PREVIEWDIFF: MULTIMODAL CRITIC-GUIDED SEARCH OVER DIFFUSION LATENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models can produce striking images and videos, but they still struggle with the compositional details that make a generation faithful to a prompt, such as object counts, attribute binding, spatial relations, and temporally grounded actions. A common way to improve prompt satisfaction is to spend more compute at test time through Best-of-N sampling, but final-sample selection is fixed. Best-of-N can only choose among completed outputs and cannot repair a promising trajectory before it fails. We introduce **PreviewDiff**, a training-free test-time search method that turns diffusion sampling from scalar search into a multimodal critic-guided search over intermediate latents. At selected denoising checkpoints, PreviewDiff decodes a partial preview, asks a multimodal judge to score and critique it, and uses the resulting natural-language feedback to branch over semantic prompt edits and locally re-noised latent continuations. These branches are scored, and then pruned and selectively rolled forward, allowing verifier compute to guide generation while the sample is still editable. Across image and video generation benchmarks, PreviewDiff consistently improves over budget-matched Best-of-N selection and strong scalar-search baselines. Ablations show that earlier interventions and increased search width provide the largest gains, while deeper search and additional semantic variants offer complementary improvements. PreviewDiff demonstrates that multimodal feedback is most useful not only as a final verifier, but as an active controller inside the denoising process.

1 INTRODUCTION

Recent text-to-image and text-to-video models have achieved remarkable photorealism and broad prompt fidelity [33, 32, 25]. Yet despite scaling across diffusion architectures and modalities, these systems remain notoriously brittle when tasked with fine-grained compositionality [2, 8, 38]. While standard holistic benchmarks like FID or CLIPScore often register outputs as well-aligned [13, 20], targeted compositional evaluations consistently reveal binding errors [16, 21]—such as miscounting instances, swapping attributes between nearby subjects, or misinterpreting relative spatial and temporal arrangements. Ultimately, a single feed-forward pass struggles to capture the iterative, error-correcting adjustments essential for faithful visual synthesis.

A natural way to improve prompt satisfaction is to spend more computation at test time. The simplest strategy is Best-of-N sampling [36]. This procedure involves drawing N independent samples and use a reward model or multimodal judge to select the best final output. Best-of-N is easy to implement and is therefore an important baseline for test-time scaling. However, it is fundamentally open-loop. It only evaluates completed samples, cannot intervene when a partially denoised sample is promising but flawed, and discards the computation spent on generations that fail late in the process. In contrast, the denoising trajectory itself contains intermediate visual evidence about whether the sample is moving toward or away from the prompt, as illustrated by the decoded checkpoint preview in Figure 1. This suggests a different use of test-time compute. Rather than spending all verifier calls on final outputs, we can use some of them earlier to diagnose and steer partial generations.

Recent work suggests that test-time feedback can improve visual generation. Self-correcting LLM-controlled Diffusion [42] closes the loop by detecting errors in generated images and applying LLM-directed latent correction operations, but this work is organized as an iterative correction loop around completed images. Other methods formulate inference-time scaling as search over sampling noises,

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

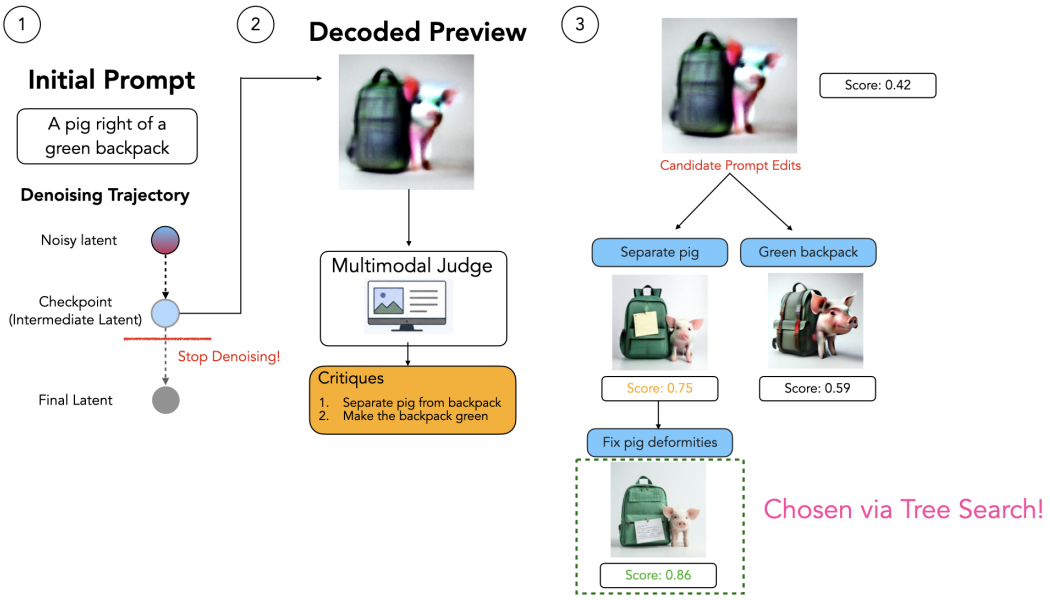


Figure 1: **PreviewDiff searches over critic-guided latent continuations.** (1) From an initial prompt, an intermediate denoising latent is decoded into a preview. (2) A multimodal judge critiques the preview and proposes semantic prompt edits. (3) PreviewDiff branches over these edits, scores the resulting continuations, and selects the best rollout through beam-pruned tree search.

noise trajectories, particles, diffusion trees, or evolutionary mutations [31, 34, 19, 35, 23, 12]. Closest to our video setting, Video-T1 [30] frames video generation as verifier-guided test-time search and introduces tree methods to adaptively expand and prune autoregressive frame continuations. These methods show that visual generation benefits from inference-time search. Yet they generally use the verifier as a reward oracle for selecting candidates. PreviewDiff is complementary. We use a multimodal judge not only to score partial generations, but also to produce interpretable semantic actions that steer diffusion itself.

We propose **PreviewDiff**, a training-free test-time search method for improving compositional prompt following in pretrained image and video diffusion models. PreviewDiff treats intermediate denoised latent estimates as search states. At selected denoising checkpoints, a multimodal model inspects a decoded preview, estimates semantic prompt satisfaction, and produces targeted natural-language feedback about missing objects, incorrect attributes, wrong counts, violated spatial relations, or temporal inconsistencies. This feedback defines semantically meaningful search actions through alternative correction prompts, local re-noising restarts, and latent continuations. The resulting branches are scored, pruned, and selectively rolled forward, turning a single open-loop denoising trajectory into a closed-loop search process guided by multimodal critique.

Our central empirical question is whether multimodal feedback is more useful when applied during generation than when reserved for final selection. We compare PreviewDiff against budget-matched Best-of-N baselines in which the same number of multimodal-judge calls is used to select among independently generated final outputs. We also compare against scalar-search baselines such as EvoSearch [12] and particle sampling inspired by Feynman–Kac and sequential Monte Carlo steering [35], and include Video-T1 [30] as a video test-time scaling baseline. We evaluate scaling along several axes: semantic branching, search width, total verifier budget, checkpoint timing, tree depth, and base model scale. We show a detailed overview in Figure 1.

In summary, this paper makes the following contributions:

1. We introduce PreviewDiff, a training-free framework that casts image and video diffusion generation as multimodal, semantically guided tree search over intermediate denoising states.
2. We use a multimodal model not only to score partial generations, but also to propose natural-language corrective actions that define search branches.

3. We compare guided intermediate search against budget-matched Best-of-N selection, scalar trajectory-search baselines, and video-specific test-time scaling methods.
4. We demonstrate robustness across image and video generation tasks, multimodal verifiers, and downstream diffusion backbones.

2 Related Work

Test-time scaling for diffusion: Search during diffusion sampling is often guided by a scalar verifier score [31, 34]. That score determines which partial trajectories receive more compute in evolutionary and particle-based methods [12, 35, 23]. Video-T1 uses verifier feedback to expand and prune video frame continuations [30]. In PreviewDiff, the critic also proposes the prompt corrections used to create new latent continuations.

