Supplementary Material for: Merging and Splitting Diffusion Paths for Semantically Coherent Panoramas

Fabio Quattrini, Vittorio Pippi,
Silvia Cascianelli, and Rita Cucchiara

University of Modena and Reggio Emilia, Modena, Italy
{name.surname}@unimore.it

In this document, we report additional results of our proposed MAD approach for inference-time adaptation of diffusion models for generating long images. We explore the effect of changing the application stage and varying the number of application steps, present more detailed information regarding the user study, and offer further comparisons with SD-L. Additionally, we report other qualitative results, also at different aspect ratios, and further discuss the limitations.

1 Effect of the Application Stage

In Tables 1 and 2, we report a quantitative analysis of the effects of applying MAD at different stages in the noise prediction model of the considered LDM and LCM. We observe that when MAD is applied only in the bottleneck (Mid blocks), FID, KID, and mGIQA values are comparable to those obtained with the baseline in which MAD is never applied and the perceptual coherence is lower than in the other settings. On the other hand, I-LPIPIS, I-StyleL, and mCLIP get better when MAD is applied more (*i.e.*, in all blocks) and in later stages of the U-Net (*i.e.*, in the upsampling blocks). The same trend can be observed from the LDM qualitatives in Fig. 2.

2 Effect of the Number of Application Steps

We report a quantitative analysis (Tab. 3) and a more exhaustive qualitative analysis (Fig. 1) on the effect of the number of inference timesteps in which the proposed MAD operator is applied (defined by the hyperparameter τ). In this experiment, we use the considered LDM run for 50 reverse process steps and apply MAD at each timestep from the first one up to a certain threshold. As we can see, 15 steps lead to a good trade-off between variety and uniformity.

3 Details of the User Study

Here, we report further details on the user study conducted to assess human preference between images generated with our approach or with two other State-of-the-Art ones (namely, MultiDiffusion [1] and SyncDiffusion [3]). For each ap-

proach, we generated 60 images for each of the six prompts used for the quantitative evaluation (for a total of 360 images per method). We then formed pairs of images with matching prompts and presented 60 random pairs to the users (plus 6 additional pairs as a vigilance task), asking them to select the image that they found to be "*the most coherent, and that conforms the most to the prompt*". As mentioned in the main paper, 139 out of 201 participants passed the vigilance task and provided a total of 7807 preference answers. In Fig. 3, we plot the overall and the per-prompt results. As we can see, our method is consistently preferred over the competitors.

4 Further comparison with Direct Inference

Here, we report further comparison on the variability between the generated images using MAD and SD-L. In particular, for each one of the six evaluation prompts, we exploit the same network used for computing the FID, KID, and mGIQA (an Inception-v3 trained on Imagenet) to extract features from square images generated with SD and embed them using t-SNE [5]. Then, we generate 512×5120 images with SD-L and MAD and embed all the squared crops using openTSNE [4, 7]. In Fig. 4, we report the embeddings, showing that images generated by MAD have more variability with respect to the images generated by the vanilla SD. Thus, MAD has increased inter-image variability compared to SD-L. For each prompt, we plot the SD embeddings using grey circles, the MAD embeddings using red dots, and the SD-L embeddings using blue dots. Darker areas indicate an overlap between the embeddings of MAD and SD-L images.

5 Combination with Attention Scaling

To adapt Stable Diffusion to the generation of larger images, a recent work has proposed to re-scale the attention operations in the U-Net [2] (which we refer to as Attn-S). We evaluate this approach both in combination with a method that generates the whole image directly and MAD (we refer to this combination as MAD-L+Attn-S). To this end, we consider the generation of 512×3072 images. The result of generation with LDMs and LCMs are reported in Table 4 and Table 5, respectively. It can be noticed that the combination with the Attn-S has limited to no benefit in terms of generation performance. This can be explained by looking at the small differences in terms of qualitative results, reported in Figure 5.

6 Further Results at Different Aspect Ratios

We report a qualitative comparison with the State-of-the-Art approaches (MultiDiffusion [1] and SyncDiffusion [3]) on the generation of images with different aspect ratios. In particular, we consider the generation of horizontal images with the LDM (Fig. 7) and with the LCM (Fig. 8) and of vertical images with the same models (Fig. 9 and Fig. 10, respectively).

7 Further Qualitative Results

We report some randomly picked qualitative results on the GPT1k prompts and the six prompts used for the quantitative analysis of the LDM in Figs. 11, 12 and 15 to 20 and of the LCM in Figs. 13, 14 and 21 to 26. Moreover, in Figs. 27 to 30, we report more qualitative results, both horizontal and vertical, on different GPT1k prompts for the LDM and LCM, respectively.

8 Plug&Play Applications

Note that our MAD operator can be used as it is in settings such as tight and rough region-based generation, and conditional image generation (with Canny edge map guidance). since it is applied to the attention layers of the noise prediction UNet. The visual quality of the resulting images, shown in Figure 31, is on par with that of the generated panoramas (which is our main focus), even for guidance spanning multiple views of the large image.

9 Further Discussion of the Limitations

Approaches for panorama image generation that adapt diffusion models pretrained on squared images, including ours, can be limited by the performance and the image distribution learned by the base model. This results in generated images whose quality depends on the backbone, as showcased in Fig. 32. Moreover, inference-time panorama generation approaches struggle with scenes or objects that do not fit well with the specified aspect ratio and prompts where the base model itself does not produce good-quality results. We provide examples in Figs. 33 and 34. As we can see, the prompt *A gothic cathedral nave* (Fig. 34) does not fit well with the horizontal aspect ratio. However, when asked to generate a vertical image, the results improve. Moreover, the prompt *A fancy living room*, which entails a usually small indoor space, is rendered at an inferior quality when the desired output width is too large.

