Fewer Denoising Steps or Cheaper Per-Step Inference: Towards Compute-Optimal Diffusion Model Deployment

Supplementary Material

In this supplementary material, we provide additional experimental results, analyses, and visualizations to complement our main paper.

- We provide an analysis on the module-level redundancy across denoising steps in Appendix A.
- We examine the impact of varying the resolution scale β used in our mixed-resolution denoising on the final performance in Appendix B.
- We show the ablation studies of hyperparameters m and k in Appendix C.
- We demonstrate the effectiveness of using a small calibration set to determine the optimal hyperparameters for mixed-resolution denoising in Appendix D.
- We further evaluate the effectiveness of the proposed PostDiff on Imagenet21K_Recaption dataset in Appendix E.
- We apply fine-tuning on top of PostDiff to further improve its performance in Appendix F.
- We provide an overview figure to illustrate our hybrid module caching strategy in Appendix G.
- We compare our PostDiff with prior works in Appendix H.
- We present additional visual examples and their corresponding prompts in Appendix I and Appendix J, respectively.

A. Motivations for Module-level Redundancy from Profiling and Previous Works

To profile the module-level redundancy across denoising steps, we calculate the average relative L_1 distances [6] of the i-th block's feature map between two consecutive steps, i.e., t and $t-1$, as $L1_{\text{rel}}(i,t) = \|F_i(x_t) - F_i(x_{t-1})\|_1 / \|F_i(x_t)\|_1$, where $F_i(\cdot)$ represents the output feature map of layer i. As shown in Fig. 1 (a), the feature maps are highly similar across all blocks throughout the entire denoising process, indicating the potential effectiveness of caching and reuse across steps. This insight is utilized by caching-based compression methods [5, 6].

Delving deeper, we calculate the average L_2 distances of different types of layers' feature maps between two consecutive timesteps, as shown in Fig. 1 (b). We observe that attention layers' feature maps have relatively low distances across all steps compared to convolution layers. This stability enables effective caching with minimal impact on image quality. Additionally, considering that (1) the text guidance provided by cross-attention layers primarily determines the image layout, i.e., the low-frequency structure of the image, and (2) the semantics-planning phase for layout generation occurs in the early denoising stages [4], as also discussed in Sec. 2.1 of our main paper, it is feasible to use the cached cross-attention feature map in the later denoising phase, as the low-frequency information has been determined early on.

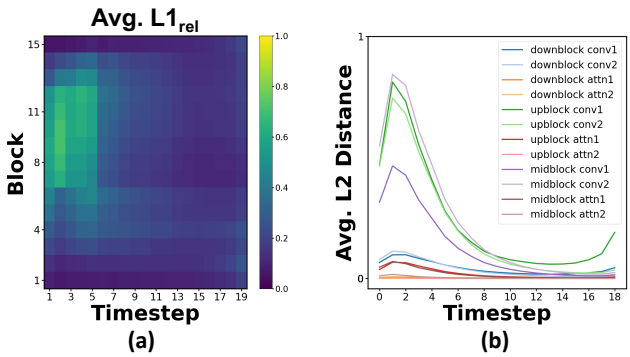

Figure 1. Visualize the feature map distance across consecutive denoising steps based on (a) block index, and (b) block types.

B. Impact of Varying the Resolution Scale β

We analyze the effect of varying β, the scaling factor applied to the resolution discussed in Sec. 2.2 of our main paper, on the performance of the mixed-resolution strategy.

Setup. We use SD V1.5 (DreamShaper-7 version) as our backbone and the MS-COCO 2014 validation dataset as the test set. The scale factor $\beta = \{0.375, 0.5, 0.625, 0.75, 0.875\}$ (from the leftmost points to the rightmost points on each curve in Fig. 2) is varied while keeping $s = 0.5$ fixed, i.e., low generation resolution is applied only in the first half of the denoising steps.