Language feedback for diffusion: Visual-feedback methods also use generated images to identify prompt violations. Self-correcting LLM-controlled Diffusion applies LLM-directed corrections after inspecting completed images [42]. Prompt refinement methods use this feedback to revise the conditioning for subsequent generations [41, 22, 4]. We apply corrections during denoising and evaluate the resulting continuations before completing the sample.

Appendix B gives an extended discussion, including prior work on verifier-based selection and compositional evaluation.

3 Methods

PreviewDiff is a training-free search procedure that uses multimodal feedback during diffusion sampling. At selected checkpoints, a critic inspects a decoded preview and proposes corrections. The method branches over these corrections and local latent restarts, retaining promising continuations for further denoising.

3.1 Preliminaries

Let p be the user prompt and G_θ a pretrained diffusion or flow model. Denoising proceeds from high-noise time T toward clean data at $t = 0$. We write a sampler transition as

$$z_s = F_\theta(z_t, (c^+, c^-), t \to s; \epsilon), \qquad s < t, \tag{1}$$

where c^+ and c^- are positive and optional negative conditioning, and ϵ is optional sampler noise. Interventions follow a checkpoint schedule $\mathcal{C} = (\tau_0, \ldots, \tau_D)$ with $T \geq \tau_0 > \cdots > \tau_D \geq 0$. At each checkpoint, the sampler provides a one-step clean-latent estimate $\hat{x}_{0,t}$. The critic observes its decoded preview $y_t = D(\hat{x}_{0,t})$, where D is the image or video decoder. Classifier-free guidance and sampler-specific clean-estimate formulas appear in Appendix C.

3.2 The PreviewDiff Framework

A node n stores the state $s_n = (\hat{x}_0^n, t_n, c_n^+, c_n^-, h_n)$, where $\hat{x}_0^n$ is the clean-latent estimate at checkpoint t_n and h_n records previous feedback. The multimodal critic $\mathcal{J}$ scores the preview against the original prompt,

$$R(n) = \mathcal{J}(D(\hat{x}_0^n), p), \tag{2}$$

and proposes short corrections to observed prompt violations. Search uses a simpler prompt-faithfulness rubric than final evaluation. The critic appends local fix notes rather than rewriting the scene specification, preserving the original prompt while targeting the next continuation [22].

A branch action $a = (\Delta c^+, \Delta c^-, r, \xi)$ combines a positive fix note with an optional negative avoid note, a restart depth r, and restart noise ξ generated from a branch seed. Child conditioning is

$$c_{n'}^\pm = c_0^\pm \oplus \mathrm{recent}(h_n, \Delta c^\pm), \tag{3}$$

where $c_0^\pm$ are the original conditionings and $\oplus$ denotes concatenation. The function recent retains the latest relevant correction notes from the branch history and the proposed action. Negative updates are omitted when unsupported. Prompt templates and the note-retention rule appear in Appendix D.

To create a child, PreviewDiff re-noises $\hat{x}_0^n$ to time $\rho = \mathrm{Restart}(t_n, r)$, obtained by moving r sampler steps toward higher noise, with $t_n < \rho \leq T$. The restart operator is

$$Q_\rho(\hat{x}_0^n; \xi) = \begin{cases} \sqrt{\bar{\alpha}_\rho}\,\hat{x}_0^n + \sqrt{1-\bar{\alpha}_\rho}\,\xi, & \text{DDIM-style,} \\ (1-\sigma_\rho)\hat{x}_0^n + \sigma_\rho \xi, & \text{flow matching,} \end{cases} \tag{4}$$

where $\bar{\alpha}_\rho$ and σ_ρ are scheduler noise parameters. The child then denoises under the corrected conditioning to the next checkpoint $t' < t_n$:

$$z_{t'} = F_\theta\big(Q_\rho(\hat{x}_0^n; \xi), (c_{n'}^+, c_{n'}^-), \rho \to t'\big)\,. \tag{5}$$

Local restarts aim to preserve partial scene structure while allowing the correction to affect generation. We restrict restart depths and decode only at checkpoints with visible semantic structure, since early clean estimates and aggressive restarts can be unreliable. Children are scored and pruned before receiving further compute. Checkpoint ablations examine the tradeoff between preview reliability and remaining editability.

3.3 Beam-pruned tree search over latent continuations

PreviewDiff adapts Monte Carlo tree search with a beam-pruned policy rather than visit-count selection. For each frontier node, the critic proposes M semantic variants. These are crossed with a finite set of restart depths R and L latent branch seeds:

$$\mathcal{A}(s_n) = \{(\Delta c_m^+, \Delta c_m^-, r, \xi_b) : 1 \leq m \leq M,\ r \in R,\ 1 \leq b \leq L\}. \tag{6}$$

An optional no-op branch adds no new correction. The finite action set bounds the cost of decoding and critic evaluation at each expansion.

Root seeds are denoised to the first checkpoint and scored, and the best B initialize the frontier. At each depth, every frontier node is expanded into corrected continuations. Children are denoised only to the next checkpoint, decoded, and scored against p. The next frontier retains the best B children across all parents.

Values are propagated through the tree by max backup,

$$V(n) = \max\left(R(n), \max_{m \in \mathrm{children}(n)} V(m)\right), \tag{7}$$

with $V(n) = R(n)$ for unexpanded leaves. Each action is expanded once before pruning, so the new frontier is selected directly from the scored children. At the final depth, surviving trajectories are completed and scored again, and the highest-scoring final output is returned. Algorithm 1 summarizes the procedure.

4 Experiment Setup

Models. We use SDXL for image scaling because it exposes intermediate clean-latent estimates and supports local restarts. Video scaling uses LTX-Video-2.3 (9B) [10]. Fixed-budget experiments use stronger backbones, with SD-3.5 and FLUX.2 for image transfer and Wan 2.2 for video [40]. All methods share base-model weights within each comparison, without finetuning.

Datasets. GenEval2 is our main image scaling benchmark, testing object counts, colors, attributes, relations, and multi-object constraints [21]. T2I-CompBench++ and T2I-CoReBench extend fixed-budget evaluation to broader compositional and reasoning-oriented prompts [17, 27]. Video scaling uses NarrLV for narrative and temporally structured prompt satisfaction [7]. T2V-CompBench and VBench 2.0 provide complementary coverage of text-video alignment, motion and action binding, temporal consistency, intrinsic faithfulness, and visual quality [38, 45].

Multimodal evaluation. We use Gemini 3.1 Flash-Lite (`gemini-3.1-flash-lite`) [9] as our primary evaluator. It scores generated media against the target prompt using four binary criteria: object and count correctness, attribute binding, relation or action correctness, and overall semantic faithfulness. We report mean scores (0–4) and exact-score-4 pass rates where appropriate. Unless stated otherwise, all search methods and baselines share the Gemini model and rubric. All prompts appear in Appendix D.

Algorithm 1 PreviewDiff

Require: Prompt p, generator G_θ, decoder D, critic $\mathcal{J}$, checkpoints $\mathcal{C}$, beam width B, root seeds S, semantic variants M, restart depths R, latent branch factor L

Denoise roots S to τ_0, decode and score their previews.
Initialize $\mathcal{F}$ with the best B roots.
for checkpoint transition $\tau_d \to \tau_{d+1}$ **do**
    $\mathcal{U} \leftarrow \emptyset$
    for node $n \in \mathcal{F}$ **do**
        Ask $\mathcal{J}$ to score $D(\hat{x}_0^n)$ and propose M corrections.
        for action $(\Delta c_m^+, \Delta c_m^-, r, \xi_b) \in \mathcal{A}(s_n)$ **do**
            Form child conditioning from the original prompt and recent notes.
            $\rho \leftarrow \mathrm{Restart}(t_n, r)$; $\tilde{z}_\rho \leftarrow Q_\rho(\hat{x}_0^n; \xi_b)$.
            Denoise $\tilde{z}_\rho$ to τ_{d+1} under the child conditioning.
            Decode and score the child preview; add the child to $\mathcal{U}$.
        end for
    end for
    Back up child values; retain the best B children in $\mathcal{F}$.
end for
Complete and score surviving trajectories; return the best final output.

