Training-free, Perceptually Consistent Low-Resolution Previews with High-Resolution Image for Efficient Workflows of Diffusion Models

Supplementary Materials

Wongi Jeong*[1] Hoigi Seo*[1] Se Young Chun[1,2†]
[1]Dept. of Electrical and Computer Engineering, [2]INMC & IPAI
Seoul National University, Republic of Korea
{wg7139, seohoiki3215, sychun}@snu.ac.kr

A. Details on compliance

We assume compliance of the downsampled trajectory when performing preview generation. A natural question is whether this compliance property genuinely holds. Zhang et al. [7] provide a proof in the appendix of their work, showing that under an L-Rectified Flow, the deviation between the compliant trajectory and the true trajectory is bounded by $O(1/L)$. Their analysis indicates that, although exact compliance is not guaranteed, approximation becomes reasonable once L is sufficiently large. Building on this result, we consider our compliance assumption to be well justified.

B. Detailed Experimental Setup

B.1. Experiment compute resources

All experiments are run on an NVIDIA A100 GPU, which serves as our primary compute resource. Floating operations (FLOPs) were measured using the `torch.profiler` tool, and the latency results in Tab. 2 are benchmarked directly on the same GPU.

B.2. Evaluation dataset

Commonly used evaluation benchmarks such as MS-COCO contain short prompts and limited scene diversity, making them inadequate for representing the aesthetic and complex prompting environments typical of designers or photographers, which are the primary targets of our study. Instead, we adopt the more challenging PixArt-30K evaluation set to better align with real-world usage. From this dataset, we randomly sample 5,000 prompts whose lengths range from a maximum of 1,885 characters to a minimum of 2 characters, with a median of 319 and a mean of 413 characters. This wide spectrum of prompt lengths and expressiveness allows us to capture the dynamic and realistic prompting scenarios encountered in practical workflows.

B.3. Metrics

To evaluate the performance of our preview generation, we employ a range of image similarity and image quality metrics. This section provides detailed explanations of each metric, including their computation procedures and specific evaluation criteria.

PIQE [5]. PIQE is a no-reference image quality metric that estimates perceptual distortion by analyzing deviations from natural scene statistics at the block level. The image is first transformed into Mean Subtracted Contrast Normalized (MSCN) coefficients, after which each non-overlapping block is classified as either spatially active or uniform based on its MSCN variance. Spatially active blocks are then examined for two types of degradations: noticeable structural distortions, detected via low-variance edge segments, and noise distortions, characterized using a center–surround deviation measure. For each distorted block, PIQE assigns a distortion score derived from the block-level MSCN variance, which is inversely proportional to perceived quality for structural distortions and directly proportional for noise. The final PIQE score aggregates block-level distortions according to

$$\text{PIQE} = \frac{\sum_{k=1}^{N_{\text{SA}}} D_{S_k} + C_1}{N_{\text{SA}} + C_1} \tag{S1}$$

where N_{SA} is the number of spatially active blocks and D_{s_k} is the distortion score of the k-th block. Lower PIQE values indicate higher perceptual quality, making it suitable for evaluating preview fidelity without requiring reference images. We leveraged block size 8, and all other parameters were set to default value.

* Authors contributed equally. † Corresponding author.

DreamSim [1]. DreamSim is a mid-level perceptual similarity metric learned from the NIGHTS dataset, designed to outperform traditional perceptual metrics such as LPIPS that mainly capture low-level or patch-level distortions. By ensembling CLIP, OpenCLIP, and DINO embeddings and fine-tuning them with LoRA, DreamSim aligns more closely with human visual judgments across variations in pose, layout, shape, and semantic content—factors that LPIPS is not sensitive to. For an image pair (x, x'), DreamSim computes similarity using the cosine distance

$$D(x, x') = 1 - \cos(f_\theta(x), f_\theta(x')) \tag{S2}$$

This formulation enables DreamSim to capture perceptual relationships that extend beyond pixel-level fidelity, yielding substantially higher agreement with human preferences compared to LPIPS and other prior perceptual metrics.

DiffSim [4]. DiffSim is a visual similarity metric that leverages the attention features of pretrained diffusion models to overcome the limitations of traditional perceptual similarity measures such as LPIPS, which primarily capture low-level patch statistics and often fail to reflect human judgments of layout, pose, and semantic consistency. DiffSim introduces the Aligned Attention Score (AAS), which aligns the latent representations of two images through the attention mechanism of the Stable Diffusion U-Net and computes a cosine-based similarity on these aligned features. For latent representations L_A and L_B extracted from a chosen attention layer, DiffSim computes

$$\begin{aligned}&\mathrm{AAS}(L_A, L_B) \\ &\quad = \cos(\mathrm{attn}(Q_A, K_B, V_B), \mathrm{attn}(Q_B, K_A, V_A)),\end{aligned} \tag{S3}$$

allowing the metric to evaluate both appearance and style similarity while compensating for spatial misalignment. Comprehensive evaluations across human-aligned, instance-level, style, and low-level similarity benchmarks show that DiffSim consistently surpasses LPIPS, CLIP, and DINO, demonstrating substantially higher agreement with human perceptual judgments.

