RAS	2	0.85	0.99	0.94	0.84	0.88	0.77	0.66
+ Prompt Scaling [50]	32	0.87	0.99	0.94	0.85	0.91	0.80	0.71
+RAS	4	0.88	0.99	0.94	0.88	0.88	0.88	0.70
ReflectionFlow [50]	32	0.91	1.00	0.98	0.90	0.96	0.93	0.72
+RAS	8	0.91	1.00	0.98	0.93	0.93	0.88	0.75
+RAS	32	0.92	1.00	0.99	0.92	0.93	0.93	0.76

jointly reflects aesthetic quality and prompt alignment; we report its mean reward across generated images as a holistic fidelity metric.

Baselines. We compare against representative test-time scaling baselines that increase inference-time compute through global exploration, including noise scaling [24, 33] and prompt rewriting [41, 50]. We also include results from reflection-tuning methods ReflectDiT [18] and ReflectionFlow [50].

Implementation Details. For candidate selection, we follow prior work and employ SANA-Verifier [44] to select the best image among N samples. For our method, we use GPT-4o [14] to generate region prompts and bounding boxes for fair comparison with related methods that adopt the same planner; we provide a more detailed analysis of planner choices in Section 3.2. Unless otherwise stated, region-level guidance is injected only during the early denoising steps, after which sampling reverts to standard

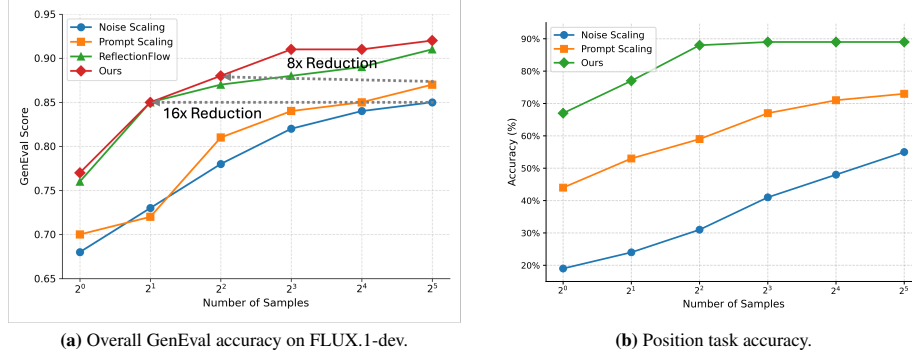


Fig. 5: Performance under different test-time scaling strategies. (a) Overall GenEval accuracy as a function of the sample budget on FLUX.1-dev. Noise scaling and prompt scaling improve gradually as the number of samples increases, while **RAS** achieves higher accuracy with fewer samples. (b) Accuracy on the Position task, where **RAS** reaches comparable or higher accuracy with substantially fewer samples than global scaling baselines.

global refinement. Specifically, we inject regional guidance for the first 3 of 30 steps on FLUX.1-dev, 4 of 20 steps on SANA-1.0-1.6B.

3.1 Main results

Comparison with Global Rewriting. Table 1 compares **RAS** with global prompt rewriting methods on the FLUX.1-dev backbone. While global rewriting (RePrompt) improves the overall GenEval score from 0.67 to 0.72, and reinforcement tuning (RePrompt+GRPO) further raises it to 0.76, their effectiveness is heavily limited in structural alignment. Specifically, even with RL-based tuning, these global methods struggle with **Position** (0.51–0.62) and **Attribution** (0.52–0.49). In contrast, **RAS** achieves an overall score of **0.77** with a single sample ($N = 1$)—outperforming all rewriting baselines without any specialized training. The performance gap is most pronounced in localized constraints: **RAS** improves Position to **0.67** and Attribution to **0.54**. These results demonstrate that while global rewriting enhances semantic specificity, it lacks the explicit spatial grounding provided by regional decomposition.

