FarsightAlign TTS: Early-Stage Test-Time Scaling for Prompt-Aligned Text-to-Image Generation

Anonymous authors
Paper under double-blind review

Abstract

Text-to-Image diffusion models have achieved remarkable progress under the guidance of “Scaling Laws”, but further performance gains are increasingly hindered by diminishing returns from scaling model size and data volume. To bypass this bottleneck, Test-Time Scaling (TTS) has emerged as a promising alternative. However, the lack of interpretable signals in the early denoising steps forces existing TTS approaches to perform nearly complete denoising process for every candidate—resulting in high computational cost. In this work, we propose FarsightAlign TTS, a novel and efficient TTS method that leverages the rich semantic signals embedded in early cross-attention maps. With just a few denoising steps, FarsightAlign TTS can extract structured semantic information, such as object presence, layout, and attributes. It then leverages a lightweight scorer to prune unaligned candidates before committing to the final generation. This design significantly reduces computational cost while improving alignment with user’s prompts. The experimental results demonstrate the effectiveness of our method. Furthermore, FarsightAlign TTS can function as a plug-and-play module, significantly boosting the semantic alignment capabilities of other advanced TTS frameworks with minimal additional computational overhead.

1 Introduction

The remarkable progress of Text-to-Image (T2I) diffusion models largely stems from “Scaling Laws”—enhancing performance by increasing model parameters and training data (Hoffmann et al., 2022; Kaplan et al., 2020). However, this approach encounters a significant bottleneck. As computational investment during training continues to grow, the improvements in model performance exhibit diminishing marginal returns. A fundamental reason for this trend is that merely increasing data and parameter counts is insufficient for models to acquire a deep understanding of complex logic, physical commonsense, and causal relationships. Consequently, even the largest models often fail to precisely align with nuanced human intent when faced with complex prompts (Agarwal et al., 2025; Mañas et al., 2024).

In contrast to allocating massive resources for marginal gains in training, users in practical applications have adopted a highly cost-effective strategy to overcome model limitations: generating multiple candidate images from the same prompt and manually selecting the best one. This widespread “Best-of-N” practice demonstrates that adding extra computation during inference time can significantly improve the final output quality (Brown et al., 2024). Inspired by this strategy, we shift our research focus from expensive model training to the more flexible and low-cost inference stage.

This perspective aligns with Test-Time Scaling (TTS) proposed in the domain of Large Language Models (LLMs) (Zhang et al., 2025a). In LLMs, to overcome the limitations of greedy decoding, researchers employ more complex search algorithms like beam search (Snell et al., 2024) or tree search (Xie et al., 2024) in test time. These algorithms actively explore a wider range of generation paths during inference, thereby producing higher-quality response with superior logical consistency and contextual coherence.

However, directly applying TTS to diffusion models is challenging. LLMs generate discrete tokens that can be decoded and evaluated at each step, whereas diffusion models operate in a high-

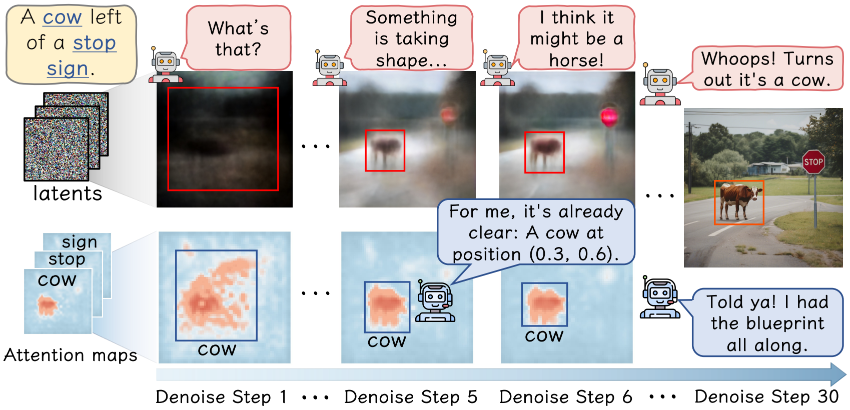

A cow left of a stop sign.
latents
What's that?
Something is taking shape...
I think it might be a horse!
Whoops! Turns out it's a cow.
sign
stop
cow
Attention maps
cow
cow
For me, it's already clear: A cow at position (0.3, 0.6).
cow
Told ya! I had the blueprint all along.
Denoise Step 1
Denoise Step 5
Denoise Step 6
Denoise Step 30

Figure 1: **Latent Decoding vs. Attention Maps.** In the early steps of diffusion, decoding intermediate latents (top row) produces perceptually blurry, semantically ambiguous images. In contrast, cross-attention maps (bottom row) already form a clear and accurate "blueprint" of the final scene, revealing precise object shape and location from the very beginning.

dimensional, non-interpretable latent noise space. The evaluation process for existing TTS methods in T2I tasks (Li et al., 2025b; Zhang et al., 2025c) requires decoding intermediate latents into full images. However, this presents a dilemma: decoding latents from early denoising stages yields blurry and uninformative images (as illustrated in Figure 1 and analyzed in Section 4.1), while decoding them from later stages incurs substantial computational cost due to the extensive number of denoising steps required. Furthermore, these methods focus only on general image quality, while overlooking whether the generated image truly aligns with the user's prompt (Agarwal et al., 2025).

These limitations demonstrate that an ideal TTS method requires **farsight**—the ability to discern a candidate's potential from the very early denoising steps. This capability can be leveraged to enhance T2I semantic **alignment**. Accordingly, we propose **FarsightAlign TTS**, a novel method built upon a key empirical observation: the cross-attention maps in the very early denoising steps already contain the critical semantic information that determines the final image composition (as shown in Figure 1).

FarsightAlign TTS operates by first sampling a large pool of initial noises. It then performs only a few denoising steps (just 5 steps) and, instead of costly decoding, directly extracts structured semantic features (object confidence, spatial location, attribute binding) from the early-stage cross-attention maps. A lightweight LLM scorer then rapidly prunes unpromising candidates based on their alignment with the prompt, allowing only the best to proceed to full generation.

Experimental results demonstrate that our method significantly outperforms existing TTS approaches on two mainstream benchmarks across three key evaluation metrics. The performance improvement is particularly pronounced in scenarios with limited computational resources.

Furthermore, the noise-filtering mechanism of FarsightAlign TTS is highly modular. We integrate this mechanism as a plug-and-play preprocessing module into other advanced diffusion TTS frameworks. Results show that with a negligible additional NFE cost, it can significantly boost their performance in semantic alignment.

To sum up, our main contributions are as follows:

- We propose FarsightAlign TTS, a computationally efficient TTS method that achieves strong semantic alignment with user prompts.
- As a plug-and-play module, FarsightAlign TTS seamlessly integrates with existing TTS methods, boosting performance with negligible computational overhead.
- Through extensive experiments, we prove that FarsightAlign TTS effectively utilizes test-time computation to consistently improve T2I performance, achieving a state-of-the-art comprehensive performance that surpasses all existing TTS methods.

2 RELATED WORKS

2.1 TEXT-TO-IMAGE DIFFUSION MODELS

In the field of T2I generation, early research primarily centered on GANs (Reed et al., 2016; Tao et al., 2022; Xu et al., 2018; Zhu et al., 2019) and autoregressive models (Ding et al., 2021; Gafni et al., 2022; Ramesh et al., 2021; Yu et al., 2022). As the scale of model parameters and training datasets has continued to expand, diffusion models have demonstrated powerful capabilities in image generation tasks (Balaji et al., 2022; Hoogeboom et al., 2023; Nichol et al., 2022; Ramesh et al., 2022; Saharia et al., 2022a). To mitigate the substantial computational costs associated with high-resolution image synthesis, Latent Diffusion Models have been introduced (Rombach et al., 2022). They enhance efficiency by performing the diffusion process within a low-dimensional latent space constructed by an autoencoder. For better alignment with the textual conditions, models like Stable Diffusion (Rombach et al., 2022; Podell et al., 2024; Esser et al., 2024) further incorporate cross-attention mechanisms (Chefer et al., 2023; Kim et al., 2025; Zhang et al., 2024b) to inject textual information. However, as we have highlighted in our introduction, despite their success, these models exhibit fundamental limitations in precisely follow complex prompts.

2.2 TEST-TIME SCALING

TTS improves model performance by leveraging extra computational resources during inference, typically through two approaches: 1) Generating diverse candidate outputs to increase the likelihood of high-quality results (Brown et al., 2024; Nguyen et al., 2024; Chen et al., 2024; Wang et al., 2023; Qiu et al., 2024; Zhang et al., 2025b), such as through Best-of-N sampling or self-consistency methods; 2) Extending reasoning depth, for example, using Chain-of-Thought prompting to guide step-by-step problem-solving for complex tasks (Wei et al., 2022; Madaan et al., 2023; Li et al., 2025a; Lightman et al., 2024; Xu et al., 2025). To further refine the outputs, researchers often combine intermediate-reward models (Yao et al., 2023; Zhang et al., 2024a; Xie et al., 2023) with search strategies (Xie et al., 2024; Zhou et al., 2024; Deng & Raffel, 2023) to dynamically guide and select among generation paths. However, due to the differences in architecture and generation mechanisms between diffusion models and LLMs, directly migrating TTS methods from LLM to T2I generation poses significant challenges.

2.3 TTS FOR T2I MODEL

A dominant implementation of TTS for T2I models is search-based guidance. This class of methods frames the multi-step denoising process as a decision tree, where each node represents an intermediate latent state, and employs search algorithms to explore the optimal generation path. For instance, researchers have successfully applied strategies like Dynamic Search (Li et al., 2025b), classic algorithms such as A* (Zhang et al., 2025c), and evolutionary algorithms (He et al., 2025). The core idea is to periodically assess the quality of intermediate nodes and prune low-quality branches, thereby focusing computation on more promising paths. Nevertheless, these approaches are computationally expensive due to the need to decode intermediate latents, and they also struggle to achieve precise semantic alignment with user prompts.