Table 1: Quantitative results on the generation of images with MAD applied at different stages of the LDM noise prediction model to generate 512×3072 images.

	mCLIP↑	**I-LPIPS↓**	**I-StyleL↓** ($\times 10^{-3}$)	**FID↓**	**KID↓** ($\times 10^{-3}$)	**mGIQA↑** ($\times 10^{-3}$)
No MAD (Baseline)	31.65±2.17	0.64±0.10	2.65±2.33	34.51±13.92	9.19±4.51	27.59±6.83
MAD in Mid Block	31.66±2.16	0.64±0.10	2.19±1.17	34.80±13.80	9.35±3.91	28.04±7.23
MAD in Down Blocks	31.83±2.21	0.62±0.10	1.98±1.03	40.46±16.58	16.40±8.83	28.11±7.42
MAD in Up Blocks	32.09±2.20	0.53±0.10	1.19±0.86	60.84±24.65	43.56±19.82	28.60±7.51
MAD in All Blocks	32.15±2.25	0.53±0.10	1.43±1.04	61.76±23.73	43.31±17.41	28.10±7.84

Table 2: Quantitative results on the generation of images with MAD applied at different stages of the LCM noise prediction model to generate 512×3072 images.

	mCLIP↑	**I-LPIPS↓**	**I-StyleL↓** ($\times 10^{-3}$)	**FID↓**	**KID↓** ($\times 10^{-3}$)	**mGIQA↑** ($\times 10^{-3}$)
No MAD (Baseline)	31.36±1.83	0.55±0.05	0.90±0.35	31.35±15.42	13.27±10.05	35.44±12.38
MAD in Mid Block	31.44±1.79	0.55±0.05	0.87±0.33	29.35±15.62	13.85±11.27	35.49±12.22
MAD in Down Blocks	31.44±1.78	0.55±0.05	0.84±0.32	29.01±14.83	13.35±11.02	35.38±12.46
MAD in Up Blocks	31.46±1.84	0.51±0.05	0.71±0.26	40.22±20.99	28.98±17.43	35.45±13.08
MAD in All Blocks	31.48±1.87	0.52±0.05	0.71±0.26	38.77±19.85	27.41±16.19	35.23±13.15

Table 3: Results on the generation of 512×3072 images with MAD applied for different numbers of timesteps from the beginning of the reverse diffusion process in the LDM.

	mCLIP↑	**I-LPIPS↓**	**I-StyleL↓** ($\times 10^{-3}$)	**FID↓**	**KID↓** ($\times 10^{-3}$)	**mGIQA↑** ($\times 10^{-3}$)
τ=**0**	31.65±2.17	0.64±0.10	2.65±2.33	34.51±13.92	9.19±4.51	27.59±6.83
τ=**5**	31.86±2.22	0.59±0.10	2.07±1.31	38.10±13.71	13.75±4.00	28.32±7.64
τ=**10**	31.95±2.24	0.57±0.10	2.00±1.34	42.60±14.94	19.38±5.56	28.36±7.83
τ=**15**	32.03±2.29	0.56±0.10	1.90±1.32	48.52±17.14	27.15±9.10	28.32±7.76
τ=**20**	32.09±2.29	0.54±0.10	1.68±1.19	54.31±19.92	34.43±12.54	28.23±7.90
τ=**25**	32.15±2.25	0.53±0.10	1.43±1.04	61.76±23.73	43.31±17.41	28.10±7.84
τ=**30**	32.16±2.17	0.52±0.10	1.22±0.90	70.29±28.24	54.40±23.19	27.94±7.73
τ=**35**	32.18±2.08	0.51±0.10	1.02±0.73	78.50±32.94	65.06±29.94	27.82±7.65
τ=**40**	32.16±1.95	0.50±0.10	0.88±0.61	86.20±37.23	76.15±37.73	27.59±7.61
τ=**45**	32.14±1.83	0.50±0.10	0.78±0.53	92.69±40.76	84.13±44.13	27.35±7.50
τ=**50**	32.14±1.72	0.49±0.10	0.71±0.47	98.01±43.64	91.51±49.95	27.05±7.28

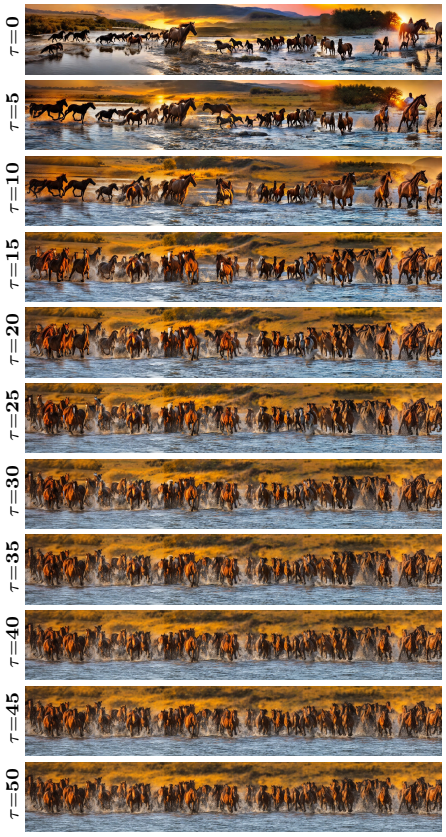

Fig. 1: Long images generated by the considered LDM with MAD applied up to different numbers of inference steps for the prompt *A herd of Mustang horses crossing a river at sunset*. When τ is too low, the view interactions are not enough to produce a globally coherent image. As τ increases, the image becomes more and more coherent, with maximal uniformity when MAD is applied at every timestep.

Fig. 2: Long images generated by the considered LDM with MAD applied in different blocks of the noise prediction model, with τ=15 for the prompt *A herd of Mustang horses crossing a river at sunset*.

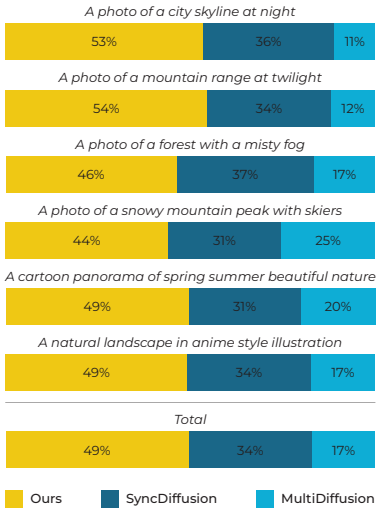

Fig. 3: User study per-prompt results of MAD, SyncDiffusion, and MultiDiffusion.