PSNR. PSNR is a full-reference image similarity metric that measures the fidelity between a generated image and its reference by comparing pixel-level reconstruction error. It is defined using the mean squared error (MSE) between two images x and x', where a higher PSNR indicates closer pixel-wise correspondence. Formally, for images with maximum possible pixel value $I_{\max}$, PSNR is computed as

$$\mathrm{PSNR}(x, x') = 10 \log_{10}\left(\frac{I_{\max}^2}{\mathrm{MSE}(x, x')}\right). \tag{S4}$$

Because PSNR is sensitive to exact pixel alignment, it strongly correlates with low-level distortions such as noise, blur, and compression artifacts but is limited in evaluating perceptual or semantic similarity, making it complementary to metrics such as DreamSim, and DiffSim.

FSIM [6]. FSIM is a full-reference metric designed to capture low-level similarity by focusing on image structures and contrast patterns that are strongly aligned with human visual sensitivity. It relies on two fundamental low-level cues—phase congruency and gradient magnitude—which respectively encode structural information (edges, corners, and local features) and local contrast. Because these cues emphasize fine-grained texture, sharpness, and structural fidelity, FSIM provides reliable measurements of low-level distortions such as blur, noise, compression artifacts, and local geometric degradation. The metric aggregates these cues using a feature-weighted formulation,

$$\mathrm{FSIM}(x, x') = \frac{\sum_p (S_{\mathrm{PC}}(p) \cdot S_{\mathrm{GM}}(p) \cdot W(p))}{\sum_p W(p)}, \tag{S5}$$

and is widely used when accurate evaluation of pixel-level and structural consistency is required.

B.4. Prompts in qualitative results

Due to space limitations in the main paper, we could not include the prompts used for the qualitative results presented in the main script, and therefore provide them here. They are summarized in Tab. S1, where the attached indices follow the left-to-right column order. The first four prompts are used for FLUX.1-dev, and the last three prompts are used for the qualitative results of SD 3.5-L. To demonstrate performance under diverse prompting conditions, we deliberately include prompts ranging from short to long.

C. Additional Experiments

C.1. Effect of the number of fixed-point iterations

In Eq. (10), the update is defined as a fixed-point iteration over k steps. For efficiency, however, we adopt the setting $k = 1$ in our main experiments. To assess the effect of using larger values of k, we additionally evaluate the method with $k > 1$.

As shown in Tab. S2, increasing k yields marginal improvement or degradation in performance, but leads to a noticeable increase in runtime. Therefore, using $k = 1$ provides a favorable choice for better efficiency.

C.2. Preview generation on lower resolution

In the main paper, we only report preview generation at a resolution of 512×512. However, this setting alone is not sufficient to fully assess the effectiveness of the preview generation. Therefore, we additionally conduct experiments at a lower resolution of 256×256.

Table S1. The prompts used for the qualitative results in the main paper are ordered from the left column to the right column.

	Prompt used in main qualitative results
1	pixelated Quantum entanglement Fentanyl digital banality, dishevelled, avant-garde multicultural diaspora humans contorted jumping twisted broken distorted lighting casts high contrast shadows lighting Hasselblad h6d - 400c, carl zeiss batis super close up 16mm f/ 2.8, ricoh r1
2	Small Dragon with Finnish elf characters, in a Finnish forest, in the style of korik kokiri, volumetric lighting, lively tableaus, mori kei, dark gold and green, vray, konica big mini. a figure and a mushroom stand among mushrooms in a forest, in the style of rendered in cinema4d, bill gekas, light gold and green, anime-inspired character designs.
3	Chinese man has invented a machine that can travel in time, High and short depth of field, stippling
4	shadoflectocyberchromos, You got a heart of stone, you can never feel.
5	Black Hole Sun + ultra high quality + beautiful colours + psychedelic + 3D rendered
6	A Stone Island X Cottweiler menswear collaboration editorial campaign shot by Johnny Dufort. Pa Salieu is modelling. Metallic fabrics. Mesh inserts. Two-tone materials. Chameleon fabrics. Color-flip materials. Technical treatments and washes to the materials. Night, Raining.
7	In the image, there are two individuals engrossed in reading a magazine. The person on the left, clad in a black sweater, is holding the magazine with both hands, indicating a deep interest in the content. On the right, another person, wearing a red shirt, is also holding the magazine with both hands, mirroring the actions of the person on the left. They are seated comfortably on a white couch, which stands out against the backdrop of a white wall adorned with a window. The window allows for natural light to filter into the room, illuminating the scene and casting soft shadows. The magazine they are reading is open, suggesting that they are actively engaged in reading it. The pages of the magazine are not visible in the image, but one can infer from their focused expressions that they are reading about something of interest. There is no text visible in the image, and the relative positions of the objects suggest a casual and relaxed atmosphere. The image captures a quiet moment of shared interest and learning between the two individuals.