Auxiliary evaluation. Where available, we test transfer beyond the search critic using BLIP VQA over disentangled object and attribute questions for T2I-CompBench++ [26] and Qwen3-VL-8B-Instruct on selected image and video runs [1]. These auxiliary evaluators do not guide the main Gemini search unless explicitly stated.

Human evaluation. Nine human raters choose the best output from budget-matched PreviewDiff, Best-of-N, and EvoSearch, separately for content (prompt adherence) and style (aesthetics). We evaluate 360 GenEval2 prompts with SDXL and 200 NarrLV prompts with LTX-Video, collecting nine independent ratings per prompt and criterion. Outputs are shown in random order with method names hidden. We report three-way preferences and PreviewDiff's pairwise win rates against each baseline. Three-way error bars use per-sample confidence intervals bootstrapped across raters and averaged over ratings. Recruitment criteria, instructions, and the interface appear in Appendix E.

Baselines. The *base model* produces one output from the original prompt without generation-time verifier calls, measuring unscaled prompt following. For *Best-of-N* [36], the critic selects among N independent completed outputs without intermediate intervention. Verifier-call matching sets N equal to PreviewDiff's critic-call count. Forward-pass matching chooses N to approximately match PreviewDiff's denoising step-equivalents.

EvoSearch [12] rolls latent particles out to completion at scheduled checkpoints and scores them with the shared Gemini rubric. Elites are preserved and selected parents produce offspring through denoising-aware latent mutations. The critic supplies scalar rewards without semantic corrections. Our *particle-sampling* baseline [12], inspired by Feynman–Kac and sequential Monte Carlo steering, resamples latent particles at intermediate checkpoints using a softmax over critic scores. It isolates scalar reward-based filtering without parent mutation or semantic correction. For *Video-T1* [30], we compare against reported verifier-guided video search settings where available, including Tree-of-Frames-style adaptive expansion and pruning. Video-T1 branches over video candidates rather than semantic prompt edits to intermediate diffusion latents.

Scaling and ablations. Unless stated otherwise, scaling uses SDXL on GenEval2 and LTX-Video on NarrLV. We measure compute in denoiser steps and verifier calls rather than wall-clock latency, since Best-of-N permits parallel generation and scoring while PreviewDiff has sequential dependencies across search depths. Larger budgets increase N for Best-of-N and combinations of root seeds, beam width, semantic variants, latent branch factor, and tree depth for PreviewDiff. Ablations vary one axis at a time. Checkpoint timing controls when previews are decoded and judged, while tree depth sets the number of sequential judge-and-branch rounds. Semantic branching controls the number of proposed fixes per expansion, and search width controls root trajectories and retained beam states. We also evaluate checkpoint timing more densely with two-checkpoint schedules and test increasing base-model scale.

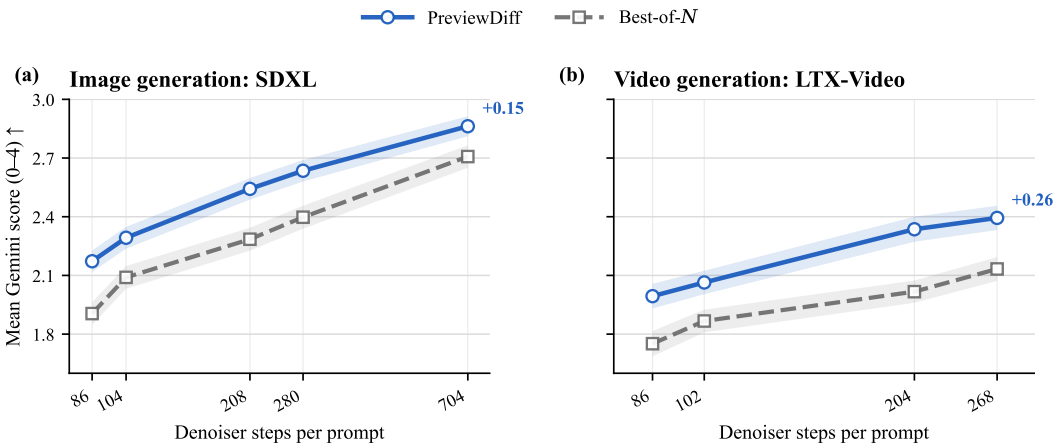


Figure 2: **PreviewDiff as a method for test-time scaling.** PreviewDiff improves with additional denoiser steps per prompt and outperforms budget-matched Best-of-N selection for both image generation with SDXL and video generation with LTX-Video.

Table 1: Image generation results across image backbones and compositional benchmarks.

Model	Method	GenEval2			T2I-CoReBench	T2I-CompBench++	
		Gemini P@4	Qwen-3 Val	Qwen-3 P@4	Gemini P@4	Gemini P@4	BLIP VQA
SDXL	PreviewDiff	**2.292**	**1.9765**	**1.975**	**2.015**	**3.474**	**0.4844**
	Best-of-N	2.090	1.5523	1.735	1.713	2.963	0.3976
	EvoSearch	2.103	–	1.6196	1.7543	3.013	0.4521
	Particle-Sample	1.902	–	1.7544	1.9154	3.003	0.4255
SD-3.5	PreviewDiff	**1.965**	**1.588**	**1.892**	**1.765**	**2.011**	**0.2566**
	Best-of-N	1.754	1.3854	1.713	1.545	1.913	0.2011
	EvoSearch	1.865	–	1.812	1.712	1.975	0.1921
	Particle-Sample	1.723	–	1.809	1.701	1.865	0.2155
FLUX.2 9B	PreviewDiff	**3.112**	**2.675**	**1.983**	**2.781**	**3.758**	**0.7851**
	Best-of-N	2.775	2.504	1.875	2.674	3.510	0.7256
	EvoSearch	2.913	–	1.923	2.752	3.601	0.7445
	Particle-Sample	2.853	–	1.920	2.730	3.456	0.7098

5 Results

We analyze PreviewDiff across all all image and video settings. Our multimodal critic models consist of Gemini-3.1-Flash-Lite and Qwen-3 when scoring denoised checkpoints. We also use Gemini-3.1-Flash-Lite and Qwen-3 as part of our evaluation, scoring the generated images/videos. We analyze these choices among others below.

5.1 PreviewDiff improves test-time scaling

Figure 2 shows the main test-time scaling result. Across both image and video generation, PreviewDiff improves as the denoiser-step budget increases and remains above budget-matched Best-of-N at every tested budget. On SDXL, PreviewDiff improves from 2.17 to 2.86 mean Gemini score as compute increases, while Best-of-N rises from 1.91 to 2.71. On LTX-Video, PreviewDiff improves from 1.99 to 2.39, while Best-of-N improves from 1.75 to 2.13. The gap persists at both small and large budgets, which suggests that intermediate feedback is not only a low-budget shortcut. It remains useful even when Best-of-N is allowed to draw many complete samples.

This result supports the central premise of PreviewDiff. Final-output selection can only choose among completed samples. PreviewDiff can diagnose a partial generation, branch over targeted

Table 2: Video generation results across video backbones and text-to-video benchmarks. Gray rows at the bottom show published benchmark reference results using each benchmark's official metric. These rows are included for context and are not compute-matched to our methods.

Model	Method	T2V-CompBench			NarrLV		V-Bench 2.0
		Gemini P@4	Qwen3-VL P@4	LLaVA Eval	Gemini P@4	Qwen-2.5 Eval	Gemini P@4
LTX-Video 9B	MCTS-CGD	**3.021**	**1.987**	**0.4523**	**2.167**	**62.17**	**3.162**
	Best-of-N	2.716	1.411	0.3922	1.544	55.92	3.123
	EvoSearch	2.814	1.901	0.3713	2.012	61.14	3.071
	Particle-Sample	2.755	1.926	0.3656	1.975	58.96	3.112
	Video-T1	2.861	1.954	0.4329	2.115	61.87	3.124
Wan 2.2 A14B	MCTS-CGD	**3.541**	**2.019**	**0.6311**	**3.411**	**71.22**	**3.554**
	Best-of-N	3.362	1.875	0.6011	2.874	68.16	3.292
	EvoSearch	3.392	1.978	0.6126	3.012	68.75	3.399
	Particle-Sample	3.386	1.944	0.5928	2.961	68.11	3.287
	Video-T1	3.465	1.901	0.6266	3.229	70.02	3.509

corrections, and allocate later computation to states that are already moving toward the prompt. The same verifier is therefore used more effectively when it guides the denoising process rather than only ranking the final outputs.