Observations. As illustrated in Fig. 2, we observe that (1) $\beta = 0.5$ is generally a close-to-optimal choice, offering relatively low FIDs while significantly reducing FLOPs. As such, we adopt this design choice in our main paper; (2) as β increases from 0.5 to 0.875, the FIDs increase slightly. This is likely due to the low-resolution part becoming closer to the full resolution, which introduces more inaccurate high-frequency components during the early denoising stage, as analyzed in Sec. 2.1 of our main paper; (3) smaller β values result in lower FLOPs, while an extremely small $\beta = 0.375$ leads to a notable FID increase, as the extremely low resolution makes it difficult to recover fine details.

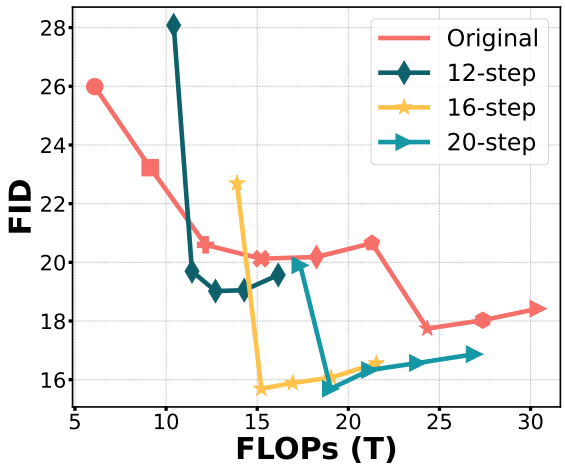

Figure 2. The FID-FLOPs trade-off achieved under different denoising steps and different β. The data points with the same number of denoising steps are annotated using the same shape.

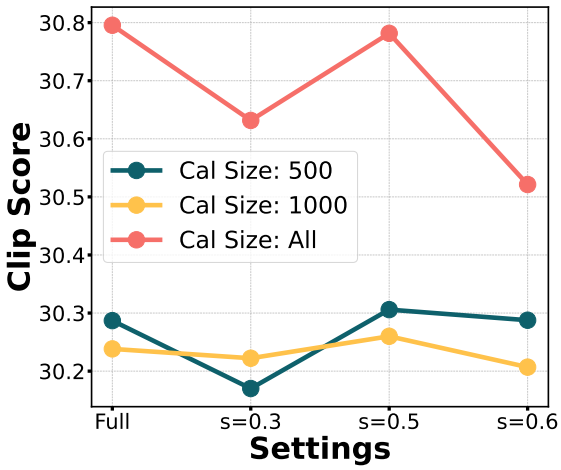

Figure 3. The achieved Clip Scores when using different calibration set sizes with varied s, where "Full" indicates no use of the mixed-resolution denoising.

C. Ablation Studies on k and m

We conducted ablation studies on k (cache update frequency) and m (the point at which CFG is abandoned) using SD V1.5 on an NVIDIA A5000 GPU. As shown in Tab. 1, a larger k (i.e., more cache reuse) or a smaller m (i.e., abandoning CFG earlier) results in greater degradation of generation quality. Based on this analysis, we adopt $k = 2$ and $m = 0.75$ in our paper, which strikes a sweet-spot efficiency-fidelity trade-off.

Table 1. Ablation studies on k and m.

Settings	FID ↓	Clip Score ↑	Latency (s) ↓
Original	18.42	30.80	2.93
$k = 2$	16.65	30.25	1.14
$k = 3$	26.96	27.68	0.97
$k = 4$	39.22	26.16	0.94
$k = 5$	51.39	25.84	0.75
$m = 0.45$	17.94	28.82	1.00
$m = 0.6$	17.21	29.50	1.08
$m = 0.75$	16.65	30.25	1.14
$m = 0.9$	16.49	30.06	1.26

D. Effectiveness of Small Calibration Sets

As mentioned in Sec. 2.4 of our manuscript, a small calibration set can be leveraged to determine the optimal hyperparameter s, i.e., the portion of denoising steps performed at a low generation resolution. We validate the effectiveness of this strategy in this section.