Table S2. Quantitative comparison across different values of k.

k	TFLOPs↓	Speed↑	PIQE↓	DreamSim↓	PSNR (dB)↑
1	1178.30	1.53×	28.55	6.83	21.182
2	1277.56	1.41×	28.79	6.85	21.132
3	1376.84	1.31×	28.88	6.76	21.132

Fig. S2 presents the qualitative results at a resolution of 256×256. While the reduced resolution inevitably introduces slight blurriness, the generated images remain highly perceptually similar to the original images. Although diffusion models exhibit significantly degraded generative capability when directly operating at such a low resolution, the use of commutative-zero guidance effectively mitigates this issue, enabling the production of visually coherent and reliable previews.

D. Additional Qualitative Results

This section presents additional qualitative results that could not be included in the main paper due to space constraints.

D.1. Qualitative Results on FLUX

Fig. S3, S4 provide qualitative comparisons demonstrating that our proposed method produces images that are perceptually more similar to the original ones compared with other baselines.

D.2. Qualitative Results on SD3.5-L

Similarly, Fig. S5, S6 show that, on Stable Diffusion 3.5 Large (SD3.5-L), our method yields results that better preserve perceptual similarity to the original images than the competing baselines.

D.3. Qualitative Results on Video Models

We conducted video synthesis experiments with Hunyuan-Video [2], as shown in Fig. S1, and observed that the reduced-timestep baseline (NFE 50→30) yields increasingly inconsistent compositions as the number of frames grows. In contrast, our method generates perceptually consistent LR videos while achieving a $1.75\times$ speedup, with generation times of 1,661 sec (Original), 1,025 sec (Reduced), and 951 sec **(Ours)** for 120-frame video generation.

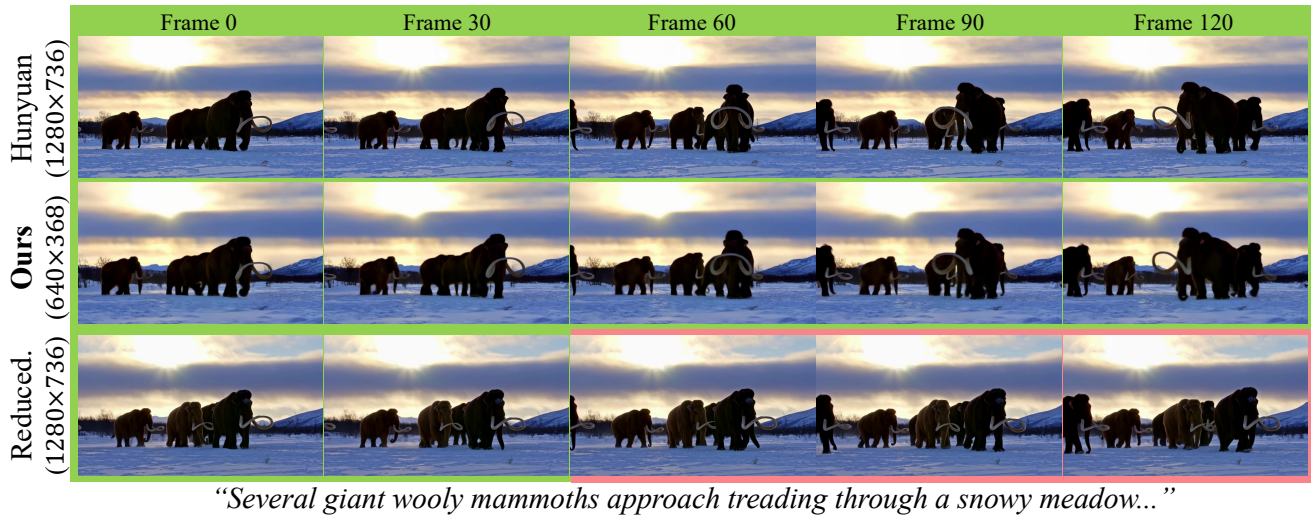

Figure S1. Qualitative results on video diffusion. Using our preview generation framework with HunyuanVideo, we compare against an alternative approach that reduces NFE from 50 to 30. While Mammoth deviates from the full-step composition beyond a certain frame, our preview generation faithfully preserves the composition of the high-resolution content.