5.2 Fixed-budget results across models and benchmarks

Tables 1 and 2 evaluate PreviewDiff under fixed budgets across image and video backbones. We include evaluations with both open- and closed-source multimodal LLMs like Gemini-3.1-Flash-Lite and Qwen-3 as well as include basic evaluations from a particular benchmark like evaluating with BLIP-VQA for T2I-CompBench++. On image generation, PreviewDiff improves over Best-of-N on most reported metrics across SDXL, SD-3.5, and FLUX.2. The gains are clearest on GenEval2, where the task stresses the same count, attribute, and relation errors that the multimodal critic is designed to identify.

PreviewDiff also improves auxiliary evaluator scores in the current table, including Qwen-3 (P@4) and BLIP-VQA columns, which suggests that the method is not merely overfitting to the Gemini critic. We also evaluate using Qwen-3 as a critic in PreviewDiff and Best-of-N rather than as only an evaluator (Qwen-3 Val). We find significant improvement in these settings as well. This indicates that we can use Qwen-3 to provide consistent feedback. We note that EvoSearch and Particle Sampling do not use critic feedback from a multimodal judge so these are intentionally left blank.

On video generation, PreviewDiff improves over Best-of-N on LTX-Video and Wan 2.2 across the main Gemini evaluations. The current LTX-Video results show especially large gains on T2V-CompBench and NarrLV, where temporal relations and prompt satisfaction are difficult to recover through final selection alone. Wan 2.2 shows the same pattern at a stronger base-model scale. We also find that using corresponding open-source models like Qwen-3, Qwen-2.5 or LLaVA as an evaluator achieves similar improvement, meaning we don't overfit to Gemini results.

We find PreviewDiff outperforms other test-time scaling baselines like EvoSearch, Particle Sampling, and Video-T1. This indicates that using denoised trajectories as part of our evaluation is important to our improvements. Similarly, having an MLLM in the loop allows us to find better trajectories than what is proposed in EvoSearch.

5.3 Qualitative behavior

Figure 3 provides a qualitative analysis of the mechanisms used in PreviewDiff. For our generations, the most common successful corrections involve missing objects, wrong counts, wrong colors, and spatial relation errors. These examples show both the decoded preview and the branch selected by the critic, so we can see that PreviewDiff is not simply sampling more outputs. It is using intermediate visual evidence to steer the continuation.

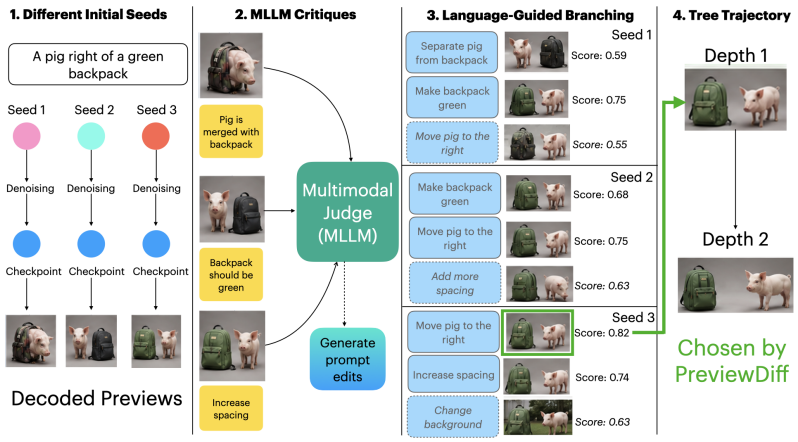


Figure 3: **PreviewDiff generation walkthrough.** Decoded previews from three seeds receive targeted prompt corrections. Candidate continuations are scored for prompt adherence on a 0–1 scale. The rightmost column shows one selected trajectory.

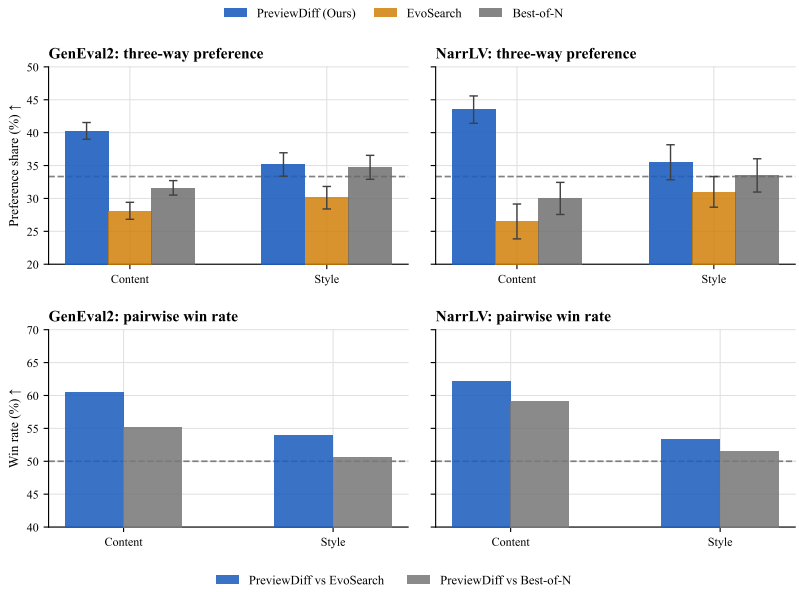


Figure 4: **Human evaluation on GenEval2 and NarrLV.** PreviewDiff's gains over EvoSearch and Best-of-N are larger for content than style.

We can observe that many corrections in PreviewDiff involve change object order, introducing spacing to aid visual style, or correcting physically implausible generations. From our analysis, language guidance is often correctly able to identify visual problems in the generated image as branching occurs. For example, an initial generation can be physically implausible e.g., the prompt requests a pig to the right of a green backpack and the resulting image merges and pig and backpack together. In the example, we find that MLLM, Gemini-3.1-Flash can identify these visual issues. Color issues are also common mistakes in accurately showing content. MLLM judges are also sensitive about spacing between objects, which is the most common stylistic complaint.

5.4 Human Evaluation

To further establish our findings with MLLM evaluators, we include a human evaluation where we compare PreviewDiff against EvoSearch and Best-of-N, rating based on style and content. We show results in Figure 4. We see that in all cases, PreviewDiff beats the other baselines including EvoSearch and Best-of-N. We find that PreviewDiff has the highest three way win rate compared to both methods, performing significantly above chance. Similarly, in head-to-head comparisons, PreviewDiff wins significantly more than EvoSearch and Best-of-N.

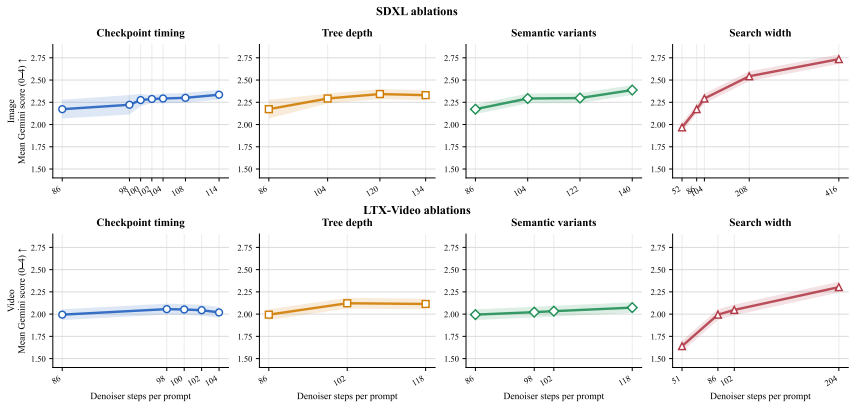


Figure 5: **Scaling across different search axes.** We ablate four axes of PreviewDiff for image generation with SDXL and video generation with LTX-Video, plotting mean Gemini score against total denoiser steps per prompt. *Checkpoint timing* varies the denoising step at which the first intermediate preview is decoded and judged. We vary our checkpoints from 2 checkpoints to 8 checkpoints. *Tree depth* varies the number of sequential judge-and-branch rounds. We vary from a depth of 1 to a depth of 3. *Semantic variants* varies the number of candidate prompt edits proposed by the multimodal judge at each expansion. We sweep from 1 variant to 4 variants. *Search width* varies the number of trajectories retained and explored. We sweep the width from 1 branch to 8 branches.