Scaling Efficiency and Performance. Beyond the single-sample setting, **RAS** demonstrates a massive leap in scaling efficiency. Compared to standard **Noise Scaling** on FLUX.1-dev, **RAS** matches the 32-sample baseline (0.85 overall) using **only 2 samples**—a $16\times$ reduction in sampling cost. Notably, **RAS** reaches a score of 0.85 in Position with only 2 samples, far exceeding the 0.52 achieved by 32-sample noise scaling. The scalability of **RAS** also generalizes to the efficient **SANA-1.0-1.6B** model, where it achieves 0.81 overall with **only 8 samples**, matching the 20-sample noise scaling baseline while reducing the sample cost by $2.5\times$. Under the high-budget regime ($N = 32$), **RAS** achieves a state-of-the-art score of **0.92** on FLUX.1-dev, proving that the “Region Axis” is a more principled scaling dimension than purely increasing global sampling

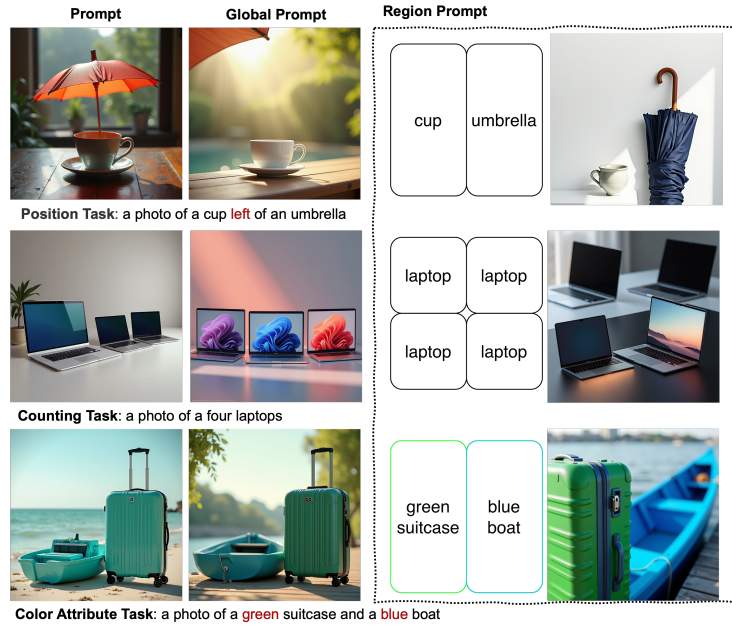


Fig. 6: Visualization of the transition from the simple prompt to the global prompt and region prompts. For each input prompt (left), our recaption module expands it into a richer and more descriptive global prompt (middle left), followed by scene decomposition into multiple semantic regions with region-level prompts (middle right). We show examples for three representative tasks: Position (cup left of umbrella) and Counting (four laptops). This visualization illustrates how RAS transforms a simple textual description into structured global and regional conditioning, enabling precise spatial and compositional control during generation.

diversity. The scaling behavior is further summarized in Figure 5, which plots overall GenEval and Position accuracy as a function of the sample budget. We provide qualitative examples of this prompt-to-region decomposition in Figure 6, illustrating how recaptioning and region prompts make compositional constraints more explicit.

Computation Cost Analysis In Table 1, we adopt the same LLM setting as prior work by using GPT-4o for planning/recaption to ensure a fair comparison. For prompt scaling, the method uses an LLM to rewrite the prompt and generate an image, then feeds the generated image back to the LLM for further refinement, following a sequential scaling paradigm (Fig. 4a). Table 2 further reports an end-to-end inference cost breakdown measured on an NVIDIA H100 GPU, where planning+recaption (**LLM**) is executed per generated sample, alongside the denoising generation cost (**Gen**) and the verifier runtime (**Verif**). We observe that the per-sample **LLM** overhead is comparable across methods and that image generation dominates the total runtime. Although RAS slightly increases the per-sample generation time due to regional guidance, its improved sample efficiency reduces the number of samples required to reach a target accuracy,

Table 2: Inference cost breakdown. Planning and recaption (**LLM**) are performed per generated sample. **Gen** denotes the denoising generation time and **Verif** the verifier runtime. All times are reported in seconds measured on an NVIDIA H100 GPU.

Method	N	LLM	Gen.	Verif	Total
Noise Scaling	32	0	9.4	0.3	310.8
Prompt Scaling	32	2.2	9.4	0.3	380.8
Ours (RAS)	4	2.2	10.2	0.3	50.8
Speedup	$7.5\times$ faster				

Table 3: Scaling configuration ablation.

We evaluate on the POSITION task. Accuracy is the prompt-level success rate (the fraction of prompts for which the configuration produces at least one correct sample).

#Noise	#Global	#Region	Acc.
8	1	–	82%
1	8	8	90%
2	4	4	92%
4	2	2	91%
8	1	1	89%

resulting in a $7.5\times$ lower end-to-end inference cost compared to global scaling in our setting. RAS achieves lower inference cost than global scaling under budget-matched comparisons.

