EXTENDED ABSTRACT: Scaling Test-Time Compute via Semantic Critique and Spectral Alignment for Visual Media Generation

Jia Xian Huang
National Yang Ming Chiao Tung University
tarohuang.ii14@nycu.edu.tw

Figure 1. **Scaling Test-Time Compute with CritiFusion.** By iteratively scaling compute during inference, our framework synthesizes high-fidelity visual media with dynamic preference optimization, significantly improving prompt alignment without retraining the foundation model.

Abstract

Recent generative models have achieved remarkable visual fidelity, yet faithfully aligning generated content with complex human semantics remains challenging. While scaling training compute has plateaued in yielding structural alignment, scaling test-time compute emerges as a strong alternative. We introduce CritiFusion, an efficient inference-time scaling framework that integrates multimodal foundation models as agentic critics. The CritiCore module scales test-time reasoning to produce high-level semantic feedback, dynamically guiding the denoising process. To prevent test-time gradient updates from corrupting global geometry, we propose SpecFusion, which merges generation states in the spectral domain. This preserves low-frequency layout constraints while injecting high-frequency semantic corrections. Preliminary results demonstrate that scaling our Multi-LLM agent ensemble monotonically improves human-aligned metrics, offering a highly robust paradigm for training-free test-time preference optimization.

1. Introduction

The paradigm of computer vision is shifting toward test-time scaling. While large-scale training of visual diffusion models [1, 3, 5] provides strong generative priors, optimizing these models for complex human preferences typically requires resource-intensive fine-tuning [4, 7]. In contrast, recent breakthroughs in reasoning models suggest that scaling compute *during inference*—by allowing models to critique and refine their outputs—can unlock significant performance gains without altering pre-trained weights.

Current training-free control methods often rely on simple scalar rewards or static heuristic schedules, which lack

the deep semantic reasoning needed to correct complex attribute binding or relational errors. Furthermore, naive test-time optimization on video or image latents frequently disrupts fragile spatiotemporal coherence and low-level visual structures.

To address this, we propose **CritiFusion**. It features a multimodal multi-agent critique system (**CritiCore**) and a Fast Fourier Transform (FFT) based stabilization mechanism (**SpecFusion**). By treating VLMs and LLMs as test-time reasoning agents, we dynamically allocate inference compute to regions requiring semantic correction. Simultaneously, SpecFusion ensures the trustworthiness of the scaled vision model by preserving the original low-frequency layout.

2. Related Work

Test-Time Optimization and Scaling. Preference alignment for diffusion models typically includes differentiable reward fine-tuning or RL-style objectives optimized against human-derived signals [4, 7]. However, training-free approaches have begun exploring how test-time interventions can close the alignment gap. By scaling inference compute—such as dynamically reweighting guidance or iteratively refining latents [9]—models can achieve preference optimization without the computational burden of retraining. We extend this by explicitly coupling inference-time compute scaling with structured multimodal feedback.

Agentic Visual Systems. Prompt-level interventions adapt textual conditioning to better match model priors. Frameworks utilizing Large Language Models (LLMs) and Vision-Language Models (VLMs) demonstrate collaborative potential for planning, critique, and revision [6]. Building on these agentic workflows, CritiFusion employs an LLM committee for semantic enrichment and a VLM for image-grounded critique, effectively acting as an autonomous test-time reasoning loop.

3. Test-Time Scaling Methodology

CritiFusion operates entirely at inference time with frozen backbones, trading a modest increase in inference compute for substantial gains in semantic alignment.

3.1. CritiCore: Scaling Semantic Reasoning

Instead of relying on a single generative pass, we scale inference compute by introducing an iterative critique loop. Given an initial base generation $\mathbf{x}^{\text{base}}$ generated from prompt c, a VLM produces evidence-grounded hints regarding missing entities or spatial misplacements. A committee of LLMs (Multi-LLM) acts as an aggregator, mapping these hints into grounded visual clauses $\mathcal{C} = \{\tau_i\}_{i=1}^{m}$.

The VLM then scores each clause $s_i = f_{\text{vlm}}(\tau_i, \mathbf{x}^{\text{base}})$. To optimize inference efficiency, we map the mean score s to a dynamic test-time compute allocation strategy via the Controlled Adaptive Denoising Rate (CADR):

$$(\lambda,\ g,\ T',\ \rho) \ = \ \Phi(s) \ = \ A + (1 - s)\, B \qquad (1)$$

where λ is the corrective strength, g is the CFG scale, T' is the corrective step count, and ρ is the SpecFusion mask parameter. Lower confidence scores trigger more intense test-time refinement, ensuring compute is scaled proportionately to the complexity of the semantic misalignment.