3 PRELIMINARIES

3.1 DENOISING PROCESS IN LATENT DIFFUSION MODELS

Diffusion models generate images by iteratively denoising a random latent vector, $z_T \sim \mathcal{N}(0, I)$. At each step t, a network ϵ_θ predicts the noise $\epsilon_\theta(z_t, t, c)$ conditioned on a text prompt c. A scheduler then computes the preceding latent z_{t-1} using this prediction:

$$z_{t-1} = \mathcal{S}(z_t, \epsilon_\theta(z_t, t, c)). \tag{1}$$

This process repeats until the final latent z_0 is reached, which a decoder $\mathcal{D}$ maps to the final image $x_0 = \mathcal{D}(z_0)$.

3.2 Cross-Attention Maps

Cross-attention layers align image generation with a text prompt by treating intermediate image features $\phi(z_t)$ as the Query (Q), and text embeddings $\tau(c)$ as the Key (K) and Value (V). At a timestep t, the attention matrix A is calculated as:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right). \tag{2}$$

Each column of A is an attention map for a specific text token. For our method, we create a single, unified attention map per token by extracting its map from all cross-attention layers, upsampling each to 64×64, and averaging the results. This aggregated map represents the token's total spatial influence and is the key signal we leverage for early-stage evaluation.

3.3 Test-Time Scaling

To extend the performance limits of the model in test time, a common strategy is to generate N candidates in parallel for a single prompt c. This is achieved by initializing a batch of independent latent vectors, $\{z_T^i\}_{i=1}^N$, to launch N parallel generation trajectories.

A reward function, $R(\cdot)$, is used to evaluate the quality of each trajectory. This evaluation can target the final image x_0^i, or intermediate states during the denoising process, enabling assessment and guidance of the generation.

The final output is the image $x_0^{i^*}$ from the trajectory with the highest cumulative reward score, S_i. The optimal index i^* is found by:

$$i^* = \underset{i \in \{1,\ldots,N\}}{\arg\max}\ S_i. \tag{3}$$

4 Method

4.1 Attention Maps vs. Decoded Latents

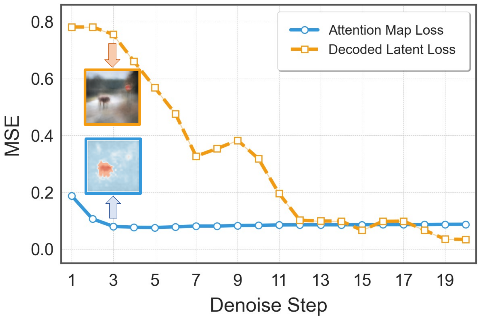


Figure 2: **Attention Map Loss vs. Decoded Latent Loss.** Attention maps (blue) provide a stable, low-error signal for object position throughout the denoising process, unlike the initially high and unstable decoded latent loss (orange).

A pivotal challenge for TTS in diffusion models lies in identifying a reliable signal at a very early denoising stage to avoid the high cost of a full denoising process for each candidate. A prevalent approach relies on decoding intermediate latents z_t into images x_t and then using reward models for evaluation. However, the critical flaw in this method is that early-stage decoded images are perceptually blurry and semantically ambiguous, as illustrated in Figure 1 (Top). This ambiguity prevents reward models from reliably assessing alignment with the text prompt, making the evaluation signal unreliable.

To address this, we propose using cross-attention maps directly as the evaluation proxy. We find that due to the inherent coupling between text tokens and their spatial layout, these maps form a clear semantic blueprint of the final image's object composition from the earliest denoising stages, as shown in Figure 1 (Bottom). This provides a far more stable and informative early signal.

To quantitatively compare these two proxies, we measure their positional error against a ground-truth position, which we obtain from the final image $D(z_0)$. We define $M_{\text{det}}(\cdot)$ as an operator that returns the center coordinate of the detected object's bounding box.

The Decode Loss, $\mathcal{L}_{\text{decode}}(t)$, is the Euclidean distance between the position predicted from the intermediate image $D(z_t)$ and the ground truth. The Attention Map Loss, $\mathcal{L}_{\text{attn}}(t)$, similarly measures the distance from the attention map's centroid to the ground truth. The losses are formulated as:

$$\mathcal{L}_{\text{decode}}(t) = \|M_{\text{det}}(D(z_t)) - M_{\text{det}}(D(z_0))\|_2 \tag{4}$$

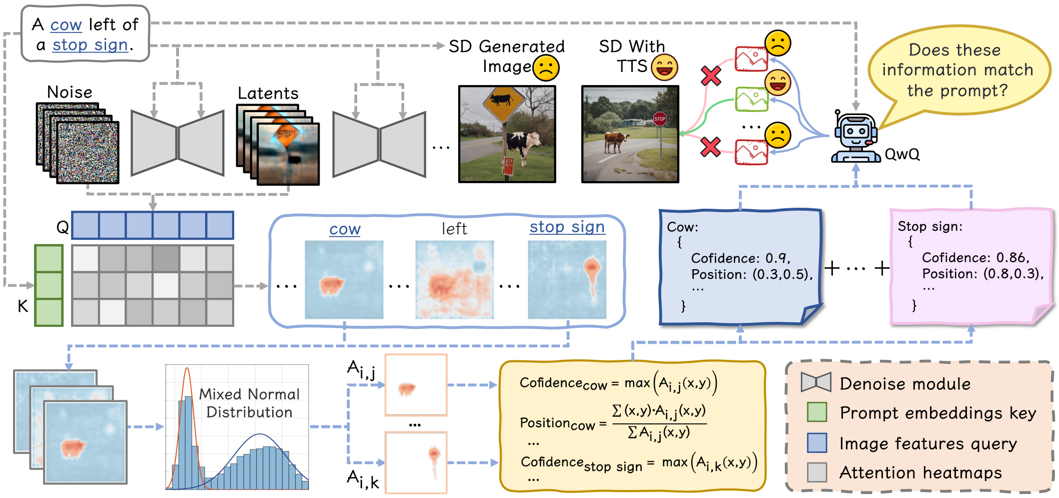


Figure 3: **An overview of FarsightAlign TTS.** To efficiently improve generation quality, we initiate multiple parallel candidates. Instead of costly full decoding for each candidate, we analyze their cross-attention maps at an early denoising stage. From these maps, we extract key semantic information including object confidence, position, and attribute binding scores. This information is used to score each candidate's alignment with the prompt, allowing us to select the highest-scoring candidate for the full generation process while discarding the rest.

$$\mathcal{L}_{\text{attn}}(t) = \|\text{Centroid}(A_t) - M_{\text{det}}(D(z_0))\|_2 \tag{5}$$

As shown in Figure 2, the results are definitive. $\mathcal{L}_{\text{attn}}(t)$ (blue line) is minimal and stable from the outset, whereas $\mathcal{L}_{\text{decode}}(t)$ (orange line) starts at a high value and diminishes slowly through the early denoising step. This stark divergence confirms that attention maps are a far superior early-stage proxy. This finding is the cornerstone of our method, enabling the efficient selection of an optimal initial latent z_T by sidestepping the need for costly denoising steps.

4.2 FarsightAlign TTS

Similar to other TTS approaches, we first initialize a pool of N initial noises for a given text prompt. After applying a few denoising steps to each, we extract the corresponding batch of attention maps, which are then processed in parallel to distill key semantic information.

4.2.1 Adaptive Object Extraction from Attention Maps

A straightforward approach to segmenting attention maps would be to apply a simple threshold. However, we find this method to be unreliable, as the optimal threshold—whether a fixed value or a percentile—varies dramatically across different tokens and denoising steps. To address this, we propose an adaptive method that requires no manually-tuned hyperparameters.

Our approach is based on the observation that the intensity distribution of an attention map typically exhibits a distinct bimodal structure, as shown in Figure 3. This bimodality is a natural consequence of the cross-attention mechanism; dot-product similarity scores, when transformed by the softmax function, inherently partition pixels into high-activation (object) and low-activation (background) groups.

Leveraging this property, we propose to explicitly model this structure by decomposing the distribution into two Normal Distribution representing the object and background, respectively. The parameters for this mixture are estimated via the Expectation-Maximization, providing a principled method for pixel-level segmentation. The probability density function for a pixel intensity x_i is given by:

$$p(x_i \mid \theta) = \alpha \cdot \mathcal{N}(x_i \mid \mu_o, \sigma_o^2) + (1 - \alpha) \cdot \mathcal{N}(x_i \mid \mu_b, \sigma_b^2) \tag{6}$$

here,(μ_o, σ_o^2) and (μ_b, σ_b^2) are the mean and variance for the high-activation object and low-activation background components, respectively, and α is the object's mixture proportion. Upon fitting the model, each pixel is assigned to the component—object or background—with the higher posterior probability. This results in a binary mask $\mathcal{R}_{i,j}$, representing the extracted object region for token j in candidate i.

Appendix A.2.3 provides a detailed performance comparison against fixed-thresholding segmentation approaches.

4.2.2 Object Generation Confidence

With the object region $\mathcal{R}_{i,j}$ extracted, we first quantify its generation confidence. A well-generated object corresponding to token j should manifest as a concentrated region of high activation in its attention map (validated in Appendix A.2.1). We therefore define the object's generation confidence as the peak intensity value within its object region $\mathcal{R}_{i,j}$. This provides a direct measure of how strongly the model focuses on the object. The confidence score is given by:

$$P_{i,j} = \max_{(x,y)\in\mathcal{R}_{i,j}} A_{i,j}(x,y) \tag{7}$$

where $P_{i,j}$ is the generation confidence for token j in candidate i, and $A_{i,j}(x,y)$ is the attention intensity at pixel coordinate (x,y).