3.2 Ablation Studies

CoT Planning and Planner Dependency. In our experiments, we find that **layout planning is a particularly challenging and error-prone component** of the pipeline. Even strong LLMs (e.g., GPT-5 [14, 32]) can produce incorrect spatial layouts, and replacing the planner with smaller open-source models such as Qwen2-VL-7B-Instruct [39] often exacerbates this issue. To disentangle the *planner* from the *region-injection mechanism*, we conduct two analyses. **(i) Planner robustness and failure analysis.** Tab. 4 shows that GPT-4o substantially outperforms Qwen2-VL-7B on layout parsing (90% vs. 52%), while rule-based parsing achieves near-perfect accuracy on simple spatial prompts. In practice, smaller LLMs frequently fail to correctly represent spatial relations, ignore relative object size, or generate overly small bounding boxes, which can lead to failure even after BoN selection. This indicates that planner quality matters, especially for more complex spatial reasoning tasks. **(ii) Component ablation.** Tab. 4 further shows that LLM rewrite and region injection provide complementary gains, indicating that the proposed RAS mechanism is not improved by the planner itself. To improve reliability given this sensitivity, we enforce *hard* rule-based constraints on bounding boxes to stabilize spatial relations. Developing stronger training-free planning methods based on open-source models remains an important direction for future work.

Noise Scaling or Prompt Scaling. We fix the total sampling budget ($\text{Noise} \times \text{Global} = 8$) and the regional layout, varying only how the budget is split across noise, global-prompt, and region-prompt scaling. This isolates the contribution of each axis under identical spatial structure. As shown in Table 3, scaling any single dimension to its extreme is suboptimal: allocating the full budget to noise (8/1/–) is the weakest setting (82%), while pushing it entirely toward prompt scaling (1/8/8) is competitive but not the best (90%). A balanced allocation (2/4/4) performs best (92%), with a moderately noise-heavy split (4/2/2) close behind (91%). This shows that noise diversity and

Table 4: Planner robustness and component ablation on the GenEval Position task. Left: layout parsing accuracy of different planners. We evaluate only the relative bounding-box relations extracted from planner outputs; “+Rule” enforces hard spatial constraints to guarantee correct relative positioning. Right: component ablation study.

Planner	Acc. $\uparrow$	Configuration	Acc. $\uparrow$
Qwen2-VL-7B	52%	FLUX vanilla	0.21
GPT-4o	90%	+ LLM Rewrite	0.40
GPT-4o + Rule	100%	+ LLM Plan + RAS	0.60
		+ Best-of-2	0.77

Table 5: Compositional generation on T2I-CompBench. Vanilla FLUX, RPG [5], and RAS, 300 images per category. All T2I-CompBench columns (Attribute Binding, Object Relationship, Complex) report GPT-4o judge scores normalized to $[0, 1]$ (higher is better, $\uparrow$); the last column is HPSv3 mean reward. The **best** per column is in **blue**.

Model	Attribute Binding			Object Relationship		Complex $\uparrow$	HPSv3 $\uparrow$
	Color $\uparrow$	Shape $\uparrow$	Texture $\uparrow$	Spatial $\uparrow$	Non-Spatial $\uparrow$		
Vanilla FLUX	0.8821	0.7304	0.8662	0.5442	0.8823	0.8308	9.22
RPG	0.8902	0.7534	0.8952	0.7042	0.9025	0.8358	9.08
RAS (Ours)	0.8921	0.7790	0.9107	0.7310	0.9133	0.8425	9.20

prompt scaling are complementary rather than substitutable: noise introduces diverse generation trajectories, while prompt scaling strengthens semantic and spatial grounding, and only their combination yields the largest gains. Regional conditioning also helps on its own—holding noise and global fixed, adding a single region prompt (8/1/— $\rightarrow$ 8/1/1) raises accuracy from 82% to 89%.

Comparison with RAG methods. We further evaluate RAS on **T2I-CompBench** (Tab. 5). RAS improves over vanilla FLUX across all compositional categories, with the largest gain on the challenging spatial relationship task (0.5442 $\rightarrow$ 0.7310). It also outperforms the region-based baseline RPG on every compositional metric, most notably on spatial relationship (0.7042 $\rightarrow$ 0.7310) and shape binding (0.7534 $\rightarrow$ 0.7790). Meanwhile, RAS stays essentially on par with vanilla FLUX in HPSv3 mean reward (9.22 $\rightarrow$ 9.20) while still exceeding RPG (9.08), indicating that its strong compositional gains come at negligible cost to overall image quality.