Setup. Using SD V1.5 as the backbone, we randomly sampled subsets of the MS-COCO 2014 validation dataset with sizes of 500 and 1000. We tested the average Clip Scores of the generated images for $s = \{0.0\,(\text{Full}), 0.3, 0.5, 0.6\}$.

Observations. As shown in Fig. 3, the performance trends for the small calibration sets align well with those for the full dataset. Specifically, a setting that achieves a higher Clip Score on a small calibration set is highly likely to achieve higher Clip Scores when using the full dataset. This strategy eliminates the cumbersome process of hyperparameter tuning on a large amount of data.

Table 2. Evaluation on Imagenet21K_Recaption dataset. Latency is measured as the single-image generation time on an NVIDIA A5000 GPU for SD V1.5 and LCM, and an NVIDIA H100 GPU for PixArt-α.

Model	FID ↓	Clip Score ↑	Latency (s) ↓
SD V1.5	30.08	32.13	2.930
w/ PostDiff	27.57	31.75	1.139
LCM	39.00	28.72	0.825
w/ PostDiff	37.74	28.22	0.651
PixArt-α	35.58	32.40	1.752
w/ PostDiff	32.69	32.04	1.382

E. Evaluation on the Imagenet21K_Recaption Dataset

In Tab. 2, we apply PostDiff to SD V1.5, LCM, and PixArt-α and evaluate the achieved performance on the Imagenet21K_Recaption dataset[1], where we randomly pick up

[1] https://huggingface.co/datasets/gmongaras/Imagenet21K_Recaption

Table 3. Performance of PostDiff w/o and w/ fine-tuning (FT).

Model	FT	FID ↓	Clip Score ↑
SD V1.5		63.64	19.47
	✓	22.28	28.68
BK-SDM-Tiny		19.87	25.87
	✓	19.38	28.44

10,000 text prompts.

The results show that applying PostDiff consistently improves the FID of different models while maintaining a comparable Clip Score and lower latency. This further demonstrates the robustness and generality of our proposed method across diverse text prompts and diffusion models.

F. Combine PostDiff with Fine-Tuning

As a common practice, diffusion models are typically trained on a single resolution, which can lead to relatively lower-quality outputs for low-resolution images. To address this, we experiment with fixed resolution schedules to fine-tune diffusion models, aiming to explore whether this approach can enhance the performance of our mixed-resolution strategy and, in turn, further improve PostDiff.

Setup. We fine-tune SD V1.5 (the original version, as we were unable to obtain the dataset used for Dreamshaper's fine-tuning) and BK-SDM-Tiny [2] (a light-weight version of SD V1.5) using the LAION-Art dataset[2]. The batch size is set to 128, with 15,000 training iterations and a learning rate of 1e-5. We set $\beta = 1/2$ and $s = 1/2$ for SD V1.5, i.e., when the corresponding low-resolution steps (i.e., 501-1000) are randomly selected, half-resolution images are used to fine-tune the model, and $\beta = 1/2$ and $s = 1/5$ for BK-SDM-Tiny.

Observations. As shown in Tab. 3, fine-tuning notably improves both FID and Clip Score, demonstrating the compatibility of our PostDiff with fine-tuning. Given that this process is fast and lightweight (about one day using one NVIDIA H200 GPU for SD V1.5), it provides a practical way to deploy PostDiff with enhanced performance while maintaining high generation efficiency.

G. Illustration of Hybrid Module Caching

In Fig. 4, we illustrate our hybrid module caching strategy, which effectively leverages feature redundancy across different timesteps, complementing the method description in Sec. 2.5 of our main paper.