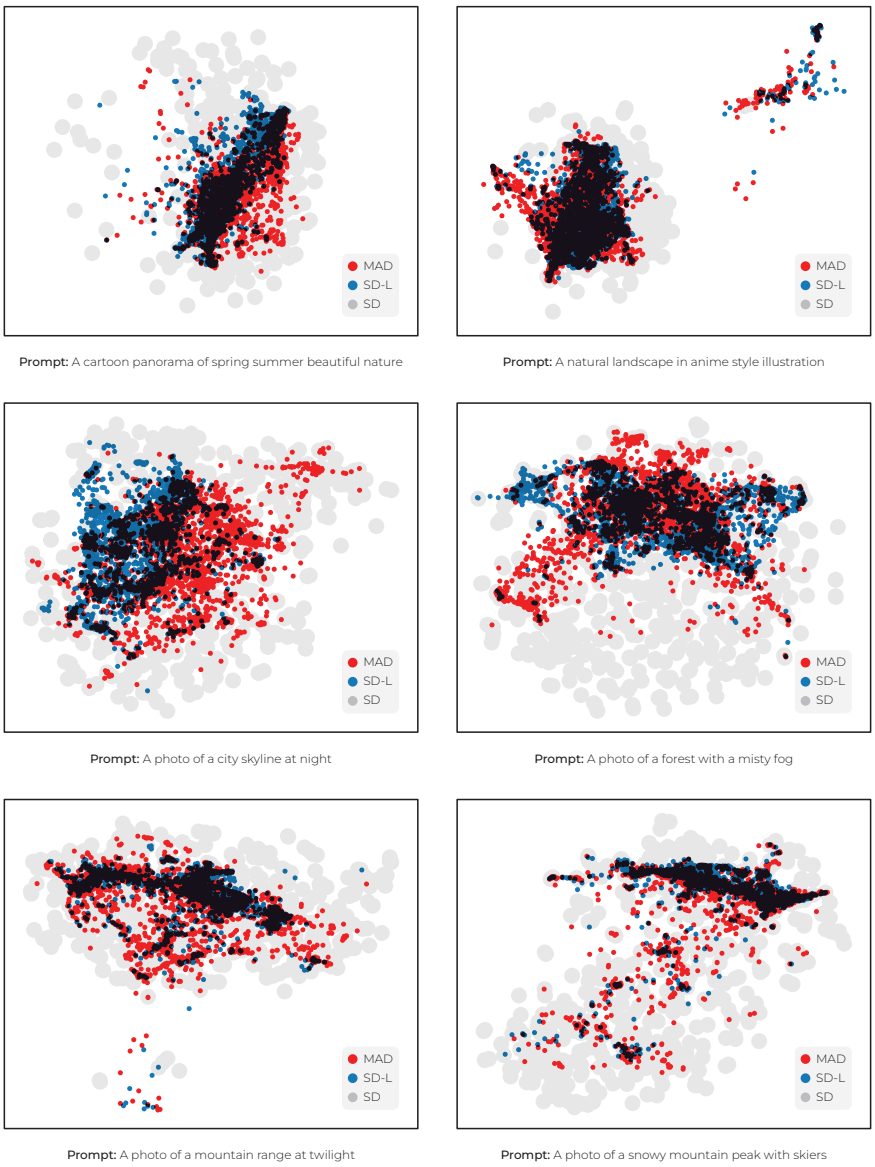

Fig. 4: Comparison on the distribution coverage of panoramas generated by MAD (red) and SD-L (blue) with respect to square images generated by SD (gray). Darker areas indicate an overlap between the embeddings of MAD and SD-L images.

Table 4: Quantitative comparison on 512×3072 panorama generation using the LDM. I-StyleL, KID, and mGIQA values are scaled by 10^3. For MAD, τ=15.

	mCLIP ↑	I-LPIPS ↓	I-StyleL ↓	FID ↓	KID ↓	mGIQA ↑
Standard Prompts						
SD-L	32.01±1.67	0.50±0.11	0.58±0.40	87.64±30.25	76.83±30.52	27.72±7.83
SD-L+Attn-S	32.02±1.66	0.52±0.11	0.73±0.47	80.16±27.11	67.54±26.00	27.89±7.67
MAD	32.03±2.29	0.56±0.10	1.90±1.32	48.52±17.14	27.15±9.10	28.32±7.76
MAD+Attn-S	32.17±2.20	0.53±0.11	1.19±0.75	64.08±28.19	47.33±23.77	28.30±0.75
GPT1k Prompts						
SD-L	31.89	0.52	0.73	68.03	6.61	12.58
SD-L+Attn-S	32.03	0.53	0.87	65.08	5.47	12.68
MAD	32.47	0.58	3.89	54.44	1.28	13.03
MAD+Attn-S	32.40	0.57	2.94	55.58	1.82	13.09

Table 5: Quantitative comparison on 512×3072 panorama generation with the LCM for different numbers of inference steps. I-StyleL, KID, and mGIQA are scaled by 10^3. For MAD, τ=1/1/2 for 1/2/4 inference steps, respectively.

	mCLIP ↑	I-LPIPS ↓	I-StyleL ↓	FID ↓	KID ↓	mGIQA ↑
1 Inference Step						
LCD-L	29.50±1.48	0.40±0.08	0.21±0.18	67.54±15.53	66.86±21.22	32.21±4.50
LCD-L+Attn-S	29.77±1.43	0.41±0.08	0.33±0.33	77.50±14.72	79.76±20.96	31.13±4.15
MAD	29.01±1.66	0.40±0.07	0.37±0.28	72.43±23.75	75.47±33.24	32.23±4.95
MAD+Attn-S	29.36±1.46	0.41±0.07	0.51±0.42	80.53±24.19	86.79±36.77	31.28±4.75
2 Inference Steps						
LCD-L	30.77±2.09	0.47±0.06	0.56±0.29	49.98±25.20	44.29±26.88	35.45±11.94
LCD-L+Attn-S	31.11±1.95	0.49±0.06	0.66±0.32	56.37±28.61	54.55±33.00	34.94±12.09
MAD	30.97±2.15	0.50±0.06	0.85±0.34	35.69±17.88	23.74±15.82	35.32±11.49
MAD+Attn-S	31.16±2.07	0.50±0.06	0.81±0.31	36.84±18.15	25.83±16.26	35.68±11.60
4 Inference Steps						
LCD-L	31.30±1.63	0.50±0.06	0.58±0.24	55.52±32.82	51.71±37.79	35.44±12.94
LCD-L+Attn-S	31.60±1.57	0.51±0.06	0.72±0.27	59.25±32.32	57.57±37.91	34.92±13.27
MAD	31.48±1.87	0.52±0.05	0.71±0.26	38.77±19.85	27.41±16.19	35.23±13.15
MAD+Attn-S	31.64±1.87	0.51±0.06	0.68±0.23	38.68±19.10	26.49±14.77	35.45±13.27

Fig. 5: Effect of attention scaling (AS) on LDMs used directly and with MAD.