When doing further analysis, we find that we improve significantly with content over style. This finding is intuitive. The human results show larger gains for content than for style. This is expected because the PreviewDiff critic is prompted to focus on prompt adherence, including objects, attributes, counts, relations, and actions. The search therefore directly optimizes semantic correctness more than aesthetic quality. Style improvements are still possible when better semantic corrections also improve visual coherence, but they are not the primary target of the current critic. SDXL and LTX-Video have weaker performance style-wise. Improvements over style will likely happen with more robust underlying models.

5.5 Ablations on search axes

Figure 5 isolates the main axes of PreviewDiff. Search width produces the largest gains for both images and videos. On SDXL, increasing width raises the mean score from 1.97 to 2.74. On LTX-Video, width raises the score from 1.64 to 2.30. This is expected because width increases both the diversity of root trajectories and the number of partial states that survive pruning. It gives the critic more chances to find a promising continuation before the sample is finalized.

Depth and semantic variants provide smaller but useful gains. Increasing depth allows the method to correct a generation more than once, although returns saturate once the search reaches later checkpoints. Increasing semantic variants gives the critic more possible fixes to try at each node, but the benefit depends on whether the preview error is specific enough for the critic to propose distinct corrections. Checkpoint timing is also important. Earlier checkpoints leave more room for the corrected conditioning to affect the sample, while later checkpoints provide clearer previews but less opportunity to change the final output.

6 Conclusion

We introduced PreviewDiff, a training-free test-time search method that uses multimodal feedback inside the denoising process rather than only after generation is complete. By decoding intermediate latent previews, asking a multimodal judge to critique them, and branching over semantic prompt edits with local latent continuations, PreviewDiff turns visual generation into an adaptive closed-loop search. Across image and video benchmarks, this strategy improves over final-sample selection and scalar-search baselines under matched compute budgets. The results suggest that multimodal judges are most useful when they act not only as verifiers, but as semantic controllers that help decide where generation should go next.

AI USE STATEMENT

AI was used for paper revising and literature review collected for this paper. All final drafting of this work has been reviewed and edited by a human author. All references have been verified by a human.

AI was also used for figure editing and initial figure drafting. AI code was used to generate initial drafts of plotting code. All outputs and numerical scores were verified by a human author before inclusion in the paper.

REPRODUCIBILITY STATEMENT

Specific details employed in this work have been included to aid reproducibility such as compute used for PreviewDiff, the exact prompts employed for the diffusion model and LLM-evaluator, and focus on open-source datasets/evaluations. We also plan to include code to reproduce results upon acceptance of the paper as well as make assets of example videos available for analysis downstream. The exact algorithm used in this paper is described in detail and can be easily written in code as well.

Our paper includes closed- and open-source models. However, our pipeline has been evaluated and verified on a open-source pipeline thoroughly including open-source image- and video-generation models with Qwen-3 as the judge. We believe this will strongly aid with reproducibility.

REFERENCES

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[2] 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 Transactions on Graphics*, 42(4):1–10, 2023.

[3] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

[4] Xinyan Chen, Jiaxin Ge, Tianjun Zhang, Jiaming Liu, and Shanghang Zhang. Learning from mistakes: Iterative prompt relabeling for text-to-image diffusion model training. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2937–2952, 2024.

[5] Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for step-by-step text-to-image generation and evaluation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[6] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, in Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *International Conference on Learning Representations*, 2024.

[7] Xinyi Feng, Haonan Yu, et al. NarrLV: Towards a comprehensive narrative-centric evaluation for long video generation models. *arXiv preprint arXiv:2507.11245*, 2025.

[8] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 52132–52152, 2023.

[9] Google DeepMind. Gemini 3.1 Flash-Lite Model Card. `https://deepmind.google/models/model-cards/gemini-3-1-flash-lite/`, 2026.

[10] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch,

Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

[11] Mariam Hassan, Bastien Van Delft, Wuyang Li, and Alexandre Alahi. Factorized video generation: Decoupling scene construction and temporal synthesis in text-to-video diffusion models. *arXiv preprint arXiv:2512.16371*, 2025.

[12] 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. *arXiv preprint arXiv:2505.17618*, 2025.

[13] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[16] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20406–20417, 2023.

[17] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3563–3579, 2025.

[18] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[19] Vineet Jain, Kusha Sareen, Mohammad Pedramfar, and Siamak Ravanbakhsh. Diffusion tree sampling: Scalable inference-time alignment of diffusion models. *arXiv preprint arXiv:2506.20701*, 2025.

[20] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9307–9315, 2024.

[21] Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. Geneval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025.

[22] Mohammed Abdul Hafeez Khan, Yash Jain, Siddhartha Bhattacharyya, and Vibhav Vineet. Test-time prompt refinement for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6506–6516, 2025.

[23] Sunwoo Kim, Minkyu Kim, and Dongmin Park. Test-time alignment of diffusion models without reward over-optimization. *arXiv preprint arXiv:2501.05803*, 2025.

[24] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

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

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

[27] Ouxiang Li, Yuan Wang, Xinting Hu, Huijuan Huang, Rui Chen, Jiarong Ou, Xin Tao, Pengfei Wan, Xiaojuan Qi, and Fuli Feng. Easier painting than thinking: Can text-to-image models set the stage, but not direct the play? *arXiv preprint arXiv:2509.03516*, 2025.

[28] Ouxiang Li, Yuan Wang, Xinting Hu, Huijuan Huang, Rui Chen, Jiarong Ou, Xin Tao, Pengfei Wan, Xiaojuan Qi, and Fuli Feng. Easier painting than thinking: Can text-to-image models set the stage, but not direct the play? In *The Fourteenth International Conference on Learning Representations*, 2026.

[29] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, 2024.

[30] Fangfu Liu, Hanyang Wang, Yimo Cai, Kaiyan Zhang, Xiaohang Zhan, and Yueqi Duan. Video-t1: Test-time scaling for video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

[31] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025.

[32] OpenAI. Video generation models as world simulators. OpenAI technical report, 2024.

[33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

[34] Vignav Ramesh and Morteza Mardani. Test-time scaling of diffusion models via noise trajectory search. *arXiv preprint arXiv:2506.03164*, 2025.

[35] 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. *arXiv preprint arXiv:2501.06848*, 2025.

[36] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[38] Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[39] Rui Tian, Mingfei Gao, Mingze Xu, Jiaming Hu, Jiasen Lu, Zuxuan Wu, Yinfei Yang, and Afshin Dehghan. Unigen: Enhanced training & test-time strategies for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.14682*, 2025.

[40] Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[41] Shiyu Wu, Mingzhen Sun, Weining Wang, Yequan Wang, and Jing Liu. Visualprompter: Semantic-aware prompt optimization with visual feedback for text-to-image synthesis. In *International Conference on Learning Representations*, 2026.

[42] Tsung-Han Wu, Long Lian, Joseph E Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting llm-controlled diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6327–6336, 2024.

[43] 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. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

[44] Jaesik Yoon, Hyeonseo Cho, Doojin Baek, Yoshua Bengio, and Sungjin Ahn. Monte carlo tree diffusion for system 2 planning. *arXiv preprint arXiv:2502.07202*, 2025.

[45] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

A Limitations

PreviewDiff depends on the critic receiving a useful intermediate preview. Early previews can be unreliable, while later checkpoints leave less time for corrections to affect the sample. Even with feedback, the method still misses exact counts and can skip required actions in videos. Our compute comparisons measure denoiser steps and verifier calls, so they do not establish a wall-clock advantage over Best-of-N. Each search depth depends on the previous round of feedback, whereas Best-of-N can generate candidates in parallel. The critic focuses on prompt adherence, and our human evaluation shows smaller gains in visual style.