[2] https://huggingface.co/datasets/fantasyfish/laion-art

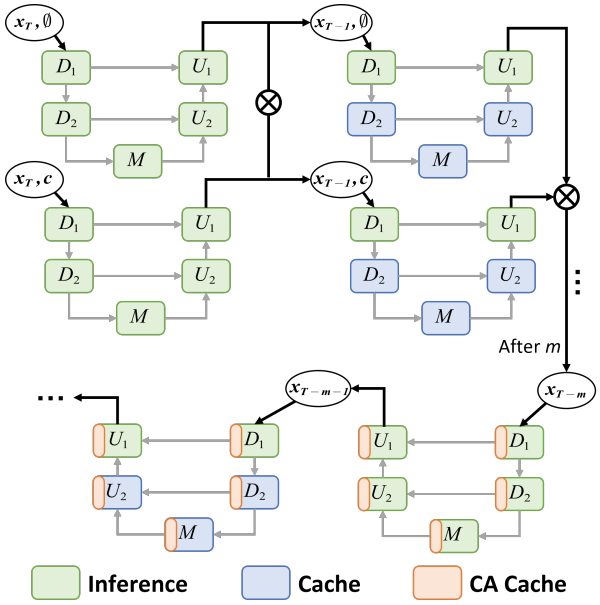

Figure 4. Overview of our hybrid module caching strategy.

H. Comparisons with Prior Works

PostDiff adopts a post-training setting, i.e., it requires no fine-tuning and is easy to use. The effectiveness of mixed-resolution denoising in this setting stems from the distinct emphasis on low- and high-frequency components in the early and late denoising stages. In contrast, Pyramid Flow [1] interprets the diffusion trajectory as a multi-stage pyramid and trains all pyramid stages end-to-end from scratch or through model-specific fine-tuning, where the denoising process is explicitly adapted to the chosen resolution. The two works leverage different aspects of resolution.

DeepCache [5] is specifically tailored for U-Net, whereas PostDiff is applicable to both U-Net and DiT, achieving a 3.72 FID improvement with a 32.7% latency reduction on PixArt-α. Furthermore, neither DeepCache nor Faster Diffusion [3] addresses redundant CFG during the late denoising phase. In contrast, PostDiff removes this redundancy to further boost efficiency, achieving 43.1%/34.4% latency reduction and improving FID by 1.2/1.1 compared to Faster Diffusion and DeepCache, respectively (see Tab. 3 of our main paper).

I. More Visualization Results

We visualize the text-to-image results generated by the original models alongside those produced using our PostDiff approach in Fig. 5, as a complement to Fig. 6 of our main paper. As evidenced by the FID Scores, PostDiff achieves significant speedups while preserving details and, in some cases, further enhances visual quality.

In addition, we include additional visual samples along

with Clip Scores at each denoising step under different mixed-resolution settings in Fig. 6, as a complement to Fig. 3 of our main paper. It is consistently observed that utilizing low-resolution images to capture low-frequency components can potentially improve the final results.

Furthermore, we also provide additional visual examples showcasing the impact of different choices for CA_{cache} with varying m values in Fig. 7, as a complement to Sec. 2.5 of our main paper. The "Cond" choice generally performs best, maintaining better consistency with the prompt.

J. Prompts for Text-to-Image Generation

In this section, we provide the prompts we used to generate images in the main paper and supplementary material.

J.1. Figure 6 in the Main Paper

J.1.1. SD V1.5

1. A mystical underwater scene with glowing coral
2. A bustling cyberpunk metropolis at night, illuminated by a kaleidoscope of neon lights and holographic advertisements. The streets are crowded with people wearing futuristic attire.
3. A forest with glowing mushrooms and creatures.
4. A fantasy castle on a hill surrounded by clouds.
5. A snowy mountain landscape with a cozy cabin and smoke coming from the chimney.
6. Imagine and detail very clearly: the exciting rebirth of an energetic being in the vast void of space, together with a glorious phoenix. It begins with a stellar background that dazzles with the beauty of the cosmos, with brilliant nebulas and resplendent constellations.