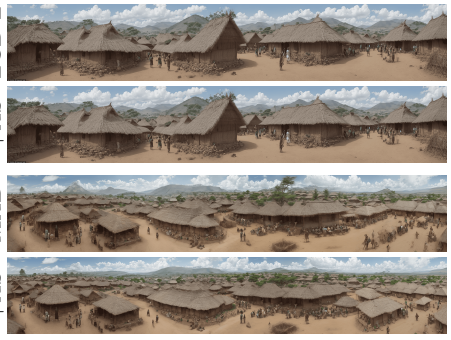

Fig. 6: Effect of attention scaling (AS) on LCMs used directly and with MAD.

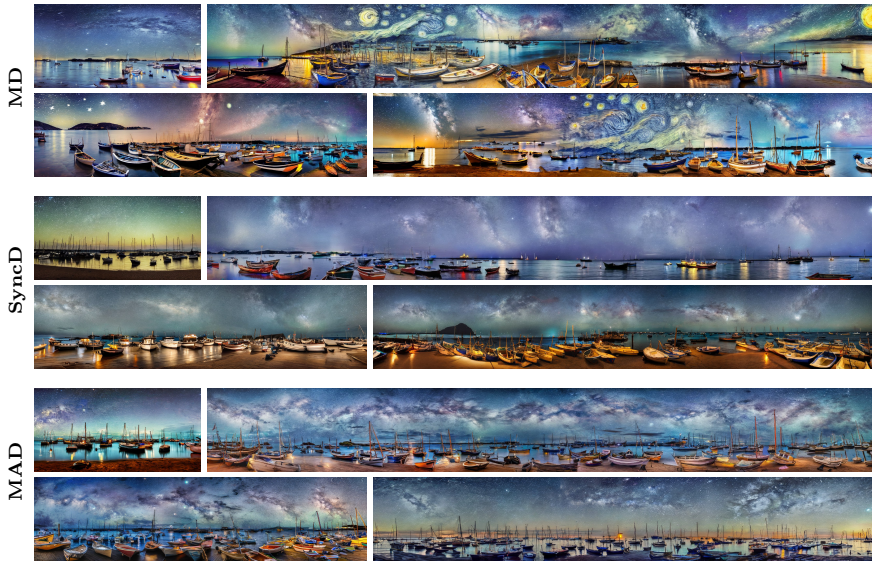

Fig. 7: Qualitative comparisons with MD and SyncD on the generation of horizontal images at different aspect ratios with the LDM model and the prompt *A photo of the seaside with some boats in a starry night.* These images are 512×1024 (top-left), 512×4096 (top-right), 512×2048 (bottom-left), and 512×3072 (bottom-right).

Fig. 8: Qualitative comparisons with respect to a baseline on the generation of horizontal images at different aspect ratios with the LCM model and the prompt *Top-view of a squared long pizza.* The images are 512×1024 (top-left), 512×4096 (top-right), 512×2048 (bottom-left), and 512×3072 (bottom-right).

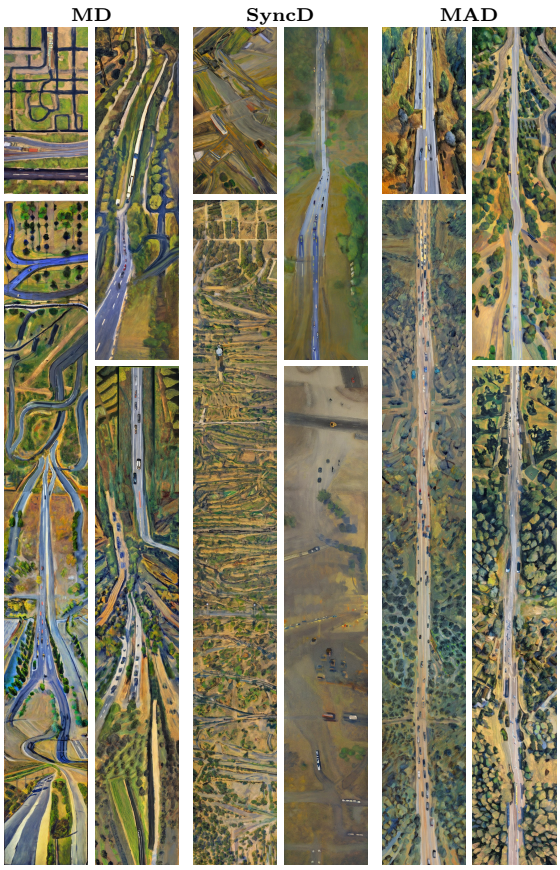

Fig. 9: Qualitative comparisons with MD and SyncD on the generation of vertical images at different aspect ratios with the LDM model and the prompt *Top view of a long road with the Manet style*. The images are 1024×512 (top-left), 2048×512 (top-right), 4096×512 (bottom-left), and 3072×512 (bottom-right).

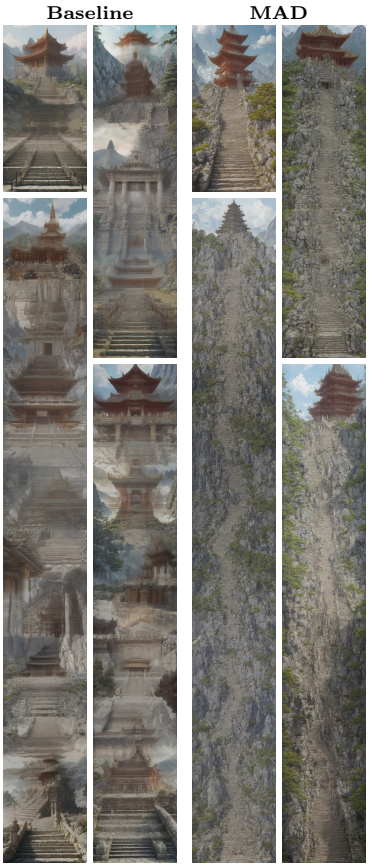

Fig. 10: Qualitative comparisons with respect to a baseline on the generation of vertical images at different aspect ratios with the LCM model and the prompt *Stairs on a mountain with a temple on top*. The images are 1024×512 (top-left), 2048×512 (top-right), 4096×512 (bottom-left), and 3072×512 (bottom-right).