B Extended Related Work

Inference-time search and steering for diffusion models. A growing line of work studies how to spend additional computation during diffusion sampling rather than only increasing the number of denoising steps. Inference-time scaling frameworks search over initial noises or sampling trajectories using verifiers and search algorithms [31, 34]. EvoSearch formulates image and video generation as evolutionary search over denoising trajectories [12]. Tree- and particle-based methods push this idea further: Monte Carlo Tree Diffusion expands and prunes partially denoised trajectories for planning [44], Diffusion Tree Sampling reuses computation by propagating terminal rewards through a diffusion tree [19], and Feynman–Kac steering and Diffusion Alignment as Sampling use particle resampling or sequential Monte Carlo to sample from reward-aligned distributions [35, 23]. Diffusion Forcing is also related in its use of flexible intermediate uncertainty and guided sequence rollouts [3]. PreviewDiff shares the adaptive inference-time view, but uses the multimodal model as more than a scalar reward. It scores partial previews and proposes semantic correction directions that become branches over latent continuations. Video-T1 [30] is the closest video-generation prior to our work. It frames test-time scaling for video generation as search over trajectories from Gaussian noise to videos, using test-time verifiers and heuristic search algorithms to guide candidate selection. Its Tree-of-Frames method adaptively expands and prunes autoregressive frame continuations, with verifier feedback used to balance quality and compute. PreviewDiff shares the goal of adaptive inference-time allocation, but differs in the search state and action space. Our states are intermediate denoised latent estimates, and our branches are critic-proposed semantic prompt corrections combined with local latent re-noising and continued denoising. Thus the multimodal model is not only a verifier for pruning candidates, but also a policy-like source of semantic actions.

Visual feedback, prompt refinement, and semantic correction. Several systems use generated images as feedback for improving prompt adherence. Self-correcting LLM-controlled Diffusion closes a loop around completed images by detecting prompt violations and applying LLM-controlled latent correction operations [42]. VisualPrompter identifies absent concepts in generated images and rewrites prompts at an atomic semantic level [41]. Test-time Prompt Refinement uses an MLLM to detect missing objects or incorrect attributes and generate refined prompts for subsequent generations [22]. Iterative Prompt Relabeling uses repeated image sampling and prompt relabeling to improve text-image alignment [4]. Attention- and program-based approaches such as Attend-and-Excite and VPGen/VPEval also address compositional failures including object neglect, counting, and spatial layout [2, 5]. PreviewDiff is closest to the feedback-driven methods, but moves the feedback inside denoising. Instead of restarting a full generation after a final-image critique, it decodes intermediate clean-latent estimates, branches over critic-proposed prompt edits plus local latent restarts, and prunes before the sample is complete.

Verifier-based selection and compositional evaluation. Best-of-N selection is a strong baseline for test-time scaling because it can improve outputs using only a final-stage verifier or reward model. Human-preference rewards such as ImageReward and PickScore provide ranking signals for text-to-image candidates [43, 24]. VQAScore shows that VQA-style alignment scores can improve generation by selecting among multiple candidates [29], and UniGen uses chain-of-thought verification for Best-of-N test-time generation in a unified multimodal model [39]. At the same time, compositional evaluation work shows why scalar final scores are often insufficient. GenEval and GenEval 2 focus on object-centric counts, colors, positions, and benchmark drift [8, 21]. TIFA, DSG, and VQAScore decompose prompt faithfulness into question-answering or graph-structured checks [16, 6, 29]. T2I-CompBench++ and T2I-CoReBench stress attribute binding, spatial relations, numeracy, complex composition, and reasoning [18, 28]. PreviewDiff uses these verifiers in a different role. Rather than only ranking finished outputs, the same multimodal signal diagnoses partial failures and turns them into interpretable search actions.

C Extended Preliminaries

Let p denote the user prompt and let G_θ be a conditional diffusion or flow model operating on latent variables. We use the standard diffusion convention in which $t = 0$ denotes clean data and larger t denotes higher noise. Reverse denoising therefore moves from a high-noise time t toward a lower-noise time $s < t$. We write a sampler transition as

$$z_s = F_\theta(z_t, c, t \to s; \epsilon), \qquad s < t, \tag{8}$$

where F_θ is the transition map induced by the pretrained model, numerical sampler, scheduler, conditioning c, and optional sampler noise ϵ. We denote the intervention checkpoints by

$$\mathcal{C} = (\tau_0, \tau_1, \ldots, \tau_D), \qquad T \geq \tau_0 > \tau_1 > \cdots > \tau_D \geq 0,$$

where each τ_d is a diffusion time and d is only the index of the checkpoint in the search schedule. In classifier-free guidance [14], we represent the conditioning as a positive prompt c^+ and a negative prompt c^-. Let m_θ denote the model prediction, which is a noise prediction for DDPM/DDIM-style samplers and a velocity prediction for flow-matching samplers. The guided prediction is

$$m_\theta^{\text{cfg}}(z_t, t, c^+, c^-) = m_\theta(z_t, t, c^-) + w\big(m_\theta(z_t, t, c^+) - m_\theta(z_t, t, c^-)\big), \tag{9}$$

where w is the guidance scale. When a model does not use a negative prompt, we set $c^- = \emptyset$.

At a checkpoint time t, the sampler exposes a clean-latent estimate $\hat{x}_{0,t}$. We assume this latent estimate to be the standard one-step clean estimate implied by the current latent and model prediction. For DDPM/DDIM-style samplers [15, 37] with cumulative noise level $\bar{\alpha}_t$, this estimate is

$$\hat{x}_{0,t} = \frac{z_t - \sqrt{1 - \bar{\alpha}_t}\, \epsilon_\theta^{\text{cfg}}(z_t, t, c^+, c^-)}{\sqrt{\bar{\alpha}_t}}. \tag{10}$$

For flow-matching samplers with noise level σ_t and velocity prediction v_θ^{cfg}, we use

$$\hat{x}_{0,t} = z_t - \sigma_t v_\theta^{\text{cfg}}(z_t, t, c^+, c^-). \tag{11}$$

The preview shown to the critic is then

$$y_t = D(\hat{x}_{0,t}), \tag{12}$$

where D is the image or video decoder.

D PreviewDiff Prompting

We describe prompts below. Specific prompts for our multimodal judge are included in Figure 6.

Diffusion Model Prompt The prompt fed to the image or video diffusion model is consist across different base generators. Given the original prompt p and a list of recent critic fix notes $\Delta c_1, \ldots, \Delta c_m$, the positive conditioning prompt is

$$p^+ = p \oplus \texttt{``FIX NOTES: ''} \oplus \Delta c_{m-1} \oplus \Delta c_m.$$

When negative prompting is available, we append recent critique fragments to the negative prompt using the prefix ``Avoid:''.

Multimodal Prompt Editor

You are guiding a diffusion model from an intermediate latent preview.

The final target prompt is:

{target_prompt} The current working prompt is: {current_prompt}

Known failure context:

{context} Look at the {media_type} preview. Propose {k} diverse, targeted correction branches. Do NOT rewrite the full prompt. Return short addenda that can be appended as FIX NOTES. Focus on missing objects, wrong counts, wrong attributes, incorrect binding, spatial/action/relation mistakes, and temporal/action mistakes for video.

Return ONLY valid JSON: {"variants": [{"emphasis": "short label", "fix_addendum": "short concrete correction", "critique": "what is wrong"}]}

Multimodal Judge Final Scoring

You are scoring a generated {media_type} for prompt satisfaction.

Return ONLY valid JSON.

Target prompt: {target_prompt} Score exactly these four binary criteria, each 0 or 1:

1. object/count correctness 2. attribute/color/material binding correctness 3. spatial/action/relation correctness 4. overall semantic faithfulness and absence of major contradictions

Return schema: {"scores": [{"criterion": "...", "score": 0 or 1, "reason": "short"}], "overall_score": integer from 0 to 4, "fail_reasons": ["short reasons"]}

Figure 6: **Prompts used in PreviewDiff**. We include prompts used for our multimodal judge in PreviewDiff. The top prompt in orange shows the prompt edits applied by a multimodal judge over the tree search. Similarly, final scoring happens using a 0-4 scale, shown in the bottom in blue.