D.4. Integration with Temporal-axis Acceleration

Tab. 3 shows that integrating our proposed method with the temporal-axis acceleration technique TaylorSeer [3] yields additional speedup with almost no performance degradation, improving the acceleration from 1.53× to 3.05×. Furthermore, Fig. S7 provides a qualitative comparison of the generated images. We observe that the outputs remain nearly indistinguishable both when applying our method alone and when incorporating TaylorSeer, indicating that the original images are preserved with high perceptual fidelity even at the 3.05× acceleration setting.

E. Limitation

The proposed approach relies on the approximate commutator-zero condition and trajectory compliance, both of which hold empirically but are not guaranteed across all architectures or sampling schedules. Since the method operates within a training-free setting, the selected downsampling operators are restricted to mutually exclusive, block-wise matrices, which may limit expressiveness and introduce sensitivity to spatial structure. The reuse of velocity predictions assumes local linearity of rectified flows, which may weaken under models exhibiting strong temporal variation or non-linear behavior, potentially reducing stability for more complex scenes or extreme prompts. Although the method improves LR–HR alignment, the corrections introduce additional computation, presenting an inherent trade-off between accuracy and acceleration. Finally, while the formulation generalizes to certain spatial manipulations, its applicability to broader transformations or non-flow-based generative frameworks remains an open direction for future investigation.

References

[1] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *NeurIPS*, 36:50742–50768, 2023. 2

[2] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 3

[3] Jiacheng Liu, Chang Zou, Yuanhuiyi Lyu, Junjie Chen, and Linfeng Zhang. From reusing to forecasting: Accelerating diffusion models with taylorseers. *arXiv preprint arXiv:2503.06923*, 2025. 4

[4] Yiren Song, Xiaokang Liu, and Mike Zheng Shou. Diffsim: Taming diffusion models for evaluating visual similarity. In *ICCV*, pages 16904–16915, 2025. 2

[5] Narasimhan Venkatanath, D Praneeth, S Channappayya Sumohana, S Medasani Swarup, et al. Blind image quality evaluation using perception based features. In *2015 twenty first national conference on communications (NCC)*, pages 1–6. IEEE, 2015. 1

[6] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011. 2

[7] Yasi Zhang, Peiyu Yu, Yaxuan Zhu, Yingshan Chang, Feng Gao, Ying Nian Wu, and Oscar Leong. Flow priors for linear inverse problems via iterative corrupted trajectory matching. *NeurIPS*, 37:57389–57417, 2024. 1