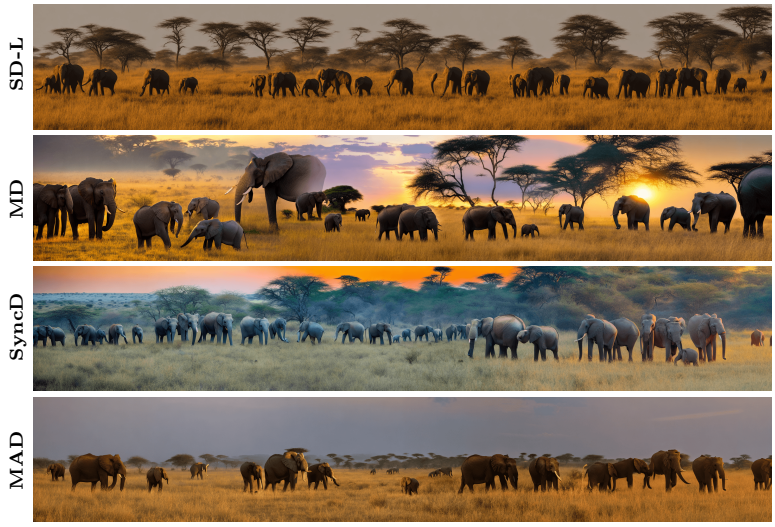

Fig. 11: Qualitative comparisons with respect to SD-L, MD, and SyncD for the prompt *A group of elephants grazing on the Savannah at sunset* from GPT1k, using the LDM model.

Fig. 12: Qualitative comparisons with respect to SD-L, MD, and SyncD for the prompt *A line of beach chairs facing the ocean* from GPT1k, using the LDM model.

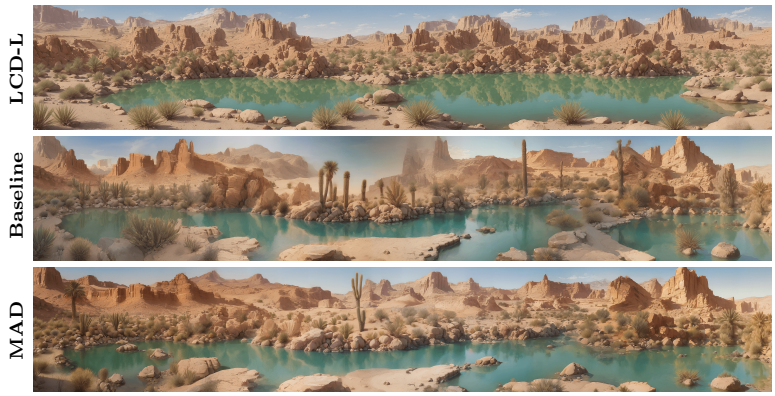

Fig. 13: Qualitative comparisons with respect to LCD-L, MD, and SyncD for the prompt *A tranquil desert oasis* from GPT1k, using the LCM model.

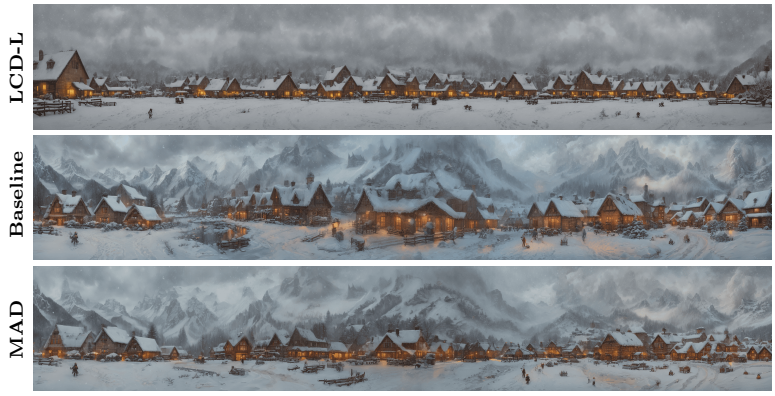

Fig. 14: Qualitative comparisons with respect to LCD-L and a baseline for the prompt *A snowstorm blanketing a small village* from GPT1k, using the LCM model.

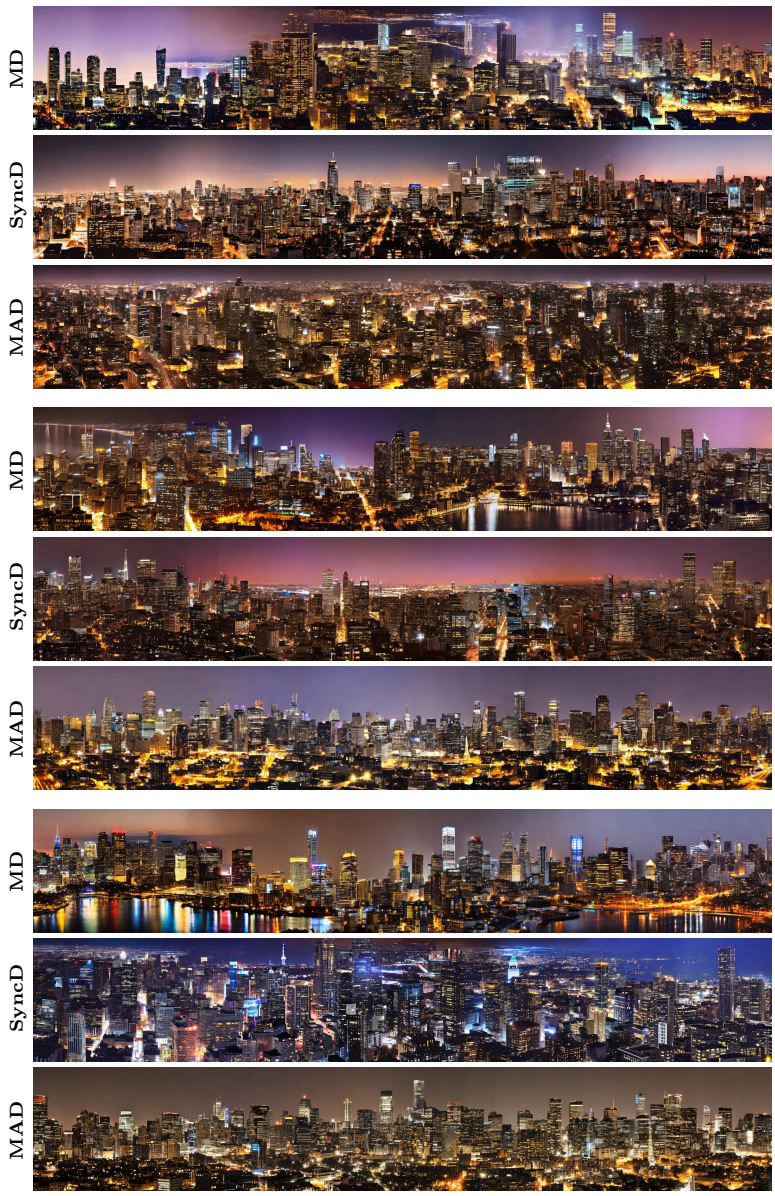

Fig. 15: Qualitative comparisons with respect to MD and SyncD for the prompt *A photo of a city skyline at night*, using the LDM model.