Final evaluator prompt. PreviewDiff uses the same multimodal judge for final image and videos. The judge receives the target prompt and the final output. It returns a JSON object with four binary criteria: object/count correctness, attribute/color/material binding, spatial/action/relation correctness, and overall semantic faithfulness. The sum gives a score from 0 to 4.

Search-time prompt. PreviewDiff uses the a multimodal judge for previewed images and videos. The judge receives the target prompt and the final output. It returns score between 0 and 1, scoring the image on faithfulness to the prompt. This is done so we don't overfit to the final evaluator prompt. The prompt follows the final evaluator prompt but does not use a JSON with the four requested evaluation axes.

Language-feedback prompt. For branch expansion, the judge is also asked to propose short correction addenda. These addenda are not full prompt rewrites. They are appended to the original user prompt under a `FIX NOTES` field, while critique fragments are optionally appended to the negative prompt using an `Avoid:` prefix. This keeps the original prompt fixed while allowing the next denoising continuation to focus on the specific error observed in the intermediate preview.

E HUMAN EVALUATION

To fully verify the improvement of PreviewDiff, we include a full human evaluation using nine human raters. Each human rater was asked to evaluate the image/video outputs of three methods for a single prompt. There were 360 image prompts, taken from GenEval2, and 200 video prompts, taken from NarrLV.

To recruit human raters, we chose adults over the age of 18. All raters were unfamiliar with the nature of the work nor were the goals of the evaluation made clear to the human raters. Raters were requested to evaluate the generations for each prompt on style and content. The instruction stated: `For each prompt, choose the best anonymous option separately for Style (visual quality/aesthetics) and Content (prompt correctness, object correctness over background correctness, consistency, etc.).`

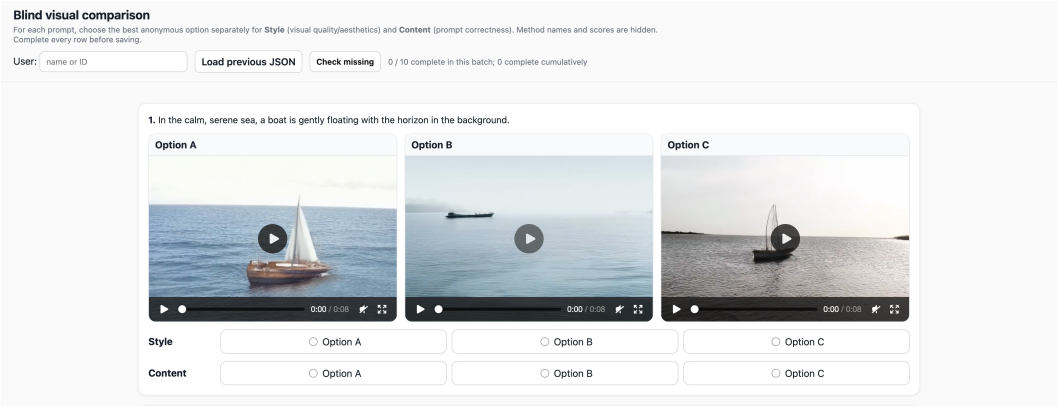


Figure 7: **Human Evaluation Page**. We provide an example of our human evaluation webpage, shown to participants for rating output images and videos. Our evaluation consists of two axes, style and content. Raters are asked to choose between the three videos based on both axes.

A screen capture of our human evaluation UI is included in Figure 7. The UI demonstrates the labels are hidden from raters. Furthermore, raters were asked to choose for both style and content.

For the purpose of this evaluation, we did not include ties. We did so to get accurate human ratings given that style comparisons often very similar. PreviewDiff wasn't prompted to improve style ratings so many times, raters would choose ties. This complicated the ratings.

F Visual Examples

We show additional image examples in Figure 8 and video examples in Figure 9. These qualitative examples highlight a consistent pattern: Best-of-N often selects a visually plausible sample, but the selected output may still miss a key prompt constraint. PreviewDiff instead uses the intermediate preview to identify the specific failure and steer the continuation toward a more faithful sample. For images, the improvements are most visible on compositional errors. PreviewDiff more often recovers the correct number of objects, fixes spatial relations such as objects being under, in front of, or to the left of another object, and improves attribute binding for colors and materials such as glass, stone, or wood. This is especially important on GenEval2, where many prompts are simple in language but difficult for open-loop sampling because the model must satisfy several object-level constraints at once.

For videos, the same mechanism improves prompt adherence over time. Best-of-N can produce videos with appealing motion or appearance, but it frequently leaves failures in scene color, lighting, object identity, or the intended action relation. PreviewDiff often corrects these issues by using the critic's language feedback to preserve promising structure while revising the parts of the trajectory that conflict with the prompt. The video examples show improvements in lighting and colorization, clearer depiction of the requested scene, and better alignment between the main subject and the described action or environment. Together, Figures 8 and 9 illustrate that PreviewDiff is not simply selecting more attractive samples. It is using intermediate visual evidence to make targeted corrections, which leads to outputs that are both visually coherent and more faithful to the prompt.

G Additional Analyses

G.1 Checkpoint timing and model scale

Figure 10 studies two complementary questions. First, it asks when the critic should intervene. Early two-checkpoint schedules consistently outperform middle and late schedules on SDXL. This indicates that intermediate previews do not need to be perfect to be useful. They only need to reveal enough semantic structure for the critic to identify a plausible correction. Intervening too late gives