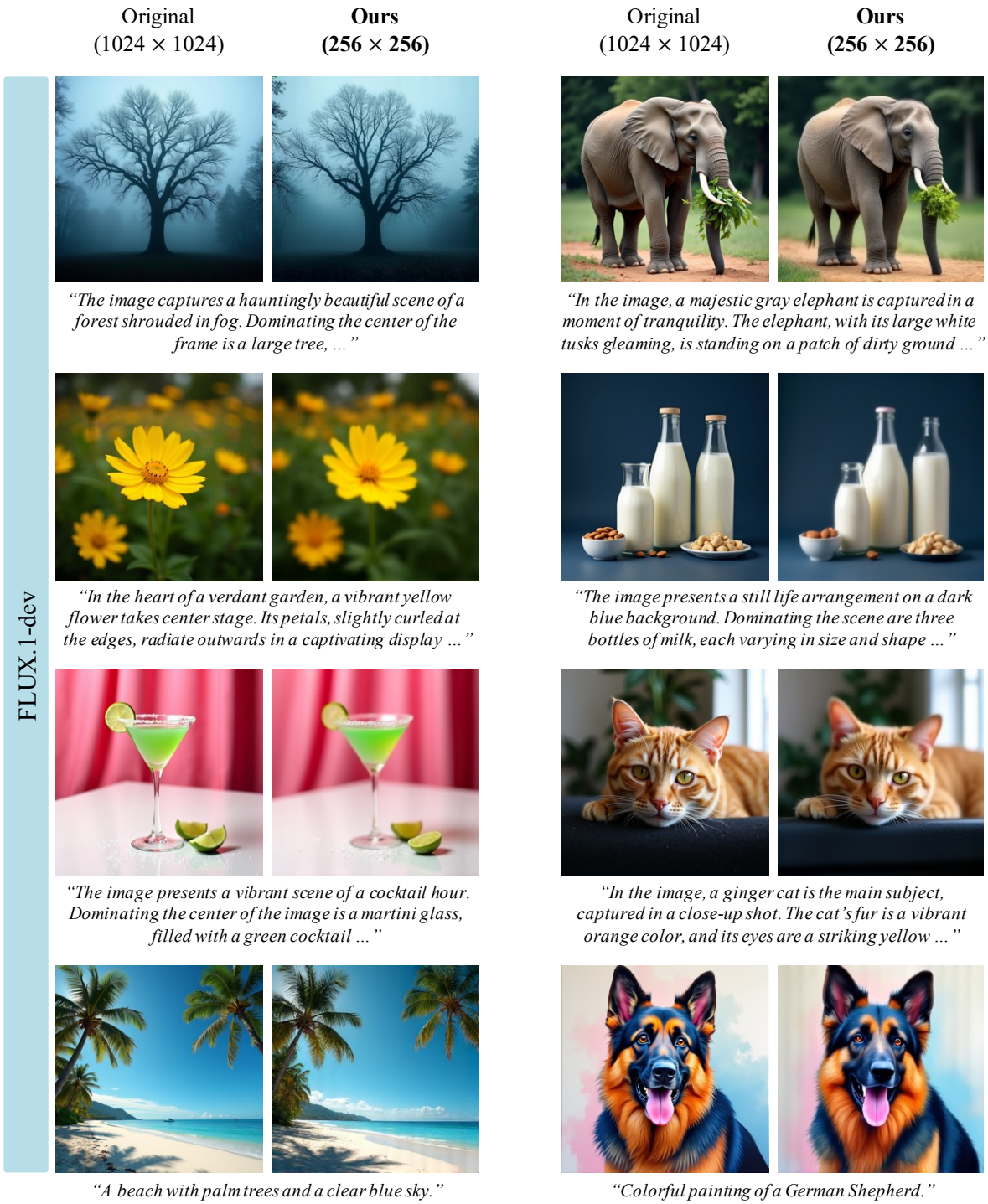

Figure S2. Qualitative results of our proposed method at a lower resolution of 256×256.