Fig. 16: Qualitative comparisons with respect to MD and SyncD for the prompt *A photo of a mountain range at twilight*, using the LDM model.

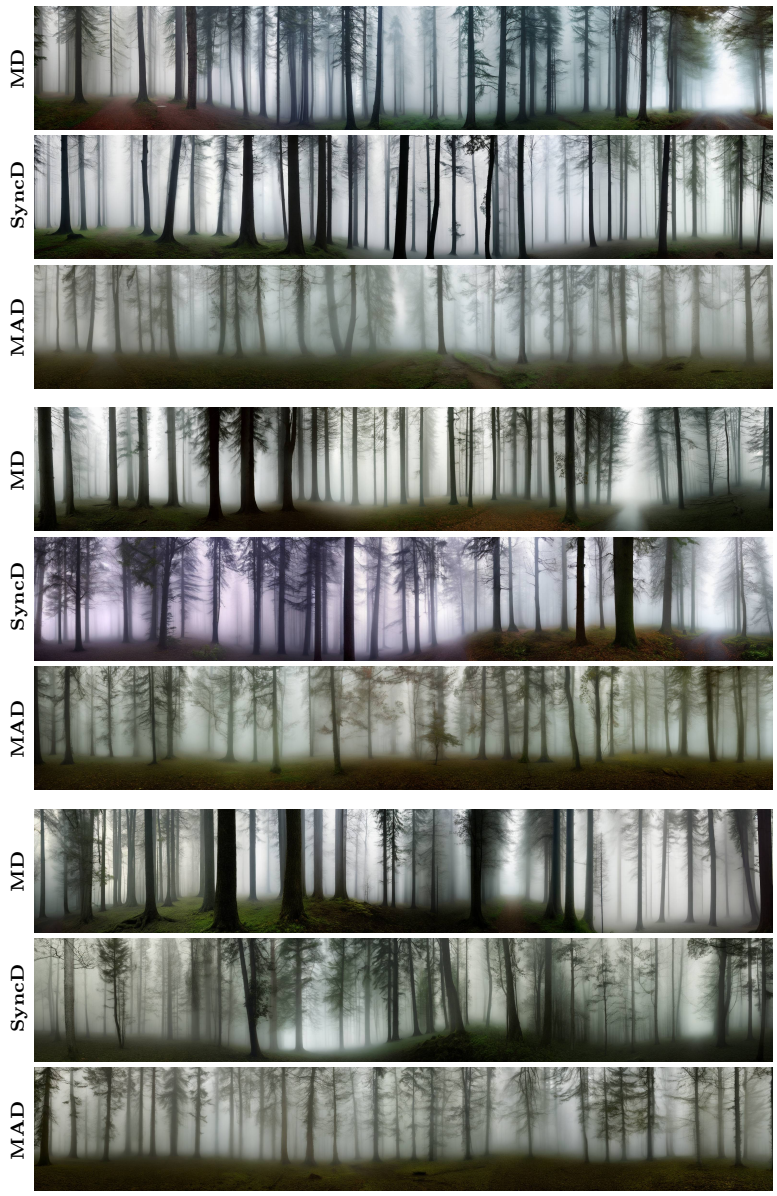

Fig. 17: Qualitative comparisons with respect to MD and SyncD for the prompt *A photo of a forest with a misty fog*, using the LDM model.

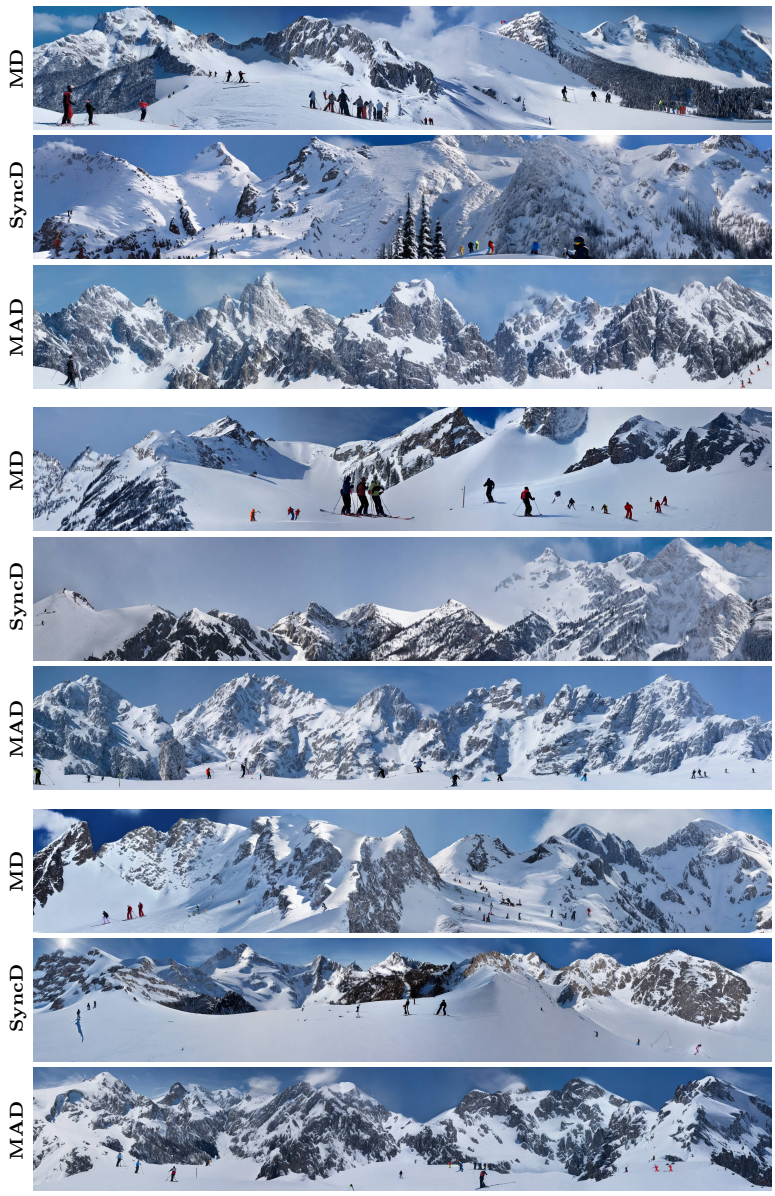

Fig. 18: Qualitative comparisons with respect to MD and SyncD for the prompt *A photo of a snowy mountain peak with skiers*, using the LDM model.