J.1.2. PixArt-α

1. A stunning Japanese-inspired fantasy painting of a lone samurai, silhouetted against a massive full moon, standing beneath a windswept, crimson-leafed tree. Falling petals swirl around him, creating a melancholic yet serene atmosphere. The dramatic chiaroscuro lighting highlights the dramatic contrast between the cool-toned background of deep blues and grays and the warm reds of the foliage. This captivating scene is reminiscent of Yoshitaka Amano's work, with the dramatic lighting of Ivan Shishkin.
2. MysticSplash. Ink-splash-style. Ink-splash-style. Extreme closeup of a dapper figure in a stylized, richly detailed black top hat, adorned with decorative golden accents, stands against a white background. The character is a skeleton with very detailed skull and long canines as vampire fangs. He cloaked in a vibrant victorian jacket, featuring intricate golden embellishments and a deep red vest underneath. He wears a large victorian monocle with a yellow-tinted lense and copper frame very reddish. Exquisite details include a shiny silver cross and a blue gem on the chest, harmonizing with splashes of paint in vivid hues of blue, gold, and red that artistically cascade around the figure, blending an impressionistic flair with elements of surrealism. The atmosphere is whimsical and opulent, evoking a sense of grandeur and mystery.
3. An ancient, overgrown temple in a dense jungle, illuminated by the soft light of early morning.
4. The image is a landscape photograph of a mountain range with a river flowing through it. The river is surrounded by a rocky shoreline with small pebbles and boulders. The water is a deep blue color and reflects the mountains and trees in the water. The mountains in the background are tall and imposing, with a mountain peak in the distance. The sky is clear and blue, and the sun is shining brightly, creating a beautiful reflection of the mountains on the water's surface. The overall mood of the image is peaceful and serene. mad-sprklngtr, flrlizer.
5. A portrait of a cybernetic geisha, her face a mesmerizing blend of porcelain skin and iridescent circuitry. Her elaborate headdress is adorned with bioluminescent flowers and delicate, glowing wires. Her kimono, a masterpiece of futuristic design, shimmers with holographic patterns that shift and change, revealing glimpses of the complex machinery beneath. Her eyes gaze directly at the viewer with an enigmatic expression.
6. A single, crazy blue and black spaceship in the sky. It overwhelms the viewer with its artistic flying skills while trailing a meteor tail. Ace pilot of the Republic who was unrivalled in the 1940s. His second name is: The Magician of the Blue Wings, a genius aviator, one of a kind in 100 years. The warriors who challenged him, were destroyed by him, were overrun by him and scattered became many stars. The Milky Way is said to be the graveyard of such aerialists.

J.2. Figure 5 in the Supplementary Material

J.2.1. LCM

1. A massive wave crashing onto a rocky shore, with a lone figure standing defiantly against the storm, holding a glowing staff.
2. A whimsical town where all the buildings are made of candy, with rivers of chocolate and lollipop streetlights glowing faintly in the dusk.
3. The long journey home, vibrant glow.
4. Parisian luxurious interior penthouse bedroom, dark walls, wooden panels.
5. Novuschroma style cup of coffee with swirling steam.
6. Scottish fold kitten, professional photo, in snow, high detail, close-up view, quantum rendering, masterpiece, professional photo.

