A. Limitations

Our evaluations were primarily conducted on PixArt-Σ and Stable Diffusion 3.5-Medium. Future work should investigate the generalizability of our approach to larger models such as Flux [11]. Additionally, while quantitative metrics were used, the qualitative assessment of generated images remains subjective, and human evaluation could provide further insights into image quality.

B. Additional Runtime Results

Table 4 shows additional runtime results for our models, the teachers, and all baselines and ablations for the relevant use-case of generating 1024×1024 images on consumer-grade hardware.

C. Finetuning

To show that our models are amenable to standard fine-tuning, we generate a high-quality synthetic dataset by generating $\approx 960,000$ images with Flux[Schnell] [11] re-using the YE-POP captions. From these, we use $\approx 24,000$ for validation and testing each. We then fine-tune our best distilled EDiT and MM-EDiT checkpoints using the remaining images while selecting the model using the validation set.

The results in Table 5 show that fine-tuning, as expected, improves FID on the test set, indicating that our models can not only be trained by disitillation but also using the standard diffusion losses of the teachers.

D. Prompts and Additional Qualitative Results

The prompts used to generate the images for the 1024×1024 pixel images generated with EDiT, PixArt-Σ and corresponding baselines and ablations. (from left to right):

1. "Pastel landscape, nature, detail"
2. "A polar bear walking through icy and snowy terrain"
3. "Painting of a wild mystical druid holding a shining golden etched dagger standing beside a crystal clear lake, woodland forest surrounded by yellow and purple coloured wild flowers, hanging moss and vines, moody lighting"
4. "A soft black and white charcoal sketch of a beautiful young mother holding her baby, blue eyes with dark circles around the irises, old canvas, minimalist, blocky, gritty"
5. "The image captures a group of people standing on a stone wall, overlooking a breathtaking mountainous landscape. The sky above them is a dramatic blend of blue and white, suggesting an overcast day. The mountains in the distance are blanketed in snow, adding a serene beauty to the scene. The perspective of the image is from behind the group, looking out towards the mountains, giving a sense of depth and scale to the landscape. The colors in the image are muted, with the blue of the sky and the white of the snow standing out against the darker tones of the mountains and the stone wall"

The prompts used to generate the images for the 2048×2048 pixel images with EDiT and PixArt-Σ

1. Realistic portrait pet portrait
2. older adult alien with feathers, no hair, in vertical striped pyjamas, white background, small ears, wearing a scarf

The prompts used to generate the images for the 1024×1024 pixel images generated with MM-EDiT, SD-v3.5 and corresponding baselines and ablations. (from left to right):

1. "In the heart of a verdant field, a majestic brown and white horse stands, its gaze meeting the camera with an air of quiet confidence. The horse's coat is a rich brown, adorned with white spots that add a touch of whimsy to its appearance. A distinctive white stripe runs down its nose, adding to its unique charm. Its mane and tail, both a deep black, contrast beautifully with its coat. The horse's eyes, full of life and curiosity, seem to be looking directly at the camera, creating a sense of connection between the viewer and the subject. The background is a lush green field, dotted with trees and bushes, providing a serene and natural backdrop to this captivating scene. The image exudes a sense of tranquility and harmony with nature."
2. " A beach with the sun rising and the waves crashing on the sand, beauty light"
3. "Super cool astronaut NFT, facing camera, stylish, bling glasses, photochromatic, plain background, futuristic 8k. $->$ Super cool astronaut NFT, facing camera, stylish, bling glasses, photochromatic, plain background."
4. " The image captures a close-up of a woman's face, bathed in a mesmerizing blend of blue and yellow lights. The woman's eyes, wide open, gaze directly into the camera. Her hair, styled in loose waves, adds a touch of softness to the intense lighting. The background is shrouded in darkness, punctuated by streaks of light from the left side of the frame. The image does not contain any discernible text or countable objects, and there are no visible actions taking place. The overall composition is balanced, with the woman's face as the central focus."
5. "The image captures a serene indoor setting, dominated by a white vase that holds a bouquet of flowers. The vase, positioned on the left side of the frame, is filled with three distinct types of flowers, each contributing to a harmonious blend of colors and shapes. The most prominent flower is a pink dahlia, its petals radiating out from a yellow center. This dahlia is situated on the right side of the vase, its vibrant color standing out against the white of the vase. In addition to the dahlia, there are two white chrysanthemums. These flowers are located on the left side of the vase, their pristine white petals adding a touch of elegance to the arrangement. The background of the image is blurred, drawing focus to the flowers in the vase. The overall composition of the image creates a sense of tranquility and beauty."