4 Related Work

Test-Time Scaling in Diffusion Models. For text-to-image diffusion models, *test-time scaling* (TTS) can modify the generation process through several training-free dimensions: (1) changing the number of denoising steps or trajectories (timestep scaling) [16, 17, 30, 43, 44], (2) sampling multiple noise latents (noise scaling) [11, 24], (3) generating prompt variants via rewriting or augmentation (prompt scaling) [6, 8, 41, 50], and

(4) ensembling multiple generators or editing models (model scaling) [40]. While these strategies expand the global sampling space, they remain *whole-image* conditioning and often fail to resolve fine-grained spatial misalignment in compositional prompts.

Related Scaling Paradigms and Our Positioning. Noise scaling produces multiple candidates and selects the best with a verifier, with recent variants biasing initialization toward more favorable latents [11, 24, 48]; however, many seeds converge to similar semantics under the same prompt, so additional sampling rarely fixes hard spatial relations (Figure 1). Prompt scaling perturbs the textual conditioning to improve semantic alignment [6, 8, 41], but global rewriting lacks explicit spatial grounding and therefore struggles with object-level constraints such as relative position or scale. Reflection-based scaling uses an auxiliary VLM to critique outputs and generate corrective instructions [18, 50], yet it introduces iterative sampling/feedback complexity and may require training a reflection-aware generator, deviating from the fully training-free setting we study. In parallel, region-aware generation and chain-of-thought reasoning improve scene construction and localized control [2, 5, 6, 8, 45], but existing frameworks can incur region-linear cost (e.g., RPG [45]) or require architectural modifications [5], and they are not designed as test-time scaling mechanisms. Our work targets this gap by turning regional decomposition into an efficient scaling extension: we inject region-level guidance during early denoising to steer structural layout while preserving global coherence, yielding a training-free, architecture-agnostic regional axis that complements conventional global scaling methods.

5 Conclusion

Our results show that conventional test-time scaling—varying global noise or prompts—often improves fidelity but remains sample-inefficient for compositional accuracy, especially in multi-object and spatially constrained scenes. Introducing a regional dimension provides localized control while preserving global coherence: applying regional guidance *early* in denoising yields consistent spatial gains with minimal overhead, and regional decomposition complements noise/prompt scaling by pushing the Pareto frontier of test-time scaling. We instantiate these insights with **Region-Aware Test-Time Scaling (RAS)**, a training-free approach that combines recaptioning, spatial decomposition, and early-stage regional conditioning, and integrates naturally with existing Best-of- N pipelines. Failure cases mainly arise in highly complex prompts with many objects and multiple pairwise relations, where reliable region planning becomes increasingly difficult and errors can propagate to generation. Improving robust planning and enabling selective, reflection-aware regional refinement are promising directions for future work.

6 Acknowledgement

This work is supported by the Imperial President’s PhD Scholarship and Global Fellows Fund, and the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2026, WBS: A-8004345-00-00).

References

1. Ahn, D., Kang, J., Lee, S., Min, J., Kim, M., Jang, W., Cho, H., Paul, S., Kim, S., Cha, E., et al.: A noise is worth diffusion guidance. arXiv preprint arXiv:2412.03895 (2024) [2](#)
2. Chen, A., Xu, J., Zheng, W., Dai, G., Wang, Y., Zhang, R., Wang, H., Zhang, S.: Training-free regional prompting for diffusion transformers. arXiv preprint arXiv:2411.02395 (2024) [3](#), [4](#), [15](#)
3. Chen, P., Shen, M., Ye, P., Cao, J., Tu, C., Bouganis, C.S., Zhao, Y., Chen, T.: Δ -dit: Accelerating diffusion transformers without training via denoising property alignment. International Journal of Computer Vision **134**(6), 276 (2026) [2](#)
4. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling (2025) [10](#)
5. Chen, Z., Li, Y., Wang, H., Chen, Z., Jiang, Z., Li, J., Wang, Q., Yang, J., Tai, Y.: Region-aware text-to-image generation via hard binding and soft refinement. arXiv preprint arXiv:2411.06558 (2024) [3](#), [4](#), [6](#), [7](#), [14](#), [15](#)
6. Duan, C., Fang, R., Wang, Y., Wang, K., Huang, L., Zeng, X., Li, H., Liu, X.: Got-r1: Unleashing reasoning capability of mllm for visual generation with reinforcement learning. arXiv preprint arXiv:2505.17022 (2025) [14](#), [15](#)
7. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Proceedings of the International Conference on Machine Learning (ICML) (2024) [2](#), [10](#)
8. Fang, R., Duan, C., Wang, K., Huang, L., Li, H., Yan, S., Tian, H., Zeng, X., Zhao, R., Dai, J., et al.: Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. arXiv preprint arXiv:2503.10639 (2025) [14](#), [15](#)
9. Gao, P., Zhuo, L., Liu, D., Du, R., Luo, X., Qiu, L., Zhang, Y., Huang, R., Geng, S., Zhang, R., et al.: Lumina-t2x: Scalable flow-based large diffusion transformer for flexible resolution generation. In: Proceedings of the International Conference on Learning Representations (ICLR) (2025) [2](#)
10. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems (NeurIPS) **36** (2024) [4](#), [9](#), [10](#)
11. Guo, X., Liu, J., Cui, M., Li, J., Yang, H., Huang, D.: Initno: Boosting text-to-image diffusion models via initial noise optimization. In: CVPR (2024) [14](#), [15](#)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems (NeurIPS) (2020) [2](#)
13. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. Advances in Neural Information Processing Systems **36**, 78723–78747 (2023) [9](#)
14. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024) [9](#), [10](#), [13](#)
15. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020) [2](#)
16. Labs, B.F.: Flux. <https://github.com/black-forest-labs/flux> (2024) [2](#), [10](#), [14](#)
17. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. <https://bfl.ai/blog/flux-2> (2025) [14](#)