4.2.3 Positional Information

Beyond confidence, we determine the object's location by computing the center of mass (centroid) of the high-activation region $\mathcal{R}_{i,j}$. This is a weighted average of the pixel coordinates, where the weights are the attention intensities, providing a robust estimate of the object's spatial position. The centroid coordinates $(\text{Pos}^i_{x,j}, \text{Pos}^i_{y,j})$ are computed as:

$$\text{Pos}^i_{x,j} = \frac{\sum_{(x,y)\in\mathcal{R}_{i,j}} x \cdot A_{i,j}(x,y)}{\sum_{(x,y)\in\mathcal{R}_{i,j}} A_{i,j}(x,y)}, \text{Pos}^i_{y,j} = \frac{\sum_{(x,y)\in\mathcal{R}_{i,j}} y \cdot A_{i,j}(x,y)}{\sum_{(x,y)\in\mathcal{R}_{i,j}} A_{i,j}(x,y)}, \tag{8}$$

where the summation is performed over all pixels within the high-activation region $\mathcal{R}_{i,j}$.

4.2.4 Attribute Binding

Quantifying the binding between an object o and an adjective a requires measuring how precisely the attribute's attention is focused on the object. Naive overlap metrics are insufficient as they ignore attention intensity. We therefore introduce a directional binding score, $B_{a\to o}$, defined as the mean attention intensity of the adjective's map, A_a, over the object's pre-segmented region, $\mathcal{R}_o$. Formally, this is expressed as:

$$B_{a\to o} = \frac{1}{|\mathcal{R}_o|} \sum_{(x,y)\in\mathcal{R}_o} A_a(x,y) \tag{9}$$

where $|\mathcal{R}_o|$ is the area of the object's region. A higher $B_{a\to o}$ score signifies a stronger semantic binding, indicating that the attribute is more precisely associated with the object.

4.2.5 LLM-Guided Candidate Selection

For each of the N initial candidates, we aggregate the extracted semantic metrics—object confidence, position, and attribute binding—into a structured JSON object. This creates a batch of N JSONs, each serving as a compact semantic summary of a potential outcome (see Figure 3 for an example).

This collection of semantic summaries is then scored by an LLM acting as a high-level semantic judge. We employ batch-processing to optimize for latency; the LLM evaluates a batch of candidate summaries against the original user prompt in each forward pass, assigning a holistic alignment score to each. For a detailed description of the LLM scorer, please refer to Appendix A.1.4.

Finally, the candidate with the highest score is selected for the full denoising process, while the remaining $N-1$ candidates are discarded.

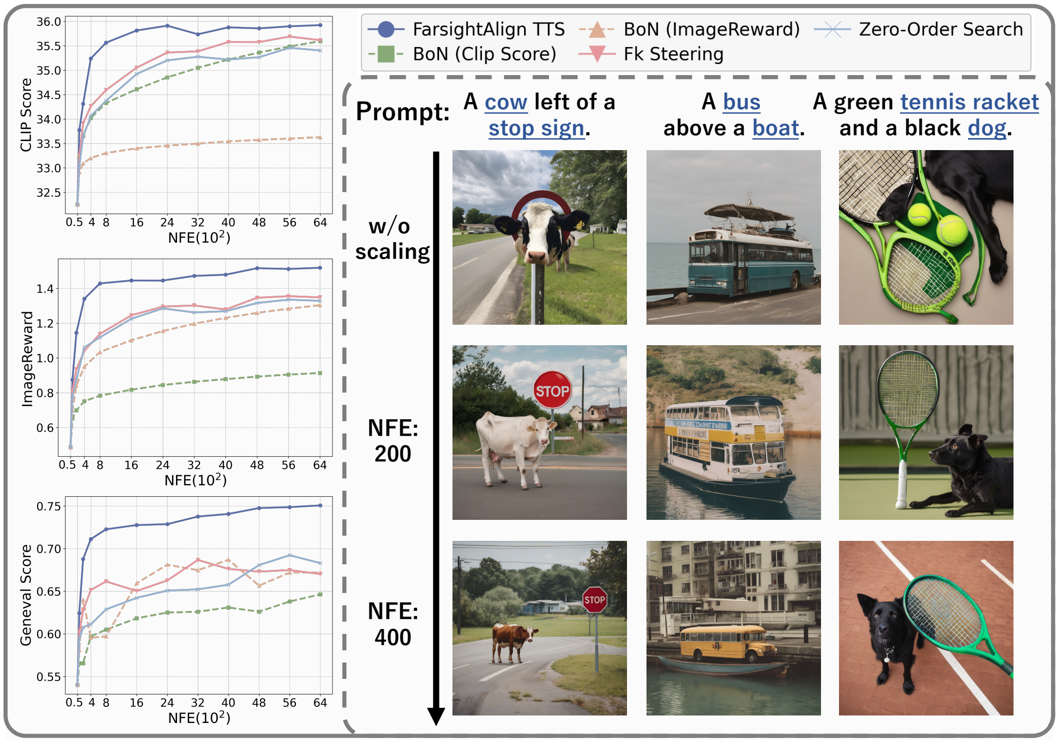


Figure 4: **FarsightAlign TTS outperforms baselines on the Geneval.** The plots (left) show superior quantitative performance across three metrics, while the images (right) demonstrate improved adherence to complex prompts. **See Appendix A.3 for additional qualitative comparisons.**

5 Experiments

5.1 Experiment Setup

Our tests are performed on two mainstream benchmarks, Geneval Benchmark (Ghosh et al., 2023) and DrawBench (Saharia et al., 2022b), using two distinct model architectures: the U-Net-based SDXL (Podell et al., 2024) and DiT-based SD3 (Esser et al., 2024). For both models, we use 50 denoising steps with other hyperparameters remaining as the default.

We compare our method against three representative TTS baselines. The first is Best-of-N (BoN), a straightforward approach that selects the best image from a pool of fully generated candidates via a reward model. The second is FK Steering (Singhal et al., 2025), a particle-sampling method that iteratively refines a set of candidates throughout the denoising process. The third is Zero-Order Search (Z-O Search) (Ma et al., 2025), an algorithm that finds an optimal solution through iterative neighborhood sampling and evaluation. Finally, to ensure an absolutely fair assessment, the metrics used for our final performance measurement—including Geneval Score, ImageReward Score (Xu et al., 2023), and Clip Score (Hessel et al., 2021)—are not integrated into FarsightAlign TTS's reward function.

5.2 Result Analysis

5.2.1 Comparative Advantages Over Existing TTS Methods

We first evaluate FarsightAlign TTS on the Geneval benchmark, measuring its test-time computation by the Number of Function Evaluations (NFE).

As illustrated in Figure 4, our method's performance consistently improves with increasing NFE, and outperforming other TTS methods. By eliminating unpromising candidates early in the denoising process, our method avoids wasting NFE on low-quality generation paths. This advantage is most significant in low-NFE scenarios, which align with practical usage constraints.

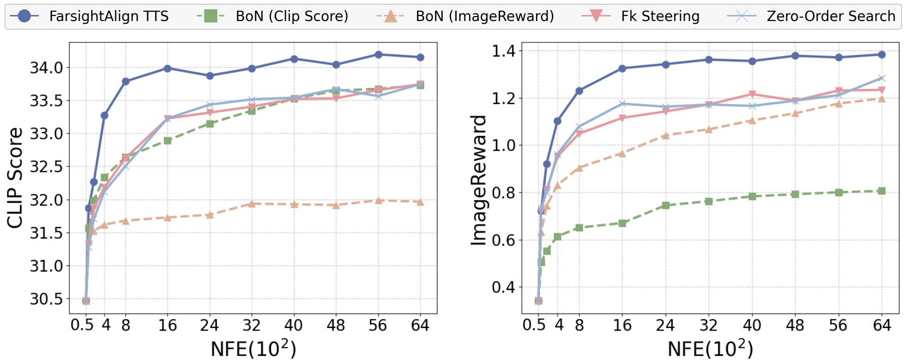


Figure 5: **Performance comparison on the DrawBench.** We also conduct experiments on the DrawBench using CLIP Score (left) and ImageReward (right).

Furthermore, we observe a critical phenomenon: while many existing TTS methods achieve stable gains on metrics they are directly optimized for (e.g., ClipScore and ImageReward), they exhibit significant performance volatility on the unseen metric: Geneval Score. This strongly suggests that these methods may be overfitting to specific reward models, failing to genuinely enhance the T2I model's ability to generate images that align with human semantics.

To further validate the generalization capabilities of our approach, we also conduct experiments on the DrawBench, using Clip Score and ImageReward as metrics. The results, illustrated in Figure 5, reaffirm that FarsightAlign TTS maintains a significant performance lead, demonstrating its robustness across different benchmarks.

5.2.2 Analysis of Efficiency

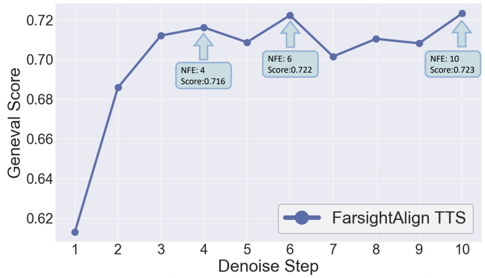


Figure 6: **Relationship between the Geneval Score and the denoising step chosen for pruning.** Performance gains saturate after just 4 steps.

The core of our method's efficiency lies in its ability to extract informative attention maps at a very early stage of the denoising process. To quantify this advantage and identify the optimal pruning point, we conduct an ablation study. Under a fixed total computational budget of 400 NFE, we evaluate the final performance when using attention maps extracted at different denoising steps (from 1 to 10) for pruning.

As shown in Figure 6, performance gains saturate after just 4 steps, indicating that we can effectively screen a candidate for only 4 NFE. In stark contrast, conventional TTS methods require the full generation process (50 NFE) to evaluate a single candidate, making our approach nearly 13 times more cost-effective. This drastic efficiency gap allows our method to explore a vastly larger search space of initial noises within the same NFE budget.

5.2.3 Synergy as a Plug-and-Play Module

Existing TTS methods often struggle in low-NFE scenarios, as the initial random sample pool may lack high-potential candidates. Leveraging the insight that initial noise critically shapes the final image (Mao et al., 2024), we propose FarsightAlign TTS as a plug-and-play pre-filtering module. It efficiently prunes a large set of initial noises at a minimal computational cost, providing a high-quality subset of candidates for a subsequent TTS search algorithm. This pre-filtering ensures the search begins with semantically aligned candidates, preventing wasted computation on poor initializations (see Appendix A.1.3 for implementation details).