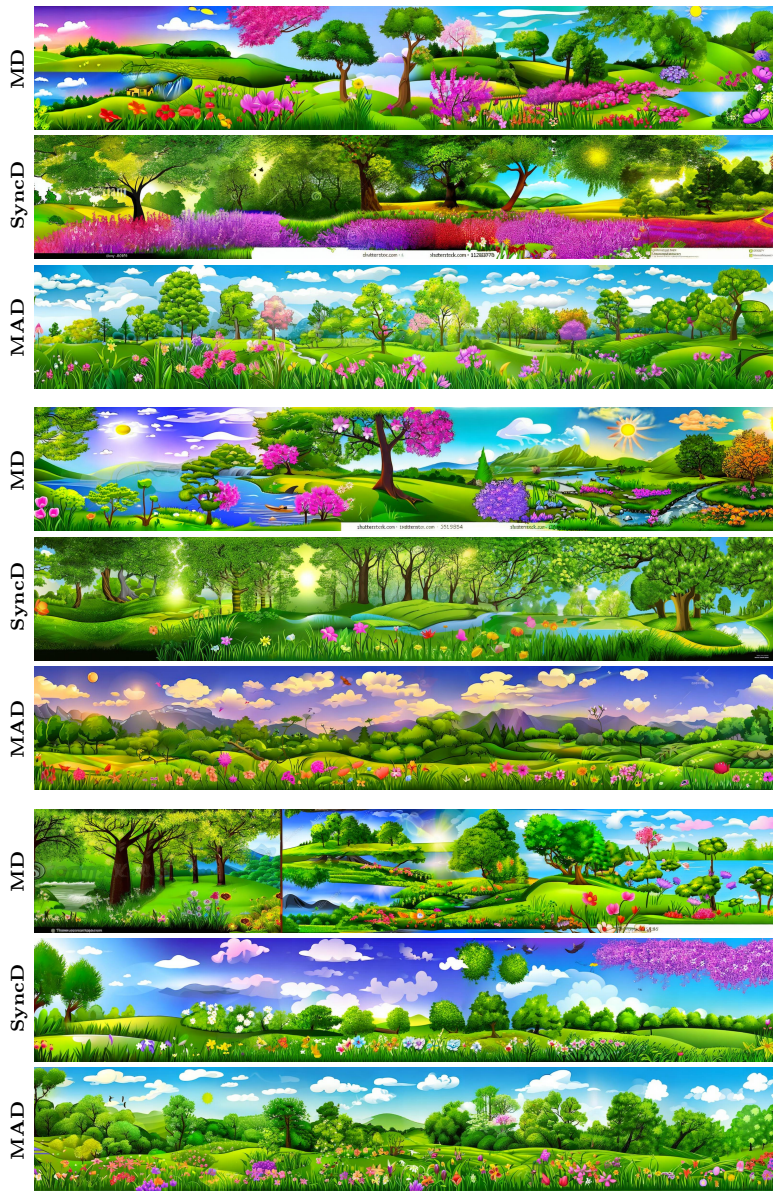

Fig. 19: Qualitative comparisons with respect to MD and SyncD for the prompt *A cartoon panorama of spring summer beautiful nature*, using the LDM model.

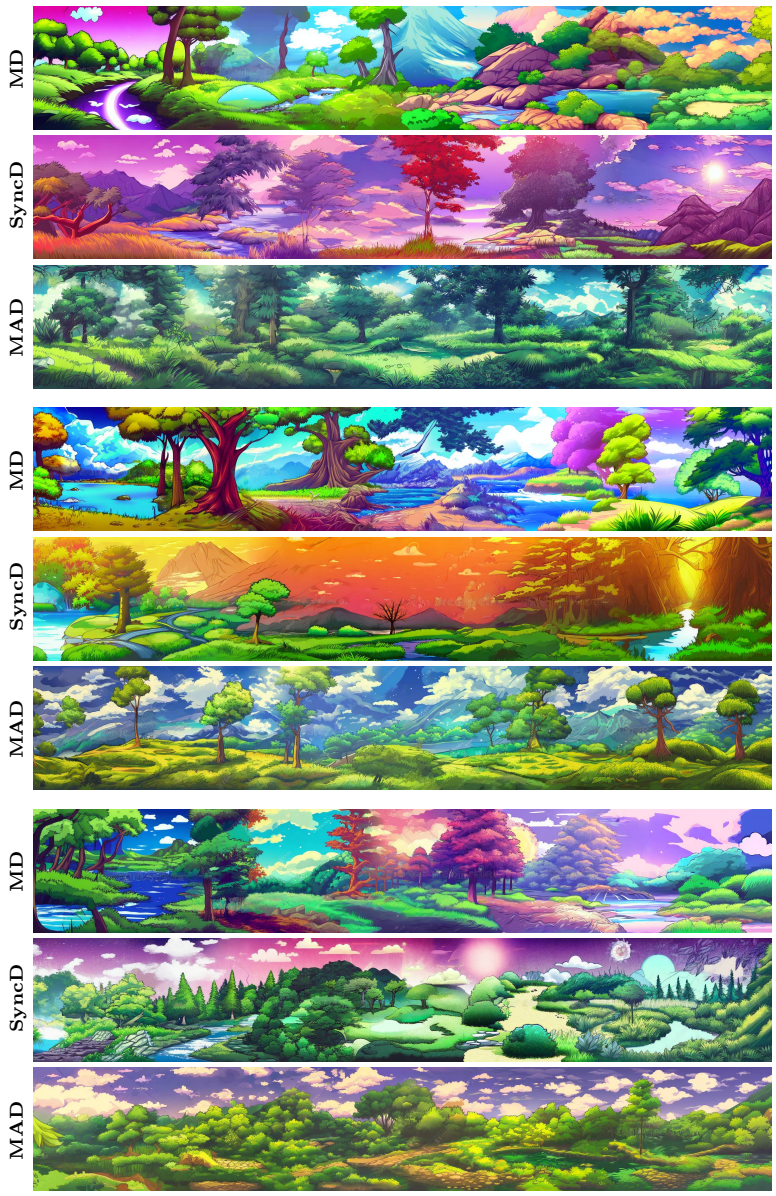

Fig. 20: Qualitative comparisons with respect to MD and SyncD for the prompt *A natural landscape in anime style illustration*, using the LDM model.