18. Li, S., Kallidromitis, K., Gokul, A., Koneru, A., Kato, Y., Kozuka, K., Grover, A.: Reflect-dit: Inference-time scaling for text-to-image diffusion transformers via in-context reflection. arXiv preprint arXiv:2503.12271 (2025) 3, 9, 10, 15
19. Li, Z., Zhang, J., Lin, Q., Xiong, J., Long, Y., Deng, X., Zhang, Y., Liu, X., Huang, M., Xiao, Z., et al.: Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. arXiv preprint arXiv:2405.08748 (2024) 2
20. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: Proceedings of the International Conference on Learning Representations (ICLR) (2023) 2
21. Liu, B., Akhgari, E., Visheratin, A., Kamko, A., Xu, L., Shrirao, S., Lambert, C., Souza, J., Doshi, S., Li, D.: Playground v3: Improving text-to-image alignment with deep-fusion large language models (2024) 10
22. Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., Zhang, D., Ouyang, W.: Flow-grpo: Training flow matching models via online rl (2025), <https://arxiv.org/abs/2505.05470> 2
23. Liu, X., Gong, C., et al.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: Proceedings of the International Conference on Learning Representations (ICLR) (2022) 2
24. Ma, N., Tong, S., Jia, H., Hu, H., Su, Y.C., Zhang, M., Yang, X., Li, Y., Jaakkola, T., Jia, X., et al.: Inference-time scaling for diffusion models beyond scaling denoising steps. arXiv preprint arXiv:2501.09732 (2025) 2, 10, 14, 15
25. Ma, Y., Shui, Y., Wu, X., Sun, K., Li, H.: Hpsv3: Towards wide-spectrum human preference score (2025), <https://arxiv.org/abs/2508.03789> 9
26. Muennighoff, N., Yang, Z., Shi, W., Li, X.L., Fei-Fei, L., Hajishirzi, H., Zettlemoyer, L., Liang, P., Candès, E., Hashimoto, T.: sl: Simple test-time scaling. arXiv preprint arXiv:2501.19393 (2025) 2
27. Peebles, W.S., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). vol. 4172 (2022) 2
28. Qi, Z., Bai, L., Xiong, H., Xie, Z.: Not all noises are created equally: Diffusion noise selection and optimization. arXiv preprint arXiv:2407.14041 (2024) 2
29. Qin, Q., Zhuo, L., Xin, Y., Du, R., Li, Z., Fu, B., Lu, Y., Yuan, J., Li, X., Liu, D., et al.: Lumina-image 2.0: A unified and efficient image generative framework. arXiv preprint arXiv:2503.21758 (2025) 2
30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022) 2, 14
31. Shen, M., Chen, P., Ye, P., Xia, G., Chen, T., Bouganis, C.S., Zhao, Y.: Md-dit: Step-aware mixture-of-depths for efficient diffusion transformers. In: Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning 2
32. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025) 13
33. Singhal, R., Horvitz, Z., Teehan, R., Ren, M., Yu, Z., McKeown, K., Ranganath, R.: A general framework for inference-time scaling and steering of diffusion models. arXiv preprint arXiv:2501.06848 (2025) 2, 10
34. Snell, C., Lee, J., Xu, K., Kumar, A.: Scaling llm test-time compute optimally can be more effective than scaling model parameters. arXiv preprint arXiv:2408.03314 (2024) 2
35. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: Proceedings of the International Conference on Learning Representations (ICLR) (2021) 2