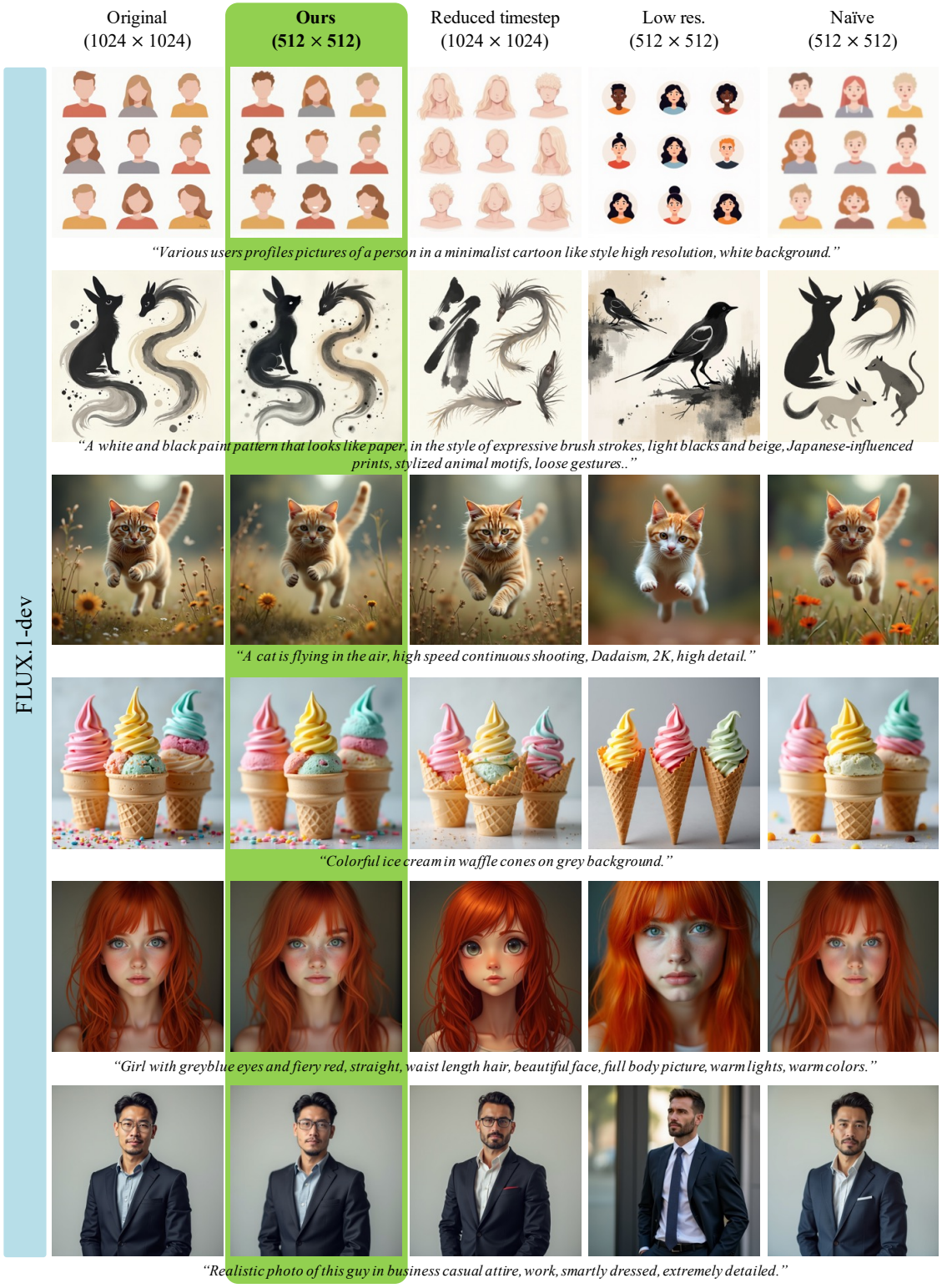

Figure S3. Qualitative comparison of our proposed method on FLUX.