Model				Latency (s)
PixArt-Σ				5.44
EDiT (ours)				4.15
SANA-DiT				4.0
LinFusion-DiT				4.2
KV Comp. ($k = 2$)				5.0
Ablations	Q	K	V	
	CF	CF	-	4.55
	-	SC	SC	3.8
	CF	-	-	4.3

Model	Latency (s)
SD-v3.5M	10.9
MM-EDiT (Ours)	8.0
MM-EDiT with η^{Lin}	8.6
MM-EDiT no ϕ_{CF} and ϕ_{SC}	7.6
SANA-MM-DiT	7.6
Linear MM-DiT-α	7.7
Linear MM-DiT-β	7.75

Table 4. Latency of generating a single 1024×1024 pixel image on a consumer-grade Nvidia 3090 RTX. PixArt-Σ, EDiT, and all the respective ablations use 20 diffusion steps. SD-v3.5M, MM-EDiT, and all the respective ablations use 28 diffusion steps. These results further emphasize that, while producing images of similar quality to their respective teachers, EDiT and MM-EDiT improve efficiency at the practically relevant resolution of 1024×1024 and strike the best compromise of runtime and quality across all considered variants.

	FID (Inception-v3)	FID CLIP	FID (Inception-v3)	FID CLIP
	512×512		1024×1024	
PixArt-Σ	7.49	2.52	8.55	3.17
EDiT (before finetuing)	7.36	2.53	10.1	3.42
EDiT (after finetuing)	5.17	1.84	4.72	1.39
SD-v3.5	11.5	4.41	8.91	5.07
MM-EDiT (before finetuing)	8.43	2.81	11.6	5.28
MM-EDiT (after finetuing)	4.46	1.69	7.15	2.89

Table 5. Quantitative comparison of fine-tuning EDiT and MM-EDiT to fine-tuned pre-trained models. The results show that EDiT amd MM-EDiT outperform fine-tuned baselines.