3.2. SpecFusion: Robust Spectral Alignment

Applying agentic test-time updates directly to the latent space often degrades global composition. We introduce SpecFusion to enforce frequency-domain consistency. Let $\mathcal{F}$ denote the 2D FFT. We extract the low-frequency spectrum from the base latent ($\mathbf{Z}_{\text{lo}} = \mathcal{F}(\mathbf{z}^{\text{base}})$) and the high-frequency spectrum from the refined latent ($\mathbf{Z}_{\text{hi}} = \mathcal{F}(\mathbf{z}^{\text{ref}})$).

We fuse these using a confidence-controlled mask $K(\rho)$, where the passband is dynamically scaled based on the test-time critique:

$$\mathbf{Z}_{\text{fuse}} = \underbrace{K(\rho) \odot \mathbf{Z}_{\text{lo}}}_{\text{global layout}} + \underbrace{\big(1 - K(\rho)\big) \odot \mathbf{Z}_{\text{hi}}}_{\text{fine details}} \qquad (2)$$

The final latent is obtained via the inverse FFT: $\tilde{\mathbf{z}} = \mathcal{F}^{-1}(\mathbf{Z}_{\text{fuse}})$. This spectral safeguard ensures that our test-time scaling improves semantics without compromising structural trustworthiness.

4. Experiments and Evaluations

4.1. Implementation Details

All experiments run on a single NVIDIA A6000 GPU using SDXL [3] as the primary diffusion backbone. Inference runs for 50 base sampling steps. We employ a VLM-augmented, multi-agent text stack accessed via API. The VLM is instantiated as Llama-4-Scout-17B, while the Multi-LLM committee comprises an ensemble of Qwen, DeepSeek, and Llama 3.3 variants. The extra cost includes one partial denoising pass and two FFTs per sample. Empirically, the runtime overhead is strictly $< 1.3\times$ compared to the baseline, ensuring highly efficient algorithm scaling. Evaluations are conducted on public prompt suites using HPSv2 [8], ImageReward [10], and PickScore [2].

4.2. Scaling Compute via LLM Ensemble Size

The core hypothesis of test-time scaling is that increasing inference compute yields better performance. Table 1 demonstrates this by varying the number of LLM agents in the CritiCore committee. We observe a monotonic gain in PickScore from 22.27 (1 LLM) to 23.13 (5 LLMs). This confirms that allocating more reasoning compute during inference directly translates to higher visual preference.