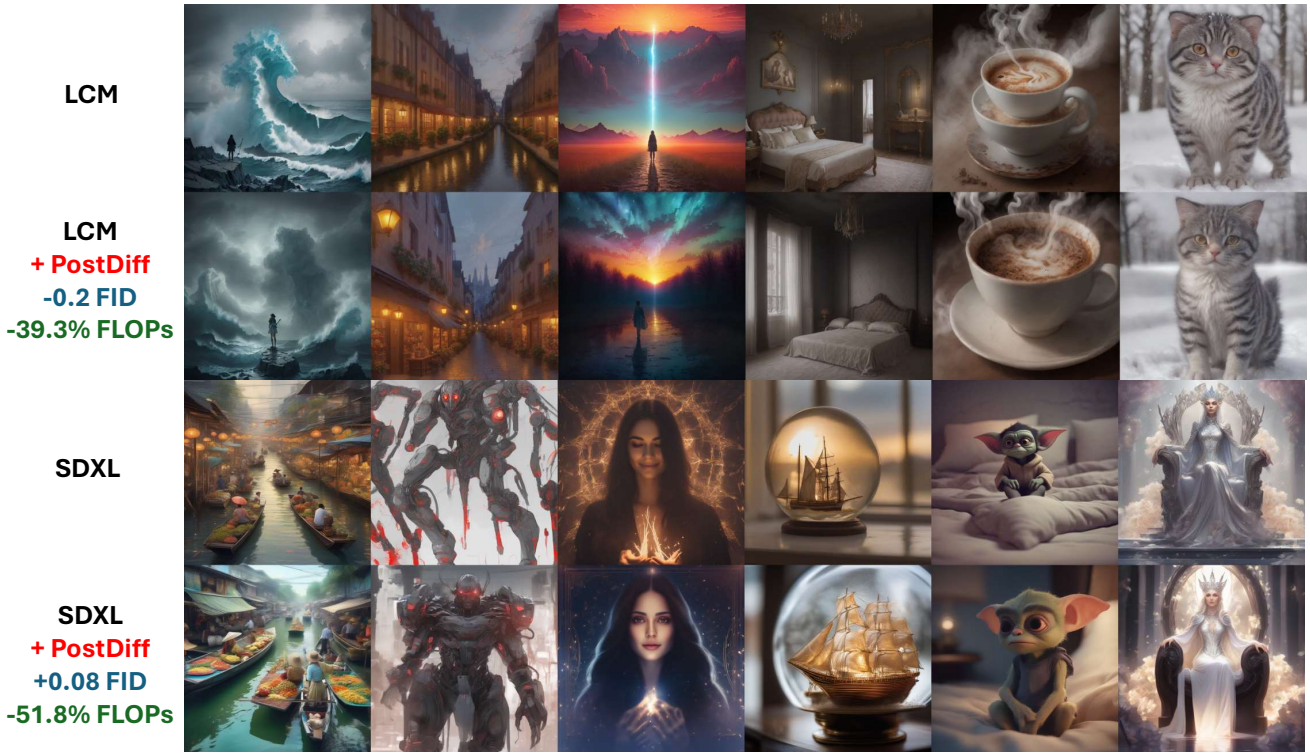

Figure 5. More visualization of generated images of different diffusion models w/o and w/ using our PostDiff.

J.2.2. SDXL

1. Floating market of old Bangkok by day, atmospheric lighting, awesome background, highly detailed, cinematicfantasy, dreaming, best quality, double exposure, realistic, whimsical, fantastic, splash art, intricate detailed, hyperdetailed, maximalist style
2. A cybernetic warrior with mechanical arms and glowing red eyes.
3. A smiling beautiful sorceress with long dark hair and closed eyes wearing a dark top surrounded by glowing fire sparks at night, symmetrical body, symmetrical face, symmetrical eyes, magical light fog, deep focus+closeup, hyper-realistic, volumetric lighting, dramatic lighting, beautiful composition, intricate details, instagram, trending, photograph, film grain and noise, 8K, cinematic, post-production.
4. Miniature sailing ship sailing in a heavy storm inside of a horizontal glass globe inside on a window ledge golden hour, home photography, 50mm, Sony Alpha a7.
5. Little cute gremlin sitting on a bed at night thinking about the world, cinematic, muted colors, faded, by pixar and dreamworks.
6. A regal elf queen sitting on a crystalline throne, her gown shimmering like liquid silver, with a crown of glowing flowers.