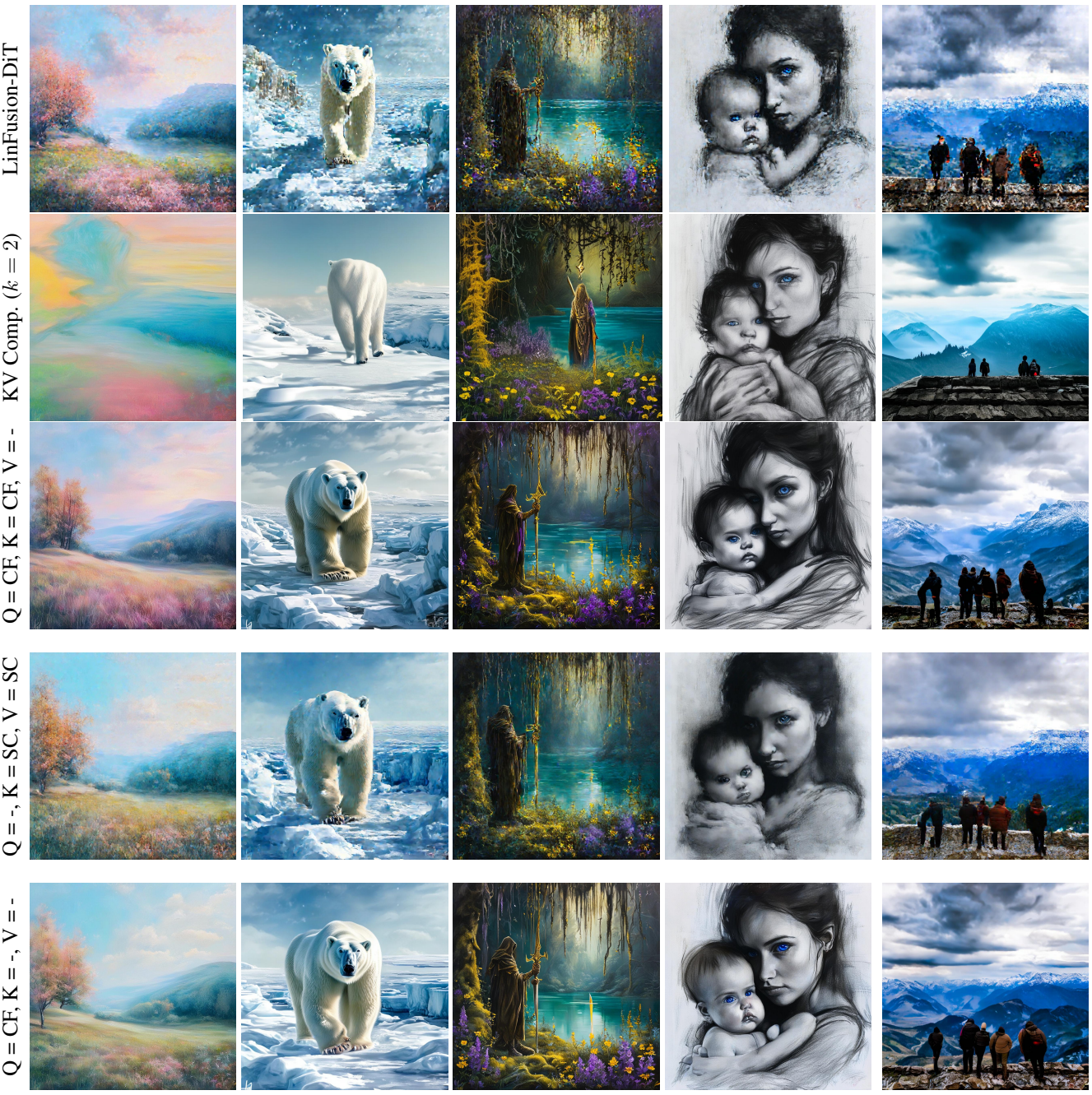

Figure 6. 1024×1024 pixel images for the EDiT ablations and baselines not shown in the main paper.

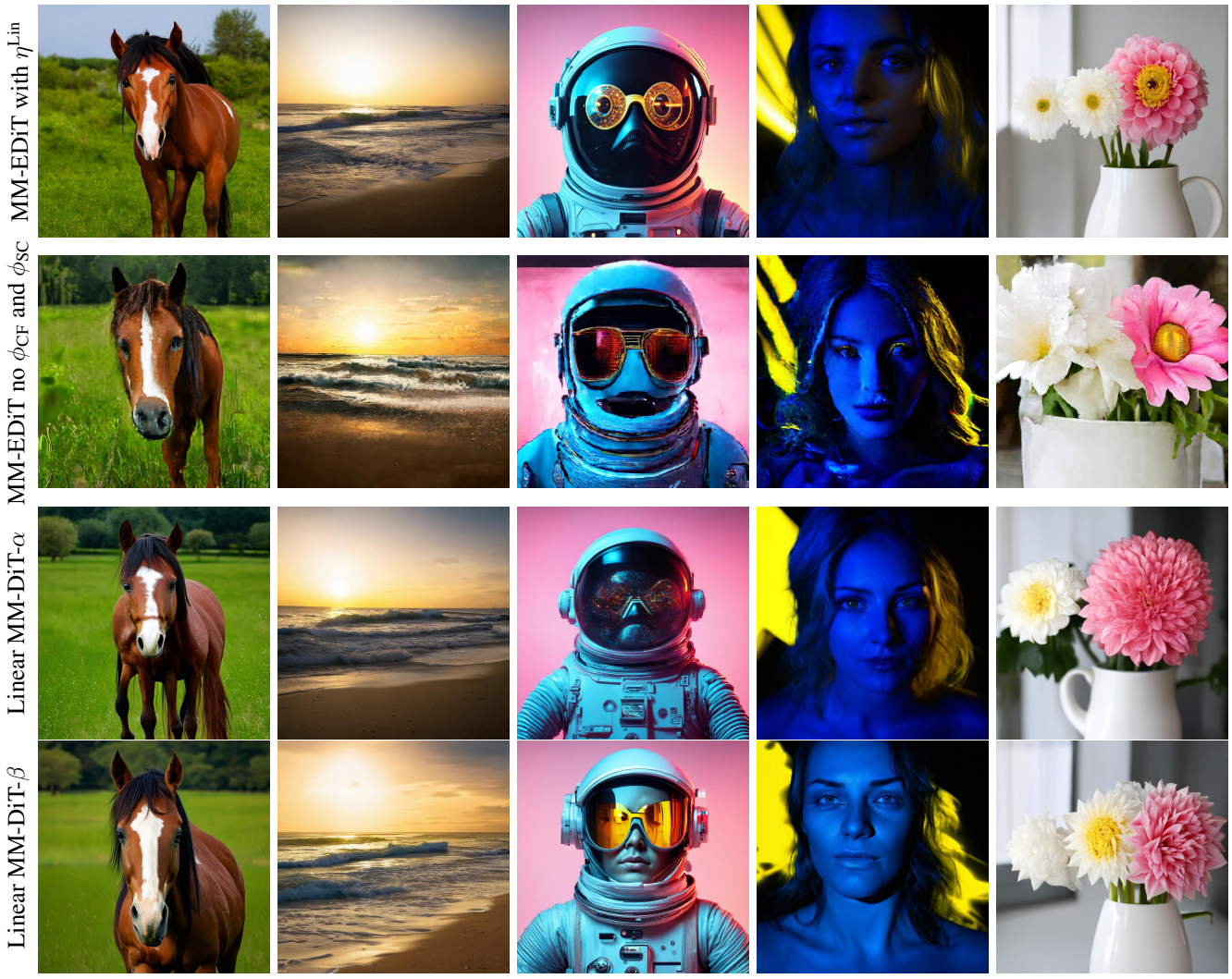

Figure 7. 1024×1024 pixel images for the MM-EDiT ablations and baselines not shown in the main paper.