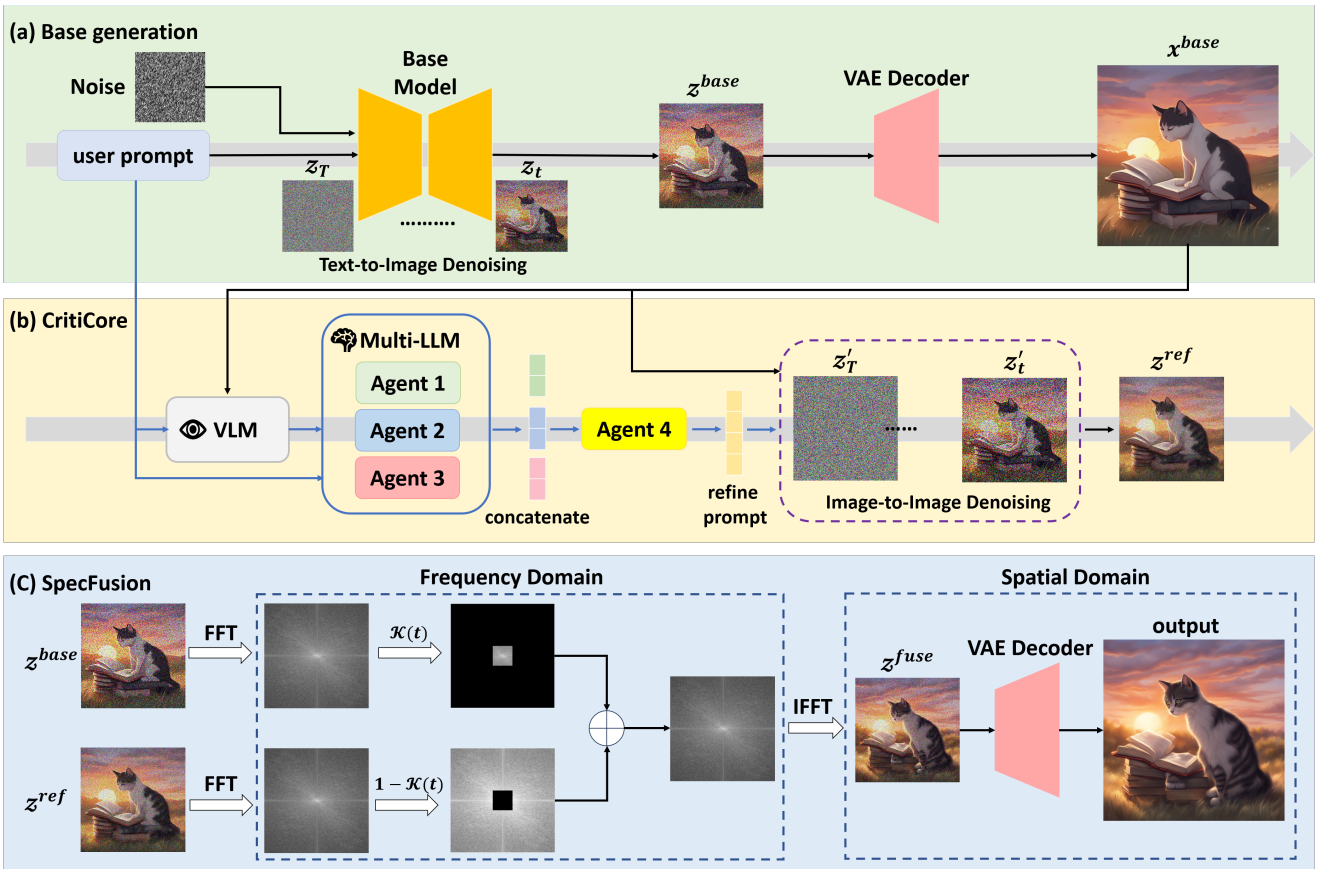

Figure 2. **Overview of our framework.** Our pipeline has two modules. (1) CritiCore fuses VLM captions with Multi-LLM feedback to produce a refined prompt embedding $\tilde{c}$. A diffusion backbone samples a base latent $\mathbf{z}^{\text{base}}$ (decoded as $\mathbf{x}^{\text{base}}$) and, guided by $\tilde{c}$, partially re-denoises it to $\mathbf{z}^{\text{ref}}$. (2) SpecFusion performs frequency gating: the high-frequency spectrum of $\mathbf{z}^{\text{ref}}$ is combined with the low-frequency spectrum of $\mathbf{z}^{\text{base}}$ to yield $\tilde{\mathbf{z}}$.

Table 1. Effect of scaling inference compute (LLM ensemble size) on generation performance.

# LLM Agents	PickScore↑	HPSv2↑	ImageReward↑
1 (Low Compute)	22.27	0.288	0.972
3 (Med Compute)	22.80	0.297	1.010
5 (High Compute)	**23.13**	**0.291**	**1.076**

4.3. Ablation Study and SOTA Comparison

Table 2 reports both the component-wise ablation and the state-of-the-art comparison. Skipping spectral alignment (*w/o SpecFusion*) degrades perceptual quality (ImageReward drops by 0.097), proving that structural safeguards are mandatory when scaling test-time optimization. Furthermore, CritiFusion consistently matches or exceeds computationally heavy training-based methods on human-aligned metrics without parameter updates. By operating solely during the reverse diffusion process, our framework democratizes preference alignment and eliminates the massive GPU overhead typically required by RL-based fine-tuning.

Table 2. Ablation study and SOTA comparison on SDXL.

Variant / Method	PickScore↑	HPSv2↑	ImageReward↑
Ablation Variants			
w/o VLM Critic	22.14	0.278	0.911
w/o SpecFusion	22.31	0.293	0.979
Baselines & SOTA			
SDXL (Baseline) [3]	21.97	0.266	0.776
Diffusion-DPO [7]	22.30	0.274	0.979
DyMO (Test-time) [9]	24.90	0.284	1.074
Ours (Full Framework)	23.13	**0.291**	**1.076**

4.4. Qualitative Analysis

Beyond quantitative metrics, Figure 3 illustrates how scaling test-time compute visually resolves deep semantic ambiguities. Attribute binding is significantly more reliable for complex spatial relations, and object counts are rigorously respected. Crucially, due to SpecFusion's spectral gating, boundaries remain crisp with fewer halo or ringing artifacts typically introduced by naive gradient-based inference updates. Global illumination is maintained con-