36. Tang, S., Cao, J., Lin, W., Hong, J., Zhang, B., Hu, S., Bai, L., Chen, T., Ouyang, W., Ye, P.: Open-source llms collaboration beats closed-source llms: A scalable multi-agent system. arXiv preprint arXiv:2507.14200 (2025) [2](#)
37. Tang, S., Lin, W., Ye, P., Ye, J., Li, H., Zhang, Y., Wang, X., Zhang, B., Hu, S., Chen, T., et al.: Beyond gemini-3-pro: Revisiting llm routing and aggregation at scale. arXiv preprint arXiv:2601.01330 (2026) [2](#)
38. VLLM, A.: Lumina-image 2.0 : A unified and efficient image generative model. <https://github.com/Alpha-VLLM/Lumina-Image-2.0/> (2025) [10](#)
39. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024) [13](#)
40. Wang, Z., Li, A., Li, Z., Liu, X.: Genartist: Multimodal llm as an agent for unified image generation and editing. *Advances in Neural Information Processing Systems* **37**, 128374–128395 (2024) [15](#)
41. Wu, M., Wang, L., Zhao, P., Yang, F., Zhang, J., Liu, J., Zhan, Y., Han, W., Sun, H., Ji, J., Sun, X., Lin, Q., Deng, W., Zhang, D., Sun, F., Zhang, Q., Ji, R.: Reprompt: Reasoning-augmented reprompting for text-to-image generation via reinforcement learning (2025), <https://arxiv.org/abs/2505.17540> [2](#), [10](#), [14](#), [15](#)
42. Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., Wang, S., Huang, T., Liu, Z.: Omnigen: Unified image generation. arXiv preprint arXiv:2409.11340 (2024) [2](#)
43. Xie, E., Chen, J., Chen, J., Cai, H., Tang, H., Lin, Y., Zhang, Z., Li, M., Zhu, L., Lu, Y., et al.: Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *Proceedings of the International Conference on Learning Representations (ICLR)* (2025) [2](#), [10](#), [14](#)
44. Xie, E., Chen, J., Zhao, Y., Yu, J., Zhu, L., Lin, Y., Zhang, Z., Li, M., Chen, J., Cai, H., et al.: Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. arXiv preprint arXiv:2501.18427 (2025) [2](#), [4](#), [10](#), [14](#)
45. Yang, L., Yu, Z., Meng, C., Xu, M., Ermon, S., Cui, B.: Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In: *Forty-first International Conference on Machine Learning* (2024) [3](#), [4](#), [6](#), [7](#), [15](#)
46. Zhang, Q., Lyu, F., Sun, Z., Wang, L., Zhang, W., Hua, W., Wu, H., Guo, Z., Wang, Y., Muennighoff, N., et al.: A survey on test-time scaling in large language models: What, how, where, and how well? arXiv preprint arXiv:2503.24235 (2025) [2](#)
47. Zhang, Y., Li, H., Chen, J., Zhang, H., Ye, P., Bai, L., Hu, S.: Beyond gpt-5: Making llms cheaper and better via performance-efficiency optimized routing. In: *Proceedings of the 2025 7th International Conference on Distributed Artificial Intelligence*. pp. 122–129 (2025) [2](#)
48. Zhou, Z., Shao, S., Bai, L., Xu, Z., Han, B., Xie, Z.: Golden noise for diffusion models: A learning framework. arXiv preprint arXiv:2411.09502 (2024) [2](#), [15](#)
49. Zhuo, L., Du, R., Xiao, H., Li, Y., Liu, D., Huang, R., Liu, W., Zhu, X., Wang, F.Y., Ma, Z., et al.: Lumina-next: Making lumina-t2x stronger and faster with next-dit. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2024) [2](#)
50. Zhuo, L., Zhao, L., Paul, S., Liao, Y., Zhang, R., Xin, Y., Gao, P., Elhoseiny, M., Li, H.: From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 15329–15339 (2025) [2](#), [3](#), [9](#), [10](#), [14](#), [15](#)