To validate this synergy, we integrated our module with several baseline TTS methods. As shown in Figure 7, this integration substantially boosts performance on the Geneval benchmark. The augmented methods show the most significant gains in the low-NFE regime, confirming our approach's effectiveness.

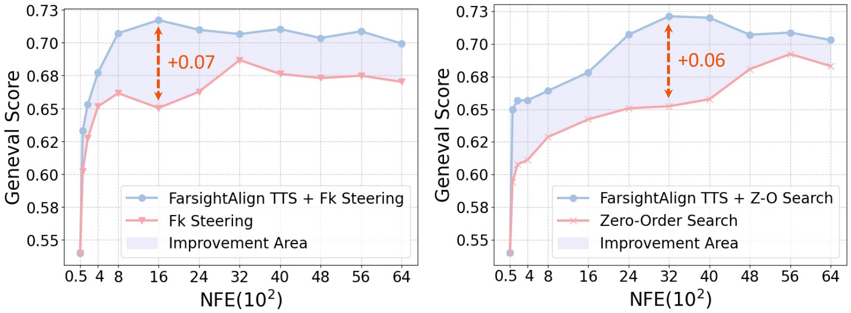


Figure 7: **Synergy with existing Test-Time Sampling methods.** FarsightAlign TTS significantly improves the Geneval Score of Fk Steering and Zero-Order Search, particularly at lower NFE.

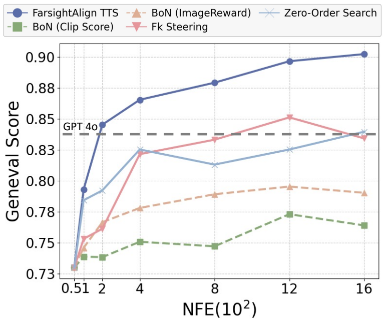


Figure 8: **Performance comparison on SD3.** Our method continues to significantly outperform all baselines on the Geneval benchmark.

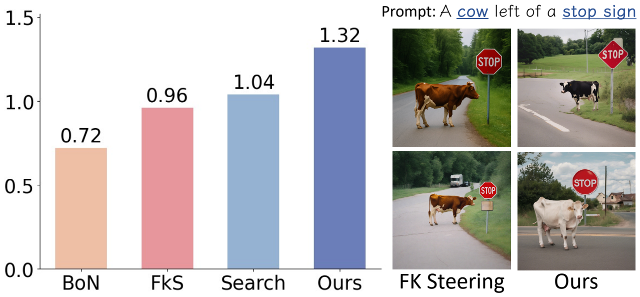


Figure 9: **Comparison of generation diversity.** The bar chart (left) provides a quantitative diversity score (based on the L2 distance of CLIP features), while the right panel shows a qualitative comparison of generated samples.

5.2.4 Generalizability Across Architectures

To further validate the architecture-agnostic nature and generalization capability of our method, we extend its application from the U-Net-based SDXL to the SD3, which is built upon DiT architecture. As shown in Figure 8, FarsightAlign TTS demonstrates superior performance on the Geneval benchmark with SD3 as well. Notably, with a computational budget of just 200 NFE (equivalent to generating 4 images), its performance even surpasses the powerful closed-source model, GPT-4o (OpenAI, 2023).

5.2.5 Maintaining Generation Diversity

A significant pitfall of TTS methods is ”reward hacking,” where over-reliance on reward models leads to reduced diversity (Ma et al., 2025). We evaluate this on the Geneval benchmark by measuring the diversity of 8 generated images (400 NFE budget) per prompt, quantified as the average L2 distance between their CLIP features. As shown in Figure 9, FarsightAlign TTS maintains significantly higher diversity than its competitors. We attribute this success to our pruning strategy, which relies on fundamental semantic alignment rather than external, potentially biased reward models that can hinder diversity.

6 Conclusion

We propose FarsightAlign TTS, a test-time scaling method that improves semantic alignment in Text-to-Image diffusion models. Instead of relying on costly full-image decoding, our approach extracts semantic cues—such as object presence and layout—from early-stage cross-attention maps, allowing efficient candidate pruning via a lightweight LLM scorer.

Through extensive experiments, we demonstrate that FarsightAlign TTS consistently surpasses existing TTS methods across multiple benchmarks, especially in low-computation settings. As a modular plug-in, it also enhances other TTS frameworks with minimal overhead.

REFERENCES

Aishwarya Agarwal, Srikrishna Karanam, and Balaji Vasan Srinivasan. Alignit: Enhancing prompt alignment in customization of text-to-image models. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2025, Tucson, AZ, USA, February 26 - March 6, 2025*, pp. 4882–4890. IEEE, 2025. doi: 10.1109/WACV61041.2025.00478. URL https://doi.org/10.1109/WACV61041.2025.00478.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *CoRR*, abs/2211.01324, 2022. doi: 10.48550/ARXIV.2211.01324. URL https://doi.org/10.48550/arXiv.2211.01324.

Bradley C. A. Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *CoRR*, abs/2407.21787, 2024. doi: 10.48550/ARXIV.2407.21787. URL https://doi.org/10.48550/arXiv.2407.21787.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Trans. Graph.*, 42(4):148:1–148:10, 2023. doi: 10.1145/3592116. URL https://doi.org/10.1145/3592116.

Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei A. Zaharia, and James Y. Zou. Are more LLM calls all you need? towards the scaling properties of compound AI systems. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/51173cf34c5faac9796a47dc2fdd3a71-Abstract-Conference.html.

Haikang Deng and Colin Raffel. Reward-augmented decoding: Efficient controlled text generation with a unidirectional reward model. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 11781–11791. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.721. URL https://doi.org/10.18653/v1/2023.emnlp-main.721.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 19822–19835, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/a4d92e2cd541fca87e4620aba658316d-Abstract.html.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=FPnUhsQJ5B.

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XV*, volume 13675 of *Lecture Notes in Computer Science*, pp. 89–106. Springer, 2022. doi: 10.1007/978-3-031-19784-0_6. URL https://doi.org/10.1007/978-3-031-19784-0_6.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a3bf71c7c63f0c3bcb7ff67c67b1e7b1-Abstract-Datasets_and_Benchmarks.html.

Haoran He, Jiajun Liang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Ling Pan. Scaling image and video generation via test-time evolutionary search. *CoRR*, abs/2505.17618, 2025. doi: 10.48550/ARXIV.2505.17618. URL https://doi.org/10.48550/arXiv.2505.17618.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 7514–7528. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.595. URL https://doi.org/10.18653/v1/2021.emnlp-main.595.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *CoRR*, abs/2203.15556, 2022. doi: 10.48550/ARXIV.2203.15556. URL https://doi.org/10.48550/arXiv.2203.15556.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13213–13232. PMLR, 2023. URL https://proceedings.mlr.press/v202/hoogeboom23a.html.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

Jeeyung Kim, Erfan Esmaeili, and Qiang Qiu. Text embedding is not all you need: Attention control for text-to-image semantic alignment with text self-attention maps. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 8031–8040. Computer Vision Foundation / IEEE, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Kim_Text_Embedding_is_Not_All_You_Need_Attention_Control_for_CVPR_2025_paper.html.

Chengpeng Li, Mingfeng Xue, Zhenru Zhang, Jiaxi Yang, Beichen Zhang, Xiang Wang, Bowen Yu, Binyuan Hui, Junyang Lin, and Dayiheng Liu. START: self-taught reasoner with tools. *CoRR*, abs/2503.04625, 2025a. doi: 10.48550/ARXIV.2503.04625. URL https://doi.org/10.48550/arXiv.2503.04625.

Xiner Li, Masatoshi Uehara, Xingyu Su, Gabriele Scalia, Tommaso Biancalani, Aviv Regev, Sergey Levine, and Shuiwang Ji. Dynamic search for inference-time alignment in diffusion models. *CoRR*, abs/2503.02039, 2025b. doi: 10.48550/ARXIV.2503.02039. URL https://doi.org/10.48550/arXiv.2503.02039.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi S. Jaakkola, Xuhui Jia, and Saining Xie. Inference-time scaling for diffusion models beyond scaling denoising steps. *CoRR*, abs/2501.09732, 2025. doi: 10.48550/ARXIV.2501.09732. URL https://doi.org/10.48550/arXiv.2501.09732.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Oscar Mañas, Pietro Astolfi, Melissa Hall, Candace Ross, Jack Urbanek, Adina Williams, Aishwarya Agrawal, Adriana Romero-Soriano, and Michal Drozdzal. Improving text-to-image consistency via automatic prompt optimization. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=g12Gdl6aDL.

Jiafeng Mao, Xueting Wang, and Kiyoharu Aizawa. The lottery ticket hypothesis in denoising: Towards semantic-driven initialization. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXIV*, volume 15132 of *Lecture Notes in Computer Science*, pp. 93–109. Springer, 2024. doi: 10.1007/978-3-031-72904-1_6. URL https://doi.org/10.1007/978-3-031-72904-1_6.

Alex Nguyen, Dheeraj Mekala, Chengyu Dong, and Jingbo Shang. When is the consistent prediction likely to be a correct prediction? *CoRR*, abs/2407.05778, 2024. doi: 10.48550/ARXIV.2407.05778. URL https://doi.org/10.48550/arXiv.2407.05778.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16784–16804. PMLR, 2022. URL https://proceedings.mlr.press/v162/nichol22a.html.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Jiahao Qiu, Yifu Lu, Yifan Zeng, Jiacheng Guo, Jiayi Geng, Huazheng Wang, Kaixuan Huang, Yue Wu, and Mengdi Wang. Treebon: Enhancing inference-time alignment with speculative tree-search and best-of-n sampling. *CoRR*, abs/2410.16033, 2024. doi: 10.48550/ARXIV.2410.16033. URL https://doi.org/10.48550/arXiv.2410.16033.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8821–8831. PMLR, 2021. URL http://proceedings.mlr.press/v139/ramesh21a.html.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022. doi: 10.48550/ARXIV.2204.06125. URL https://doi.org/10.48550/arXiv.2204.06125.

Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In Maria-Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 1060–1069. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/reed16.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10674–10685. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.org/10.1109/CVPR52688.2022.01042.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022a. URL http://papers.nips.cc/paper_files/paper/2022/hash/ec795aeadae0b7d230fa35cbaf04c041-Abstract-Conference.html.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022b. URL http://papers.nips.cc/paper_files/paper/2022/hash/ec795aeadae0b7d230fa35cbaf04c041-Abstract-Conference.html.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. *CoRR*, abs/2501.06848, 2025. doi: 10.48550/ARXIV.2501.06848. URL https://doi.org/10.48550/arXiv.2501.06848.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *CoRR*, abs/2408.03314, 2024. doi: 10.48550/ARXIV.2408.03314. URL https://doi.org/10.48550/arXiv.2408.03314.

Ming Tao, Hao Tang, Fei Wu, Xiaoyuan Jing, Bing-Kun Bao, and Changsheng Xu. DF-GAN: A simple and effective baseline for text-to-image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 16494–16504. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01602. URL https://doi.org/10.1109/CVPR52688.2022.01602.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL https://qwenlm.github.io/blog/qwq-32b/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural*

Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, James Xu Zhao, Min-Yen Kan, Junxian He, and Michael Qizhe Xie. Self-evaluation guided beam search for reasoning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/81fde95c4dc79188a69ce5b24d63010b-Abstract-Conference.html.

Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P. Lillicrap, Kenji Kawaguchi, and Michael Shieh. Monte carlo tree search boosts reasoning via iterative preference learning. *CoRR*, abs/2405.00451, 2024. doi: 10.48550/ARXIV.2405.00451. URL https://doi.org/10.48550/arXiv.2405.00451.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/33646ef0ed554145eab65f6250fab0c9-Abstract-Conference.html.

Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. Chain of draft: Thinking faster by writing less. *CoRR*, abs/2502.18600, 2025. doi: 10.48550/ARXIV.2502.18600. URL https://doi.org/10.48550/arXiv.2502.18600.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 1316–1324. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00143. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Xu_AttnGAN_Fine-Grained_Text_CVPR_2018_paper.html.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Trans. Mach. Learn. Res.*, 2022, 2022. URL https://openreview.net/forum?id=AFDcYJKhND.

Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Zhihan Guo, Yufei Wang, Irwin King, Xue Liu, and Chen Ma. What, how, where, and how well? A survey on test-time scaling in large language models. *CoRR*, abs/2503.24235, 2025a. doi: 10.48550/ARXIV.2503.24235. URL https://doi.org/10.48550/arXiv.2503.24235.

Qiyuan Zhang, Yufei Wang, Yuxin Jiang, Liangyou Li, Chuhan Wu, Yasheng Wang, Xin Jiang, Lifeng Shang, Ruiming Tang, Fuyuan Lyu, and Chen Ma. Crowd comparative reasoning: Unlocking comprehensive evaluations for llm-as-a-judge. *CoRR*, abs/2502.12501, 2025b. doi: 10.48550/ARXIV.2502.12501. URL https://doi.org/10.48550/arXiv.2502.12501.

Xiangcheng Zhang, Haowei Lin, Haotian Ye, James Y. Zou, Jianzhu Ma, Yitao Liang, and Yilun Du. Inference-time scaling of diffusion models through classical search. *CoRR*, abs/2505.23614, 2025c. doi: 10.48550/ARXIV.2505.23614. URL `https://doi.org/10.48550/arXiv.2505.23614`.

Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. Chain of preference optimization: Improving chain-of-thought reasoning in llms. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024a. URL `http://papers.nips.cc/paper_files/paper/2024/hash/00d80722b756de0166523a87805dd00f-Abstract-Conference.html`.

Yasi Zhang, Peiyu Yu, and Ying Nian Wu. Object-conditioned energy-based attention map alignment in text-to-image diffusion models. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLII*, volume 15100 of *Lecture Notes in Computer Science*, pp. 55–71. Springer, 2024b. doi: 10.1007/978-3-031-72946-1_4. URL `https://doi.org/10.1007/978-3-031-72946-1_4`.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning, acting, and planning in language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL `https://openreview.net/forum?id=njwv9BsGHF`.

Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 5802–5810. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00595. URL `http://openaccess.thecvf.com/content_CVPR_2019/html/Zhu_DM-GAN_Dynamic_Memory_Generative_Adversarial_Networks_for_Text-To-Image_Synthesis_CVPR_2019_paper.html`.

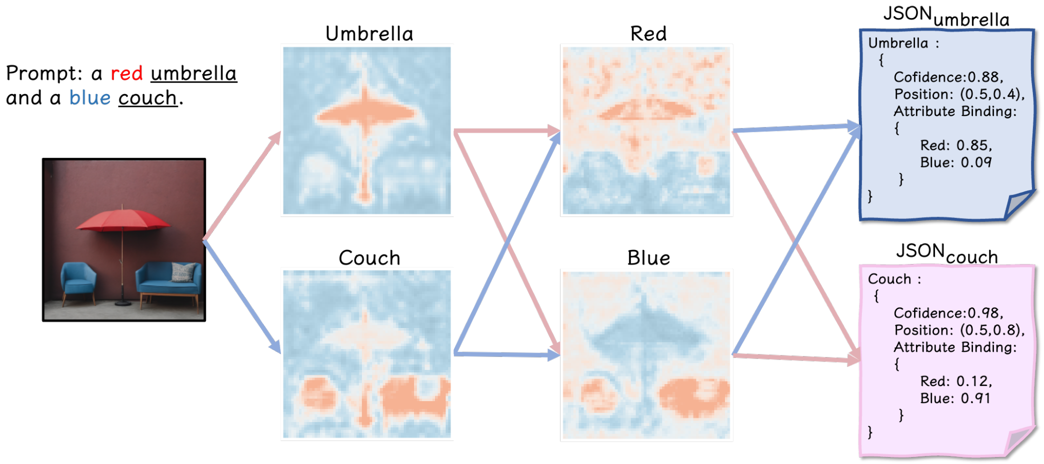
A Appendix

Appendix Directory

A.1 Experimental Setup and Implementation

A.1.1 Datasets

We provide further details on the benchmarks used for our evaluation.



```
Umbrella :
  {
     Cofidence:0.88,
     Position: (0.5,0.4),
     Attribute Binding:
     {
          Red: 0.85,
          Blue: 0.09
      }
}
```

```
Couch :
 {
     Cofidence:0.98,
     Position: (0.5,0.8),
     Attribute Binding:
     {
          Red: 0.12,
          Blue: 0.91
      }
}
```

Figure 10: Information Extraction Pipeline. An illustration of how FarsightAlign processes cross-attention maps to produce a structured JSON summary. This summary quantifies object confidence, position, and attribute bindings, forming the basis for our efficient, early-stage candidate pruning.

Geneval Benchmark: An object-centric framework to evaluate T2I models on complex prompts. It contains 500 prompts across four categories: attribute binding, object relationships, spatial reasoning, and multi-object composition. Geneval provides a more granular, instance-level analysis than holistic metrics like FID, helping to identify specific model failure modes.

DrawBench: A diverse benchmark with 200 challenging prompts designed to probe the limits of T2I models. It specifically tests difficult aspects like compositionality, negation, and unusual object interactions.

A.1.2 Evaluation Metrics

We employed three metrics for a multi-dimensional comparison. For a fair evaluation, none were used to guide the FarsightAlign TTS process.

CLIP Score: Measures semantic consistency between the prompt and the image using the cosine similarity of CLIP embeddings (ViT-B/32).

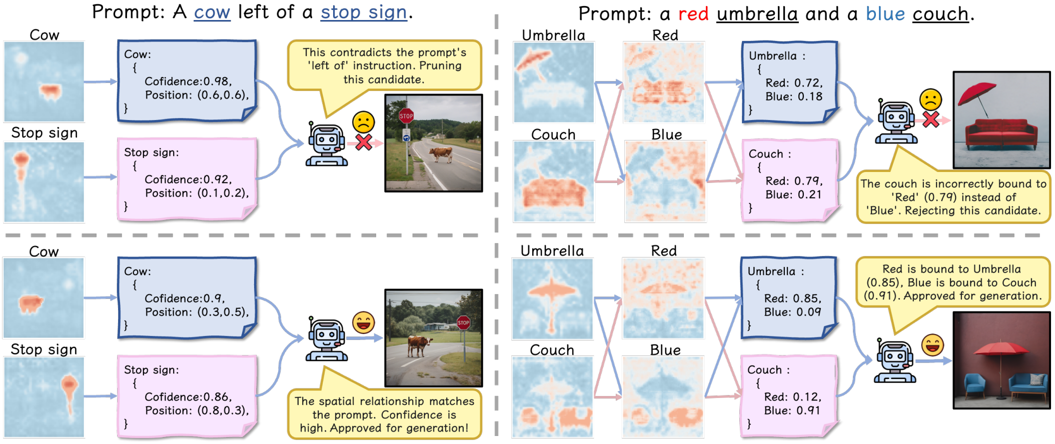
ImageReward: A reward model trained on human preferences to evaluate overall quality, including image aesthetics, realism, and text-to-image alignment.

Geneval Score: The native, object-centric eva luation suite for the Geneval benchmark. It performs fine-grained, automated verification of semantic elements (e.g., presence, attributes, position) using a vision-based pipeline. This provides a strict measure of prompt fidelity, less susceptible to reward hacking than holistic scores.

A.1.3 Plug-and-Play Pipeline Implementation

In this section, we provide a detailed description of the implementation for integrating FarsightAlign TTS as a plug-and-play module. We outline the two-stage pipeline that enables FarsightAlign TTS to act as an efficient pre-filtering stage, thereby optimizing the initial candidate pool for subsequent complex search algorithms like FK Steering and Zero-Order Search.