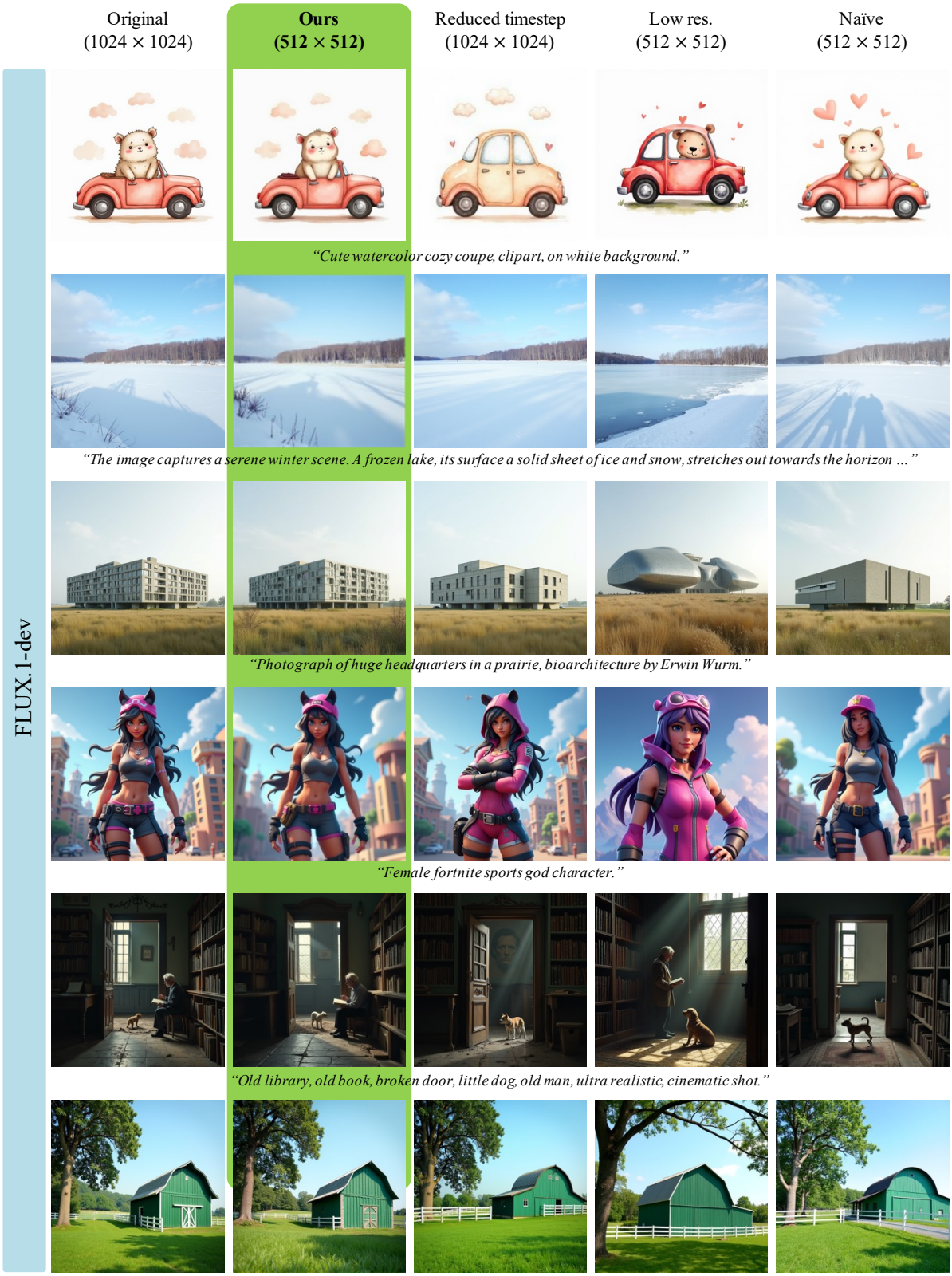

Figure S4. Qualitative comparison of our proposed method on FLUX.

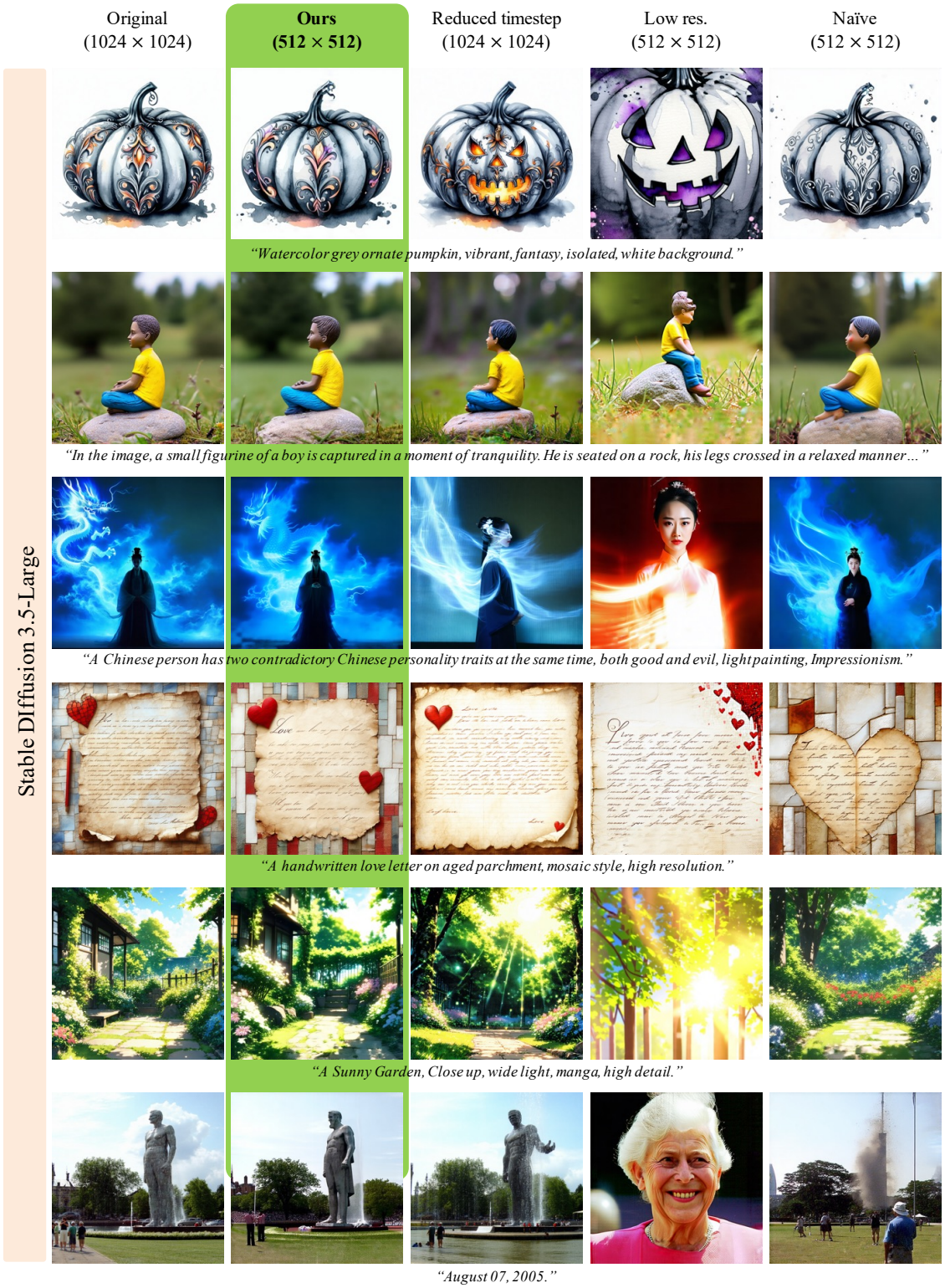

Figure S5. Qualitative comparison of our proposed method on SD3.5-Large.