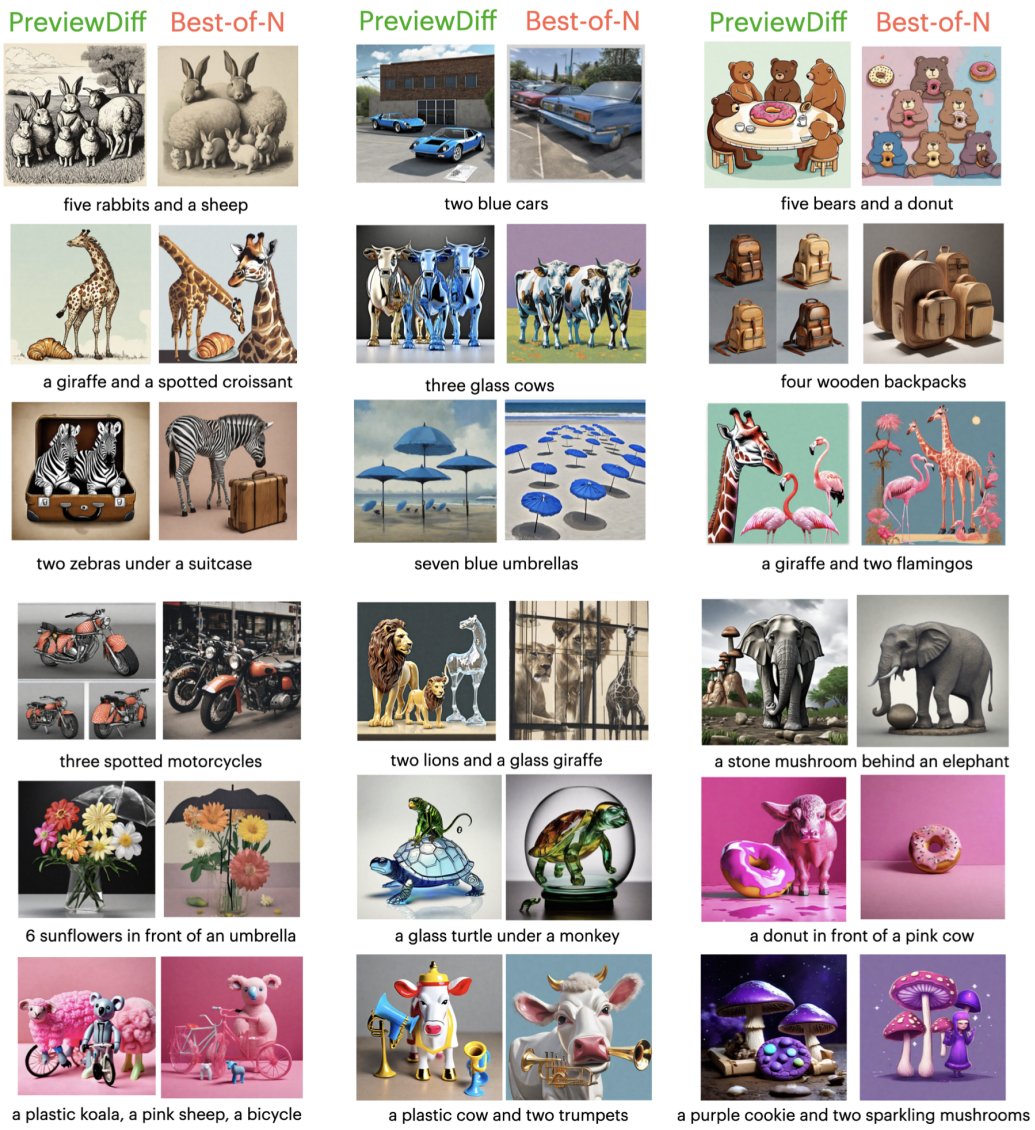


Figure 8: **Image Examples from PreviewDiff vs Best-of-N.** We include 18 text-to-image prompts comparing PreviewDiff vs Best-of-N. The base model is SDXL and the prompts are sampled from GenEval2 and T2I-CoReBench. Across all examples, using PreviewDiff leads to improved visual components with respect to style and content. For example, we correct numerical counts (five rabbits, two blue cars) and include corrections with texture (three glass cows). While numerical mistakes still occur in PreviewDiff, the errors are less severe than what is found in Best-of-N (eight umbrellas vs ¿ten).



Figure 9: **Video Examples from PreviewDiff vs Best-of-N.** We include 6 text-to-video prompts comparing PreviewDiff vs Best-of-N. The base model is LTX-Video and the prompts are sampled from NarrLV. Across all examples, using PreviewDiff leads to improved visual components with respect to style and content. For example, we correct colorization (orange glow, brightly lit) and include corrections with lighting (shines under the bright lights). While mistakes still occur in PreviewDiff, the errors are less severe than what is found in Best-of-N (e.g. harbor area is not bustling).

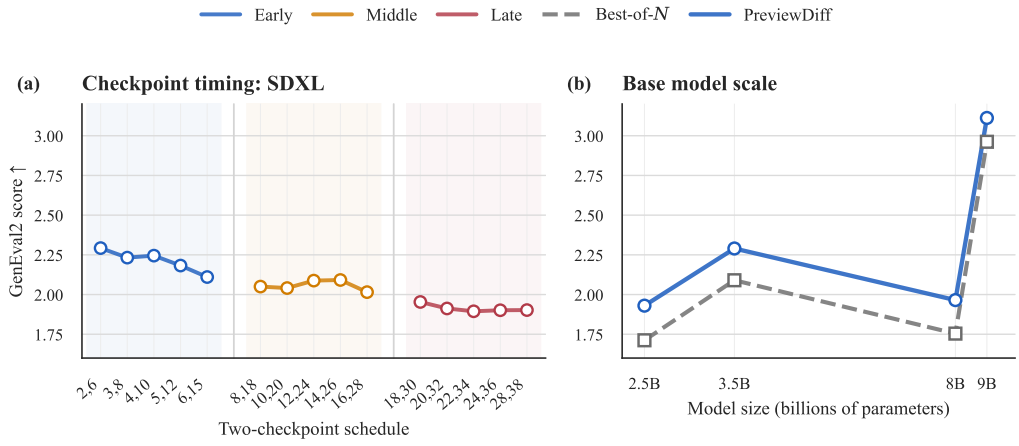


Figure 10: **Checkpoint timing and base-model scaling. (a)** Effect of checkpoint timing on PreviewDiff for SDXL on GenEval2. Each point corresponds to a two-checkpoint schedule (c_1, c_2) indicating the denoising steps at which intermediate previews are decoded and judged. **(b)** Scaling with base model size, comparing budget-matched Best-of-N selection against PreviewDiff at a fixed test-time budget.

Table 3: Component ablations for PreviewDiff. Image results use SDXL on GenEval2, and video results use LTX-Video on NarrLV. Initial-root ablations change the default two-root setting.

Experiment	Image	Video
	SDXL on GenEval2 Gemini P@4	LTX-Video on NarrLV Gemini P@4
PreviewDiff	**2.292**	**2.167**
w/o language feedback	2.063	2.014
w/o prompt editing	2.285	2.125
initial roots $2 \rightarrow 1$	2.290	2.156
initial roots $2 \rightarrow 3$	2.310	2.178

the judge a clearer preview, but leaves too little denoising time for the correction to reshape the sample.

Second, Figure 10 asks whether PreviewDiff is only useful for weaker base models. Across the tested model scales, PreviewDiff improves over fixed-budget Best-of-N. The gains remain present at the largest model size, where the base model already has stronger prompt-following ability. This suggests that critic-guided search is complementary to model scaling. Larger models produce better trajectories, but intermediate feedback still helps allocate test-time compute to the most promising continuations.

G.2 Additional ablations

In Table 3, we ablate specific components of PreviewDiff. PreviewDiff can be modified to stop using language feedback as part of the generation process, so the multimodal judge only provides a numerical score. Language feedback provides useful guidance for ensuring generation quality across the tree search using intermediate denoised checkpoints. Similarly, we test whether prompt editing via the feedback from the multimodal judge improves results. Similar to findings in Figure 10, prompt editing to add or change the prompt doesn't have very strong effects on the final generation quality but does cause an improvement. Finally, the number of initial sampled roots can affect the final generation. We could start with multiple noised latents as the roots of our tree in MCTS. We find that this has little effect on our final result. Using 1, 2, or 3 nodes doesn't have a huge effect on our final result in comparison to branching.

H Failures Cases

Although PreviewDiff improves prompt following across image and video generation, it does not eliminate all compositional failures. Figure 11 shows representative failure modes. For images, the remaining errors often involve constraints that are discrete, relational, or highly specific. PreviewDiff may recover the right object categories while still missing exact numeracy, as in the prompt requiring four clocks and five spotted turtles. It can also invert spatial relations, such as depicting sheep under umbrellas rather than umbrellas under a sheep. Attribute binding remains imperfect as well: in the blue-giraffe example, the method preserves the requested layout but fails to bind the blue color to the giraffe.

For videos, failures are most common when the prompt requires a precise temporal sequence rather than a static scene. PreviewDiff can generate a plausible goal area but still omit the goalkeeper and the downpour event. It can also skip intermediate actions, such as showing a candle already lit rather than depicting the full sequence of striking a match, lighting the candle, and blowing it out. Finally, temporal color ordering remains difficult. In the crystal example, the object changes appearance, but not in the requested transparent-to-blue-to-green-to-yellow-to-orange-to-purple sequence. These cases suggest that PreviewDiff is strongest when semantic errors can be corrected by local prompt-guided continuation, but weaker when success requires exact counting, relation grounding, or long-horizon temporal state tracking.



four clocks and five spotted turtles



five metal umbrellas under a sheep



two green suitcases to the left of a blue giraffe



On a sunny afternoon, the football match goal area is lively, with the goalkeeper standing alert. Dark clouds begin to gather, casting shadows across the field. Moments later, a torrential downpour drenches the entire scene, turning the vibrant field into a slippery, challenging terrain.



In a dimly lit ancient room, there is a candlestick with an unlit candle on a table, and an elderly man stands beside the table holding a matchstick. Subsequently, the old man strikes the match. Then, he lights the candle on the candlestick with the matchstick. Finally, he blows out the candle.

Figure 11: **PreviewDiff weaknesses and failures**. We show examples where PreviewDiff does not generate the correct result particularly with cases of numeracy, spatial relations, and attribute binding.

Table 4: Published benchmark reference scores.

Benchmark	Model	Metric	Score
T2I-CompBench++ [17]	DaLLE-3	BLIP-VQA	0.6966
T2V-CompBench [11]	Veo-3	LLaVA-Eval	0.6844
T2V-CompBench [11]	Wan2.2	LLaVA-Eval	0.5742
NarrLV [7]	Wan 2.2	Qwen-2.5 Eval	70.6

I Published Benchmark References

Table 4 collects published scores under the original benchmark metrics. These results provide context for Tables 1 and 2. Compute is not matched to our experiments.