References

[1] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *ICLR*, 2025. 3

[2] Bo-Kyeong Kim, Hyoung-Kyu Song, Thibault Castells, and Shinkook Choi. BK-SDM: A lightweight, fast, and cheap version of stable diffusion. In *ECCV*, 2024. 3

[3] Senmao Li, Taihang Hu, Fahad Shahbaz Khan, Linxuan Li, Shiqi Yang, Yaxing Wang, Ming-Ming Cheng, and Jian Yang. Faster diffusion: Rethinking the role of unet encoder in diffusion models. In *NeurIPS*, 2024. 3

[4] Haozhe Liu, Wentian Zhang, Jinheng Xie, Francesco Faccio, Mengmeng Xu, Tao Xiang, Mike Zheng Shou, Juan-Manuel Pérez-Rúa, and Jürgen Schmidhuber. Faster diffusion through temporal attention decomposition. *TMLR*, 2025. 1

[5] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *CVPR*, 2024. 1, 3

[6] Felix Wimbauer, Bichen Wu, Edgar Schoenfeld, Xiaoliang Dai, Ji Hou, Zijian He, Artsiom Sanakoyeu, Peizhao Zhang, Sam Tsai, Jonas Kohler, et al. Cache me if you can: Accelerating diffusion models through block caching. In *CVPR*, 2024. 1