Stage 1: FarsightAlign Pre-filtering. The process begins by sampling a large pool of $8 * N$ initial noise latents $\{z_{T,i}\}_{i=1}^{8*N}$. Instead of a full denoising process, we perform only a small number of denoising steps, T_{prune} (typically 5 steps), for all $8 * N$ candidates. At step $T - T_{\text{prune}}$, we extract the cross-attention maps for each candidate and use our scorer to evaluate its semantic alignment with the prompt. We then rank all candidates by their scores and select the initial noise latents of the top-N candidates, forming a high-potential subset $\{z^*_{T,j}\}_{j=1}^{N}$.


Prompt: A cow left of a stop sign.
Cow
Stop sign
Cow: { Cofidence:0.98, Position: (0.6,0.6), }
Stop sign: { Cofidence:0.92, Position: (0.1,0.2), }
STOP
Cow
Stop sign
Cow: { Cofidence:0.9, Position: (0.3,0.5), }
Stop sign: { Cofidence:0.86, Position: (0.8,0.3), }
STOP
Prompt: a red umbrella and a blue couch.
Umbrella
Red
Couch
Blue
Umbrella : { Red: 0.72, Blue: 0.18 }
Couch : { Red: 0.79, Blue: 0.21 }
Umbrella
Red
Couch
Blue
Umbrella : { Red: 0.85, Blue: 0.09 }
Couch : { Red: 0.12, Blue: 0.91 }

This contradicts the prompt's 'left of' instruction. Pruning this candidate.

The spatial relationship matches the prompt. Confidence is high. Approved for generation!

The couch is incorrectly bound to 'Red' (0.79) instead of 'Blue'. Rejecting this candidate.

Red is bound to Umbrella (0.85), Blue is bound to Couch (0.91). Approved for generation.

Figure 11: Early-Stage Pruning Mechanism. Our method successfully identifies and prunes semantically flawed candidates at an early stage by reasoning over structured data derived from attention maps. It correctly rejects candidates with incorrect spatial layouts (left) and mismatched attribute bindings (right), while approving logically sound ones.

Stage 2: Main TTS Method Execution. This pre-screened, high-potential subset of N initial latents is then passed as the starting candidate pool to the main TTS method (e.g., FK Steering or Z-O Search). The main method then executes its full, more computationally intensive algorithm—be it iterative particle refinement or neighborhood search—but its exploration is now confined to this semantically strong set of candidates.

A.1.4 LLM Scorer Implementation

The LLM scorer is the high-level reasoning engine of FarsightAlign TTS. Its primary role is to interpret the structured semantic data extracted from early-stage attention maps and evaluate its alignment with the complex, compositional intent of the user's prompt.

For each candidate, the extracted metrics—object confidence, position, and attribute binding scores—are aggregated into a structured JSON object. This "semantic summary", as illustrated in Figure 10, provides a compact, machine-readable representation of the potential image.

To optimize resource consumption and minimize latency, we aggregate the summaries from multiple candidates into a single batch and feed them to the LLM in one forward pass. The model then returns a corresponding batch of evaluation scores.

To balance high throughput with model performance, we process the semantic summaries in batches. A primary constraint is the context length of the LLM, as excessively long inputs can degrade reasoning quality. We therefore set our batch size to 64. This choice is calibrated based on the capabilities of powerful models such as QwQ, which have been shown to maintain excellent performance even with context lengths approaching 8192 tokens (Team, 2025). A batch of 64 JSON summaries (totaling approximately 6000 tokens) remains comfortably within this high-performance range, ensuring both processing efficiency and reliable evaluation.

To elicit the reasoning shown in our examples (as shown in Figure 11), we use the specific zero-shot system prompt detailed in Figure 12. This template, which defines the LLM's role, evaluation logic, and required output format, is applied to each candidate's summary within the batch.

A.2 Ablation Studies and Component Analysis

A.2.1 Validation of Object Generation Confidence

This section provides the visual evidence for our core claim in Section 4.2.2: that the peak intensity of an early-stage attention map is a highly reliable predictor of whether a corresponding object will be successfully generated.

```
You are an expert evaluator for a Text-to-Image
model. Your task is to evaluate a LIST of
candidate semantic summaries against a single
user prompt. For each summary in the list, you
must evaluate it across three distinct criteria,
providing a score from 1 to 10 for each.

Your evaluation criteria for EACH candidate are:
1. Object Presence: Are all key objects present
   with high confidence?
2. Spatial Relationships: Are spatial instructions
   (e.g., 'left of') correctly represented?
3. Attribute Bindings: Are attributes (e.g., 'red')
   correctly assigned to the right objects?

USER PROMPT:
{user_prompt}

LIST OF CANDIDATE SUMMARIES (JSON ARRAY):
{list_of_json_summaries}

Based on your evaluation, provide a JSON array.
Each object in the array should contain the three
scores for the corresponding candidate summary.
The order of objects in your output array MUST
match the order of the summaries in the input
array. Your output must be only the JSON array.

Example Output (for a batch of 2):
[
  {
     "object_presence_score": 8,
     "spatial_relations_score": 9,
     "attribute_binding_score": 10
  },
  {
     "object_presence_score": 10,
     "spatial_relations_score": 2,
     "attribute_binding_score": 7
  }
]
JSON ARRAY OUTPUT:
```

Figure 12: The prompt template for batched, factorized scoring. It instructs the LLM to process a list of candidate summaries and return a corresponding list of JSON objects containing the evaluation scores, ensuring high throughput and parsable output.

Figure 13 demonstrates this direct correlation by contrasting two sets of outcomes. The key distinction is not merely the quality of the final image, but whether the prompted object is present or absent—a factor dictated by the strength of its attention map.

- **Top Row (Low Peak Intensity → Failed Generation):** This row illustrates cases where the model failed to generate a key object. For instance, the relevant concepts required by the prompt (e.g., "cow" or "dog") are malformed, incoherent, or entirely absent in the final images. Their corresponding attention maps are visibly **diffuse, scattered, and lack any distinct high-intensity peak**. This weak, unfocused attention signals the model's failure to ground the concept, directly leading to the object's failed generation in the final output.

- **Bottom Row (High Peak Intensity → Successful Generation):** In stark contrast, this row showcases successful generations where all prompted objects are accurately rendered. The attention maps corresponding to these objects—such as the "cow" or "dog"—exhibit a **sharp, localized, and high-intensity peak**. This focused activation indicates that the model has confidently materialized the concept onto the canvas, directly resulting in a well-formed and clearly recognizable object.

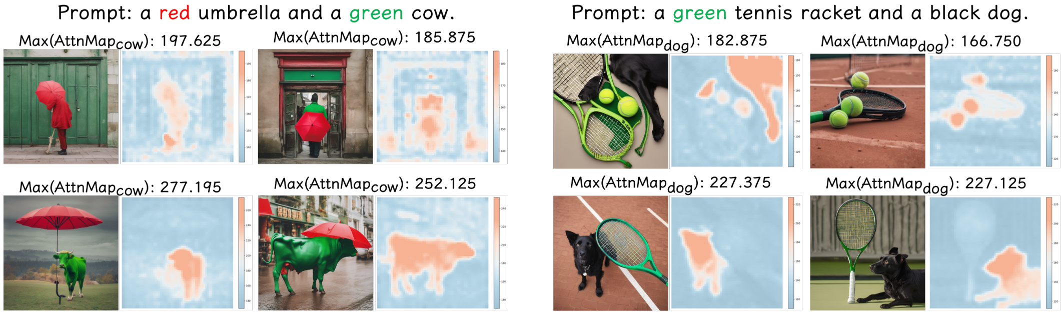

Prompt: a red umbrella and a green cow.
Max(AttnMap$_{cow}$): 197.625
Max(AttnMap$_{cow}$): 185.875
Max(AttnMap$_{cow}$): 277.195
Max(AttnMap$_{cow}$): 252.125
Prompt: a green tennis racket and a black dog.
Max(AttnMap$_{dog}$): 182.875
Max(AttnMap$_{dog}$): 166.750
Max(AttnMap$_{dog}$): 227.375
Max(AttnMap$_{dog}$): 227.125

Figure 13: **Visual Evidence for FarsightAlign's Confidence Metric.** A comparison showing that successfully generated objects (bottom row) correspond to early-stage attention maps with high-intensity, focused peaks, while failed generations (top row) correspond to diffuse, low-intensity maps.

Method	Geneval Score (%)
w/o TTS	20.03
BoN (Clip Score)	**28.99**
BoN (Peak Intensity)	**28.90**

Table 1: **Performance on GenEval Benchmark (Color Attribution).** Comparison of our BoN approach using peak intensity as a reward signal against a baseline model.

Comparing the attention peaks of the top and bottom rows reveals that the peaks for the successfully generated targets in the bottom row are significantly higher than those in the top row.

Furthermore, we conducted a supplementary study to further validate the efficacy of attention peak intensity as a predictor of generation success. In this experiment, we employed peak intensity as a reward signal within a BoN sampling strategy.

For each prompt, we generated 32 candidate samples. We then calculated the early-stage attention peak intensity for the specified object in each sample and selected the image with the highest peak intensity as the final output. This methodology was evaluated on the GenEval benchmark, with a particular focus on the ”Color Attribution” category of prompts, which are notoriously prone to object omission.

As shown in Table 1, our method demonstrates a significant improvement on the GenEval benchmark (Color Attribution), achieving performance comparable to BoN guided by CLIP scores. Critically, our approach provides this competitive performance without requiring the computationally expensive step of full image denoising for candidate selection.

In summary, these experiments demonstrate that early-stage attention peak intensity is a powerful and efficient metric. The visual evidence confirms it is a reliable predictor of successful object generation. Furthermore, when applied as a reward signal, it becomes a computationally cheap optimization tool that improves output quality without the need to fully generate every candidate image, unlike more costly methods such as CLIP-based scoring.