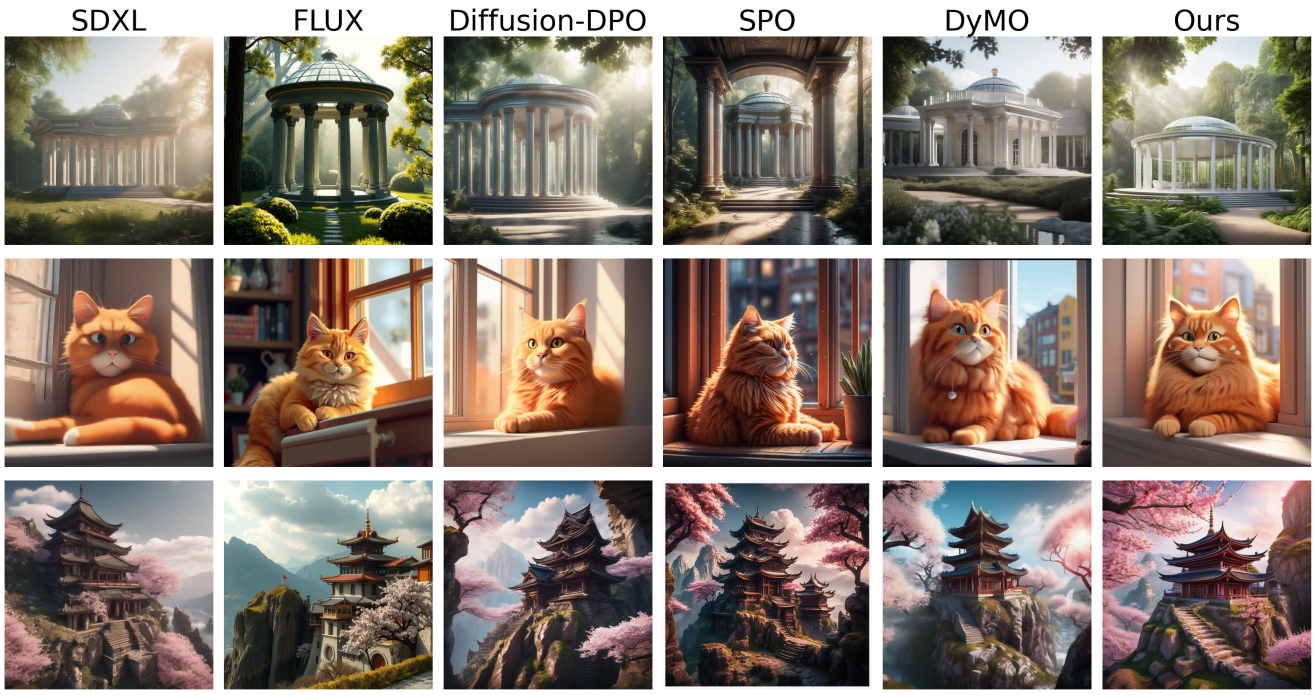

Figure 3. **Qualitative Comparison of Test-Time Scaling.** We compare images generated by SDXL, FLUX, Diffusion-DPO, SPO, DyMO, and our CritiFusion method across various subjects such as *architecture*, *animals*, *natural scenes*, and *characters*. Each group shares the same prompt for fairness. Our results (far right in each group) show superior realism and semantic alignment: natural color tones, spatial coherence, and faithful prompt adherence. Competing methods often suffer from artifacts or miss subtle details, while CritiFusion maintains balanced composition and high-fidelity rendering. Zoom-in reveals texture and object accuracy, illustrating the benefits of semantic critique and spectral fusion.

sistently with the base generation, proving the robustness of frequency-domain fusion in preserving low-level vision elements while editing semantics. Specifically, across diverse subjects like architecture and natural scenes, CritiFusion successfully mitigates the texture degradation and color shifting often observed in competing baselines.

5. Conclusion

We present CritiFusion, demonstrating that scaling test-time compute via multimodal agentic critique and spectral alignment is a highly effective paradigm for visual generation. By decoupling preference optimization from model training, our framework offers a flexible, robust, and computationally scalable path forward. Future work will extend these inference-time scaling and FFT principles to multi-view and video generation domains.

References

[1] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[2] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy. Pick-a-Pic: an open dataset of user preferences for text-to-image generation. In *NeurIPS*, 2023.

[3] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.

[4] M. Prabhudesai, A. Goyal, D. Pathak, and K. Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv*, 2023.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[6] Y. Shen, K. Liu, Z. Zhang, J. Cheng, K. Zhao, D. Zhao, and W. Chen. HuggingGPT: solving AI tasks with ChatGPT and its friends in Hugging Face. *arXiv*, 2023.

[7] B. Wallace, M. Dang, R. Rafailov, L. Zhou, A. Lou, S. Purushwalkam, S. Ermon, C. Xiong, S. Joty, and N. Naik. Diffusion model alignment using direct preference optimization. In *CVPR*, 2024.

[8] X. Wu, Y. Hao, K. Sun, Y. Chen, F. Zhu, R. Zhao, and H. Li. Human preference score v2. *arXiv*, 2023.

[9] X. Xie and D. Gong. DyMO: training-free diffusion model alignment with dynamic multi-objective scheduling. In *CVPR*, 2025.

[10] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong. ImageReward: learning and evaluating human preferences for text-to-image generation. *arXiv*, 2023.