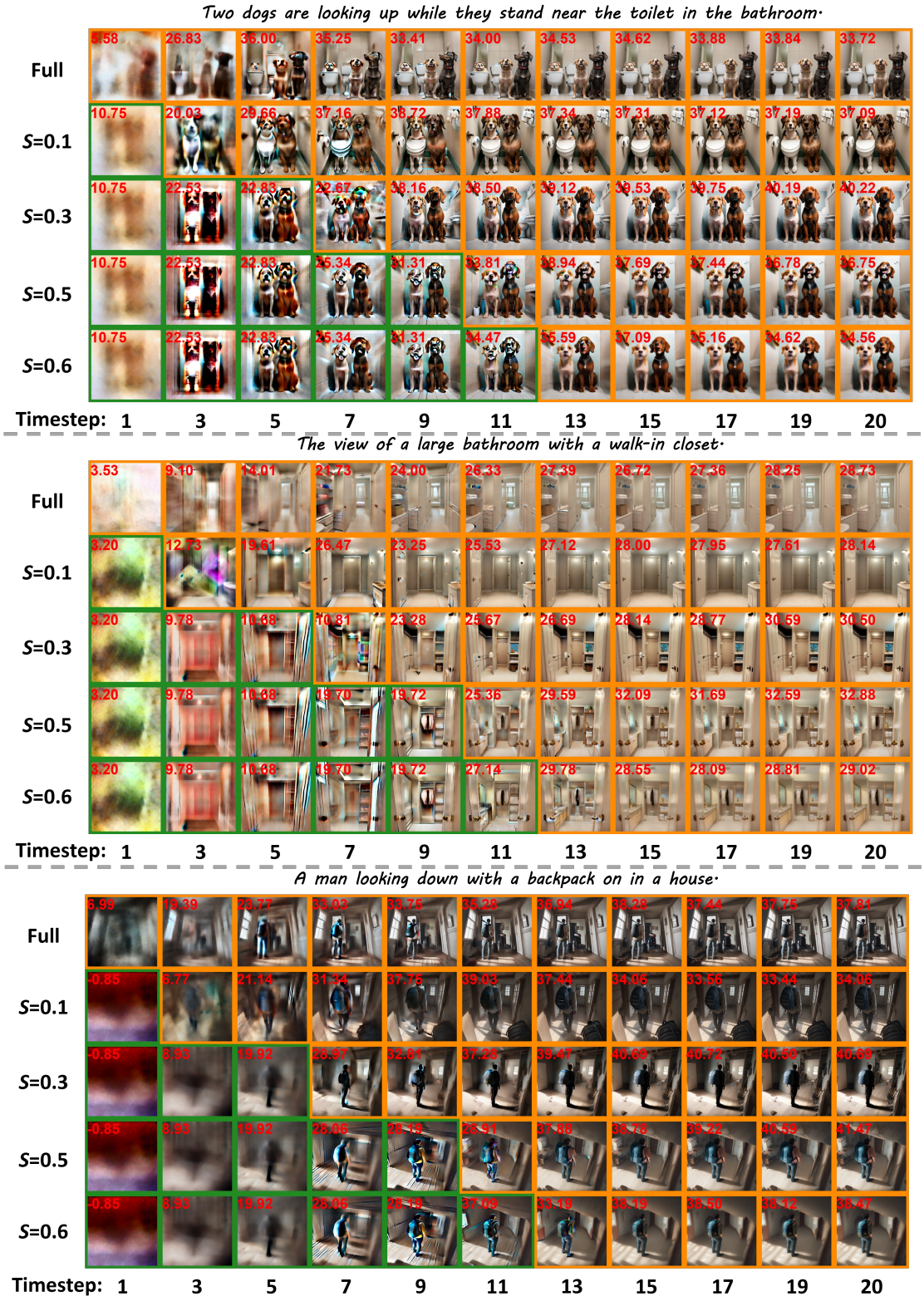

Figure 6. Visualize the Clip Score after each denoising step using full resolution (row 1) and mixed resolution (rows 2-5). The steps using low generation resolution are highlighted in green and those using high generation resolution are highlighted in orange.

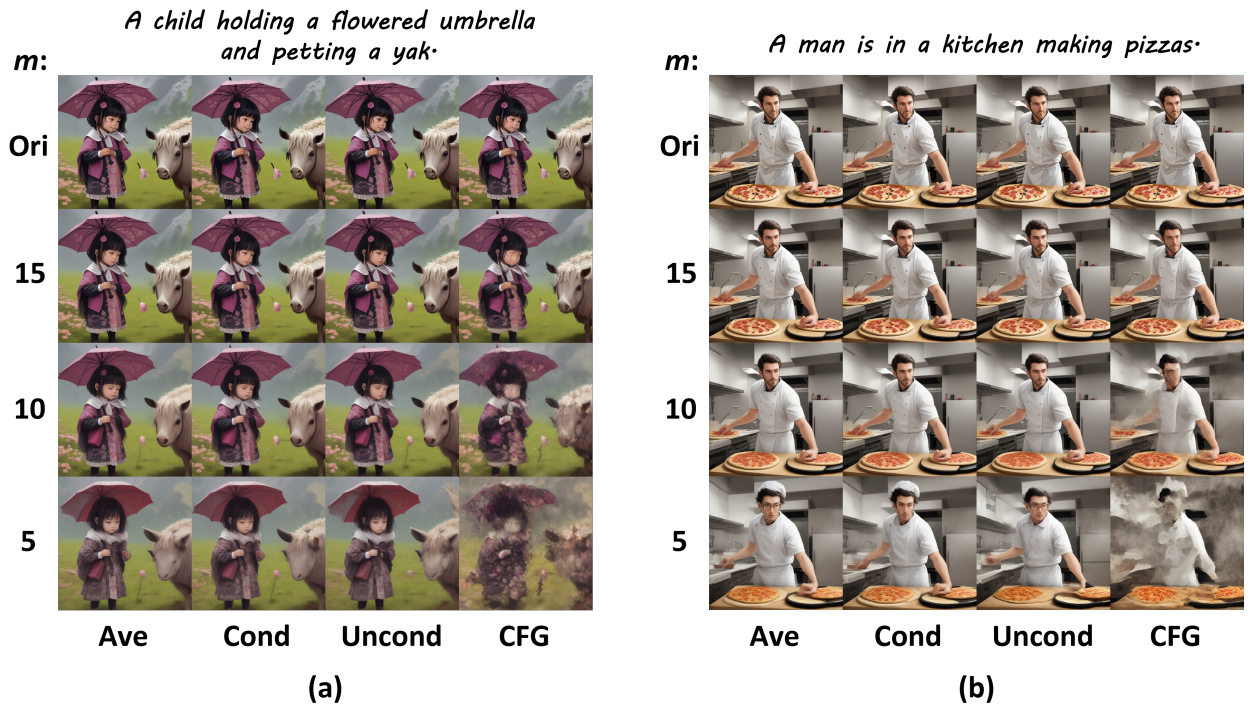

Figure 7. Generation results using different choices of CA_{cache} with varying m. "Ori" indicates no use of cross-attention cache.