A.2.2 Ablation on Scorer Choice

To determine the optimal LLM scorer for our framework, we conducted an ablation study to evaluate the impact of different models on final generation quality.

The study was performed under a fixed computational budget of NFE=400 for the generation process, ensuring a fair comparison where the primary variable was the scorer model. We evaluated four distinct models:

- QwQ-32B (Full Precision)
- QwQ-32B (8-bit Quantized)

Scorer Model	Geneval Score
Qwen2.5-7B	0.685
Qwen3-30B-A3B	0.723
QwQ-32B (Full Prec.)	0.720
QwQ-32B (8-bit Quant.)	0.716

Table 2: Ablation study on scorer model choice under a fixed NFE=400 budget. Performance is measured by the final Geneval Score. The 8-bit quantized QwQ-32B provides the best balance of performance and efficiency.

- Qwen2.5-7B
- Qwen3-30B-A3B

The results, summarized in Table 2, reveal several key insights. Interestingly, most of the larger models (QwQ-32B, QwQ-32B 8-bit, and Qwen3-30B-A3B) yielded comparable final generation quality, with only marginal differences in performance metrics. The smallest model, Qwen2.5-7B, performed slightly worse, suggesting that a certain threshold of reasoning capability is necessary for the task.

Crucially, the 8-bit quantized version of QwQ-32B achieved performance nearly identical to its full-precision counterpart. This demonstrates that for our specific task of evaluating structured JSON data, the quantization process does not lead to a significant degradation in the model's critical reasoning abilities. Based on this analysis, we selected the 8-bit quantized QwQ-32B as the final scorer.

A.2.3 Analysis of Adaptive Object Extraction

As stated in the main paper (Section 4.2), a core component of our framework is the ability to reliably segment an object's region from its cross-attention map. This section provides the detailed performance comparison that justifies our choice of an adaptive segmentation method over simpler thresholding techniques.

To comprehensively evaluate the quality of our adaptive segmentation method, we designed two distinct experiments. The first assesses the positional accuracy of the extracted masks at an early stage, while the second measures the downstream impact of the segmentation method on the final image generation quality.

Experiment 1: Positional Accuracy of Extracted Objects A high-quality segmentation should accurately predict the final position of an object, even from an early denoising step. To measure this, we designed the following evaluation:

1. At an early denoising step (5 step), we use each segmentation method to extract the object mask and compute its geometric center (centroid).
2. After the full denoising process is complete, we run a object detector on the final generated image to obtain the ground-truth bounding box for the object.
3. We then calculate the Mean Positional Error, defined as the average Euclidean distance between the predicted centroid from the early stage and the center of the final ground-truth bounding box.

The results, averaged over the Geneval benchmark, are presented in Table 3. A lower error indicates that the segmentation method provides a more accurate and stable localization signal early on. Our adaptive method achieves a significantly lower positional error, demonstrating its superior ability to pinpoint the object's final location.

Experiment 2: Impact on Final Generation Quality To assess how segmentation quality affects the end-to-end performance of our framework, we conducted a direct comparison. We integrated both the Percentile Thresholding and our Adaptive method into the full FarsightAlign pipeline and evaluated their final output on the Geneval benchmark under a fixed computational budget of NFE=400.

Segmentation Method	Mean Positional Error
Fixed Threshold (T=200)	0.38
Percentile Threshold (P=90%)	0.15
Ours	**0.07**

Table 3: Positional accuracy of different segmentation methods. The error is the L2 distance between the centroid predicted at an early stage and the final object center. Lower is better.

Segmentation Method	Geneval Score
Percentile Threshold (P=90%)	0.697
Ours	**0.716**

Table 4: End-to-end performance of FarsightAlign using different segmentation modules under a fixed NFE=400 budget. Higher is better.

As shown in Table 4, the choice of segmentation method has a direct and notable impact on the final result. By providing more accurate semantic information to the LLM scorer, our adaptive approach leads to a clear improvement in the final Geneval Score.

A.3 Qualitative Results

We present extensive qualitative results below. We compare FarsightAlign TTS against competing TTS methods on Geneval benchmark (Figures 14 and 15) and DrawBench (Figures 16 and 17). Across these scenarios, our method consistently selects candidates that exhibit superior semantic alignment.

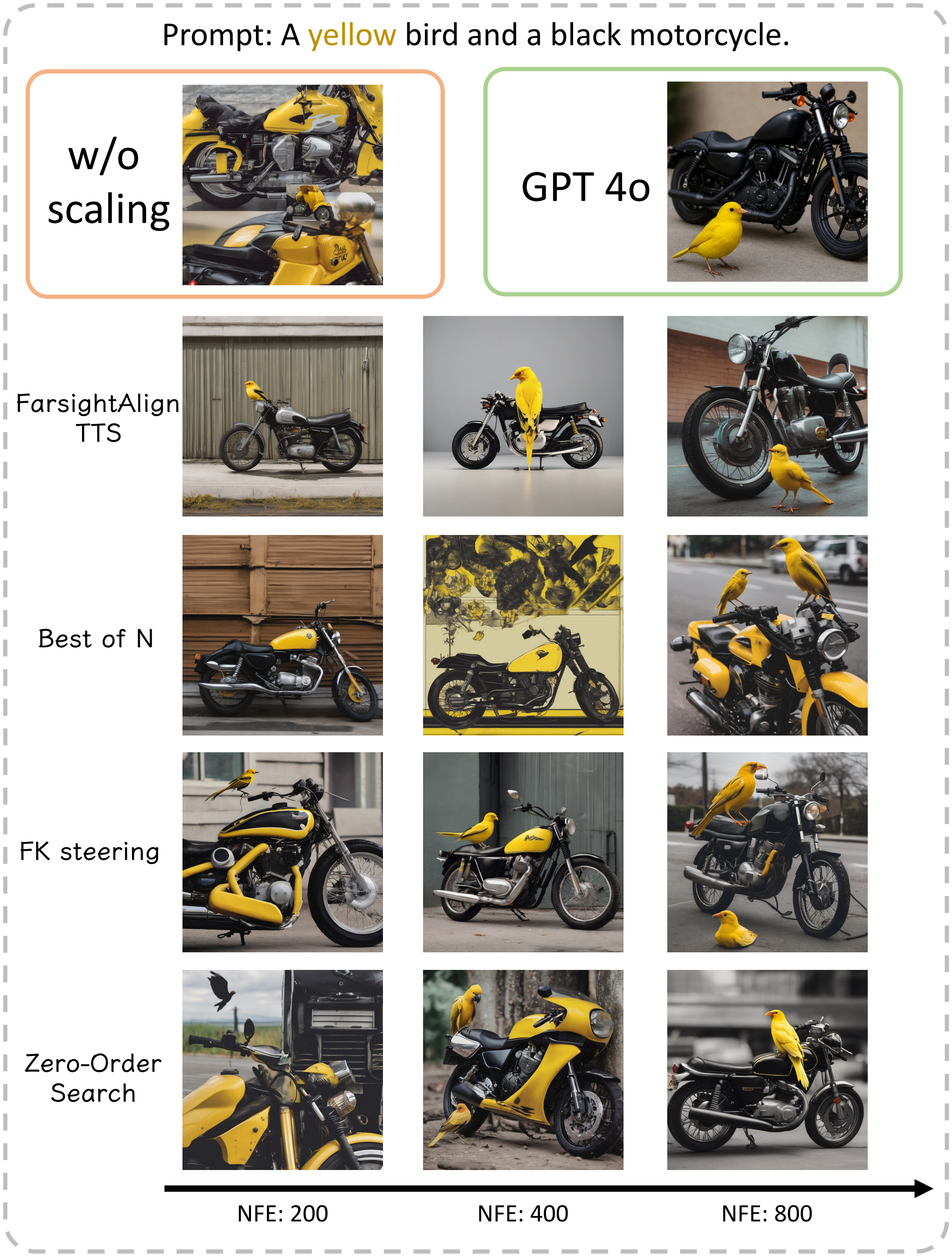


Figure 14: Qualitative comparison on the Geneval benchmark.