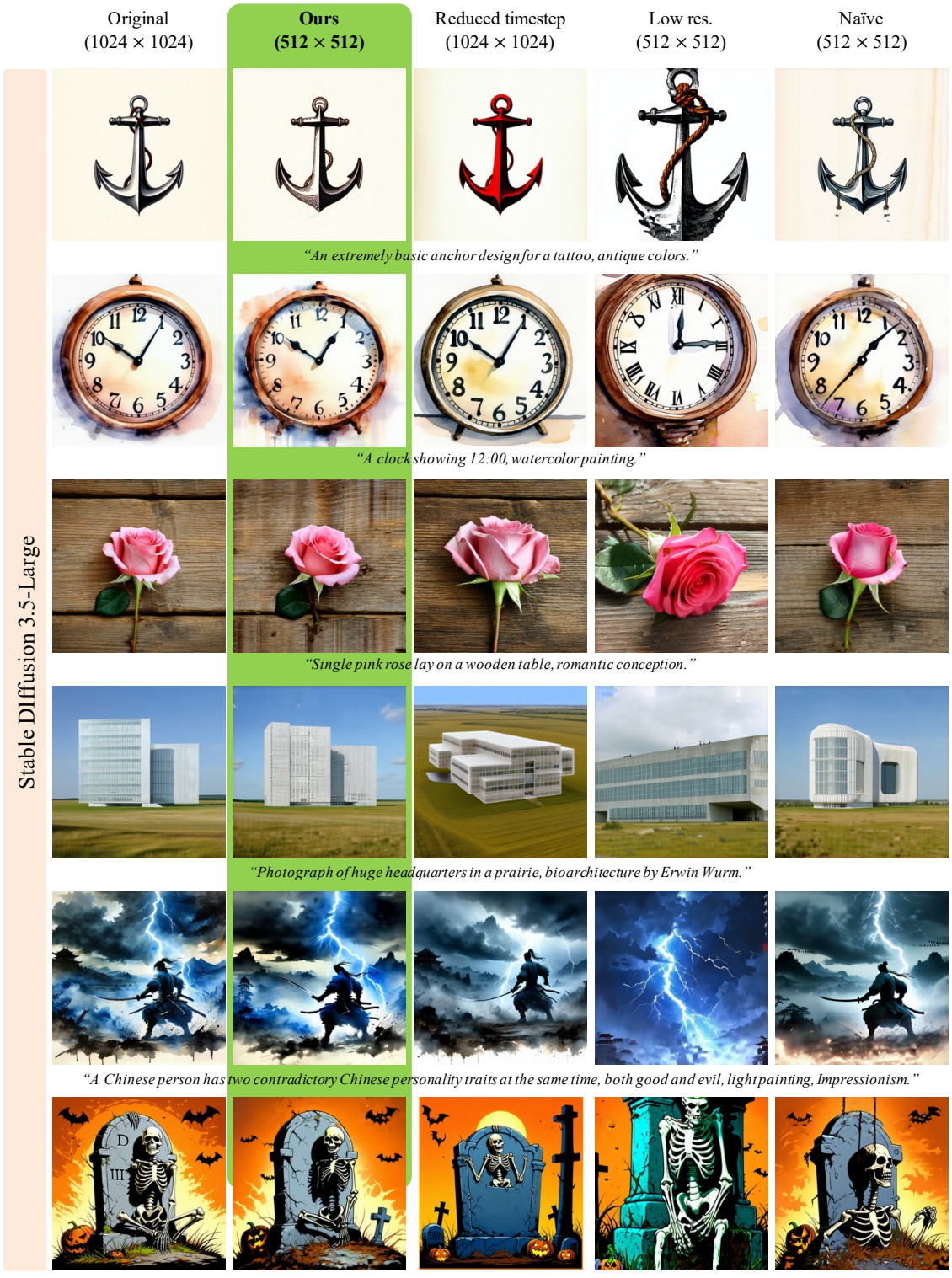

Figure S6. Qualitative comparison of our proposed method on SD3.5-Large.

Method	Original	Ours	Ours + TaylorSeer
Resolution	1024 × 1024	512 × 512	512 × 512
Computation	1800 TFLOPs	1178 TFLOPs	590 TFLOPs
Speedup	1.00 ×	1.53 ×	3.05 ×

Figure S7. Qualitative comparison after integrating temporal-axis acceleration on FLUX.