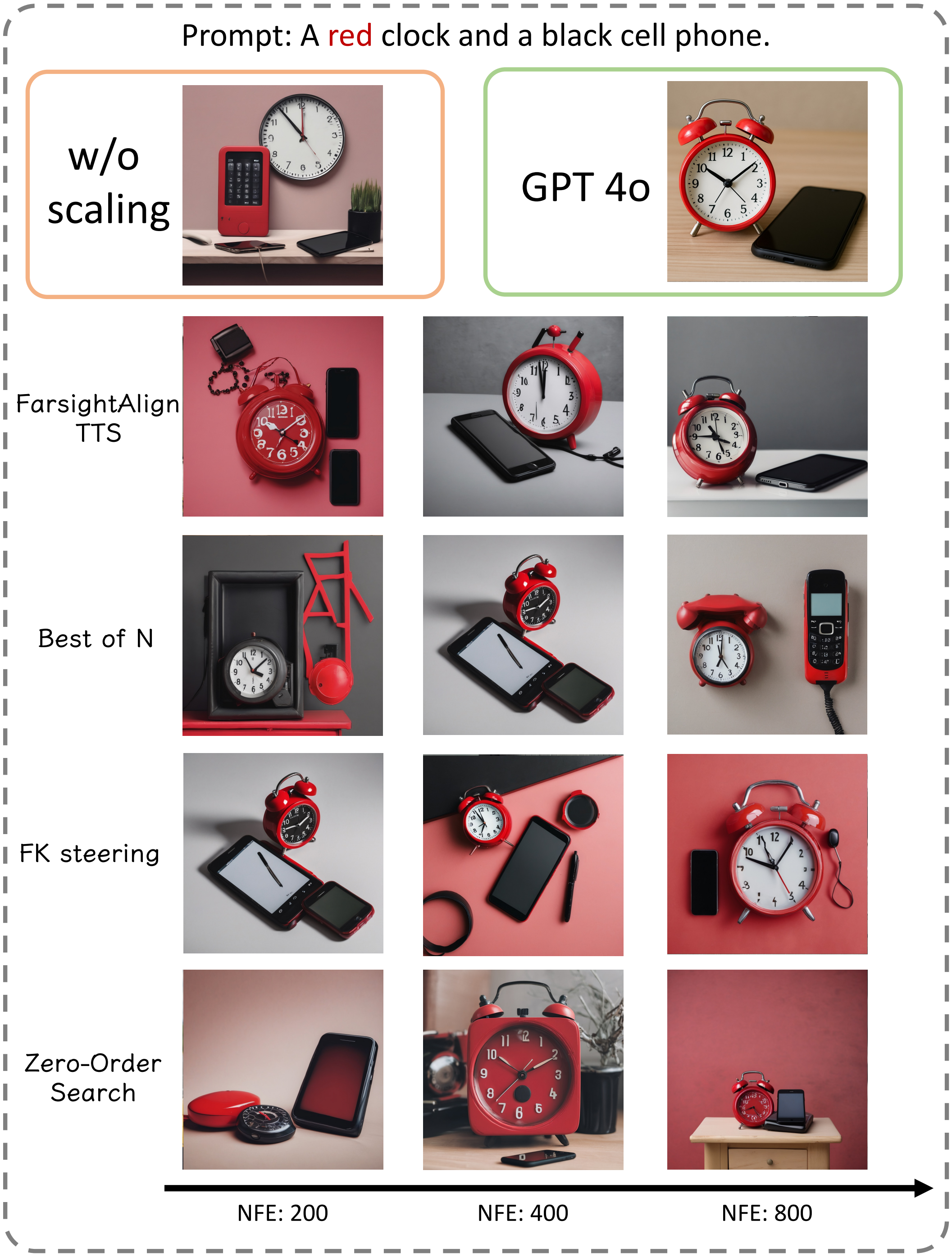


Figure 15: Qualitative comparison on the Geneval benchmark.

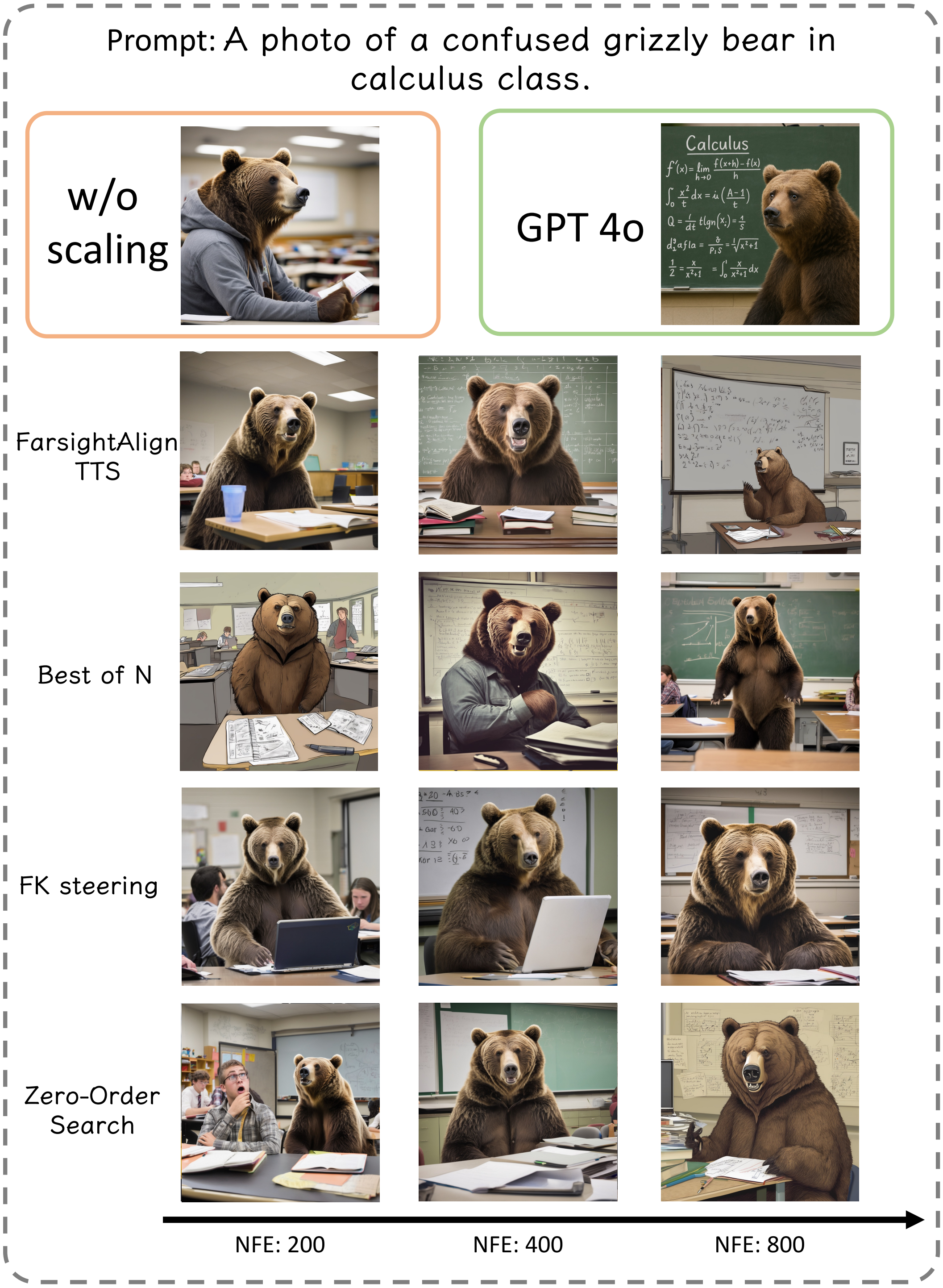


Figure 16: Qualitative comparison on the DrawBench benchmark.

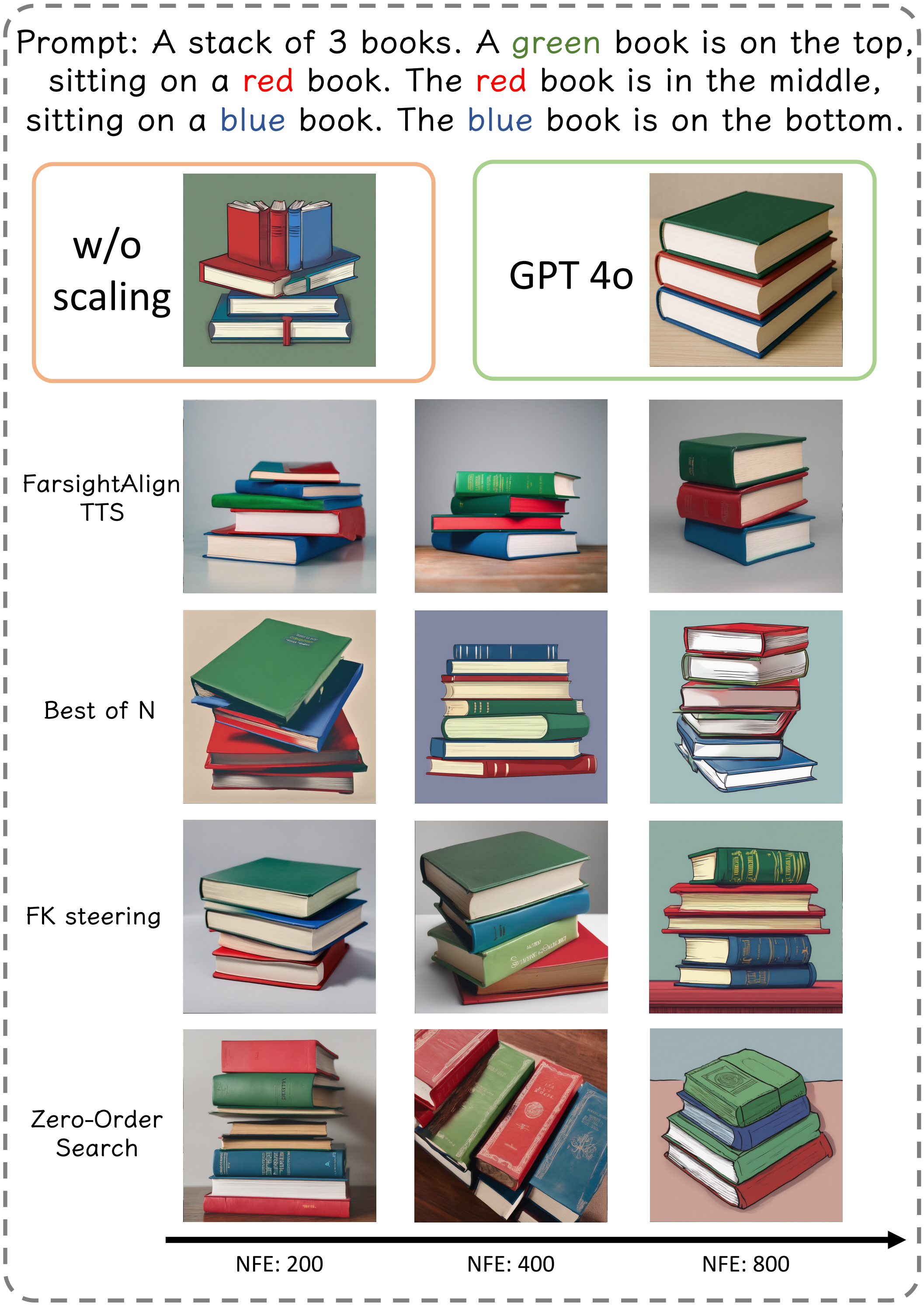

Prompt: A stack of 3 books. A green book is on the top, sitting on a red book. The red book is in the middle, sitting on a blue book. The blue book is on the bottom.
w/o scaling
GPT 4o
FarsightAlign TTS
Best of N
FK steering
Zero-Order Search
NFE: 200
NFE: 400
NFE: 800

Figure 17: Qualitative comparison on the DrawBench benchmark.