Fig. 21: Qualitative comparisons with respect to a baseline for the prompt *A photo of a city skyline at night*, using the LCM model.

Fig. 22: Qualitative comparisons with respect to a baseline for the prompt *A photo of a mountain range at twilight*, using the LCM model.

Fig. 23: Qualitative comparisons with respect to a baseline for the prompt *A photo of a forest with a misty fog*, using the LCM model.

Fig. 24: Qualitative comparisons with respect to a baseline for the prompt *A photo of a snowy mountain peak with skiers*, using the LCM model.

Fig. 25: Qualitative comparisons with respect to a baseline for the prompt *A cartoon panorama of spring summer beautiful nature*, using the LCM model.

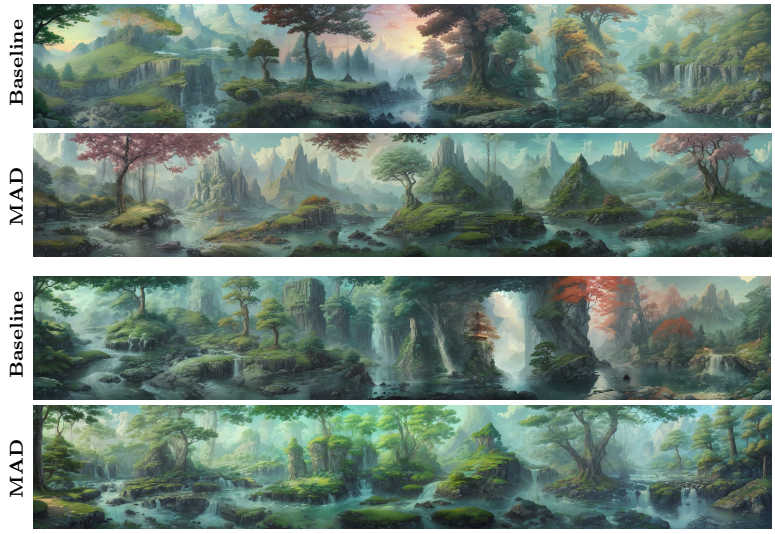

Fig. 26: Qualitative comparisons with respect to a baseline for the prompt *A natural landscape in anime style illustration*, using the LCM model.

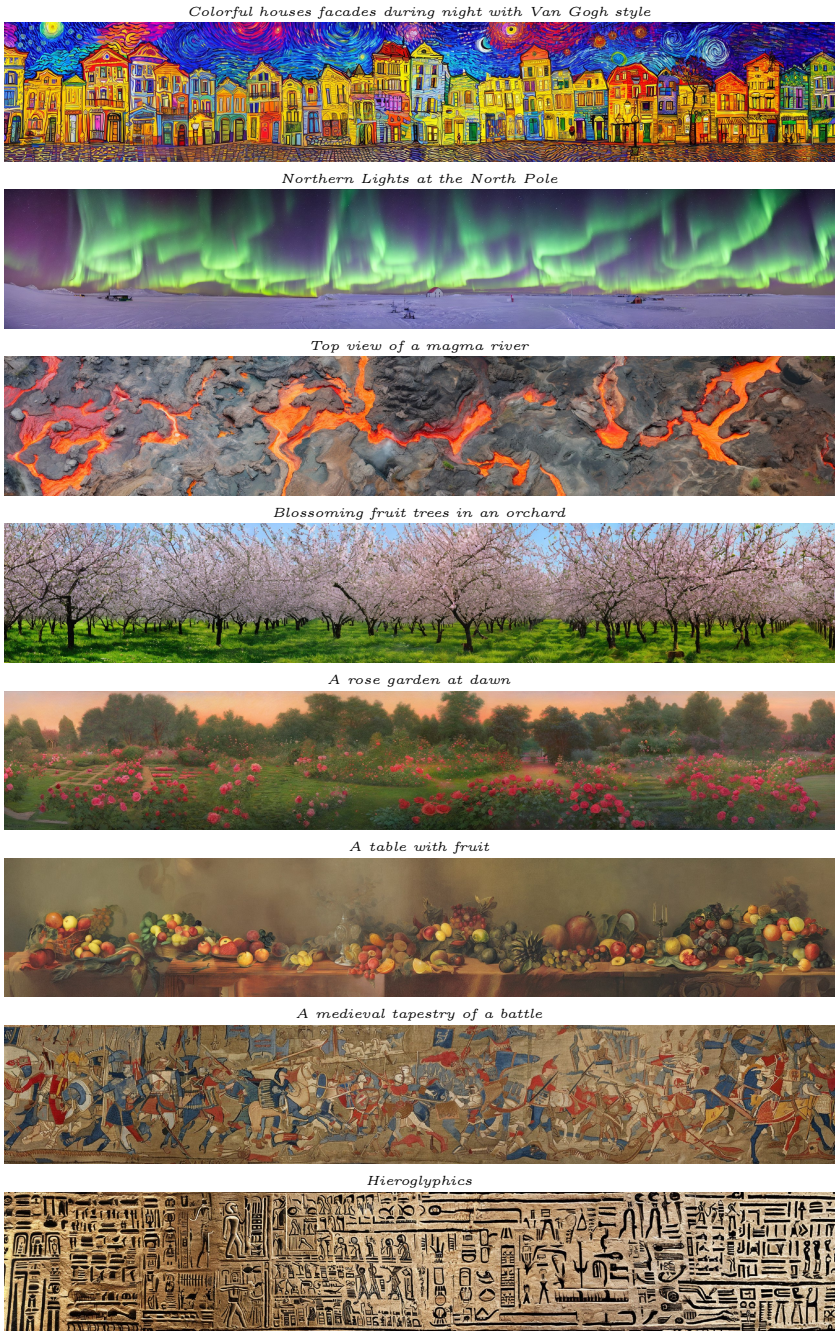

Fig. 27: Horizontal panorama images with various prompts and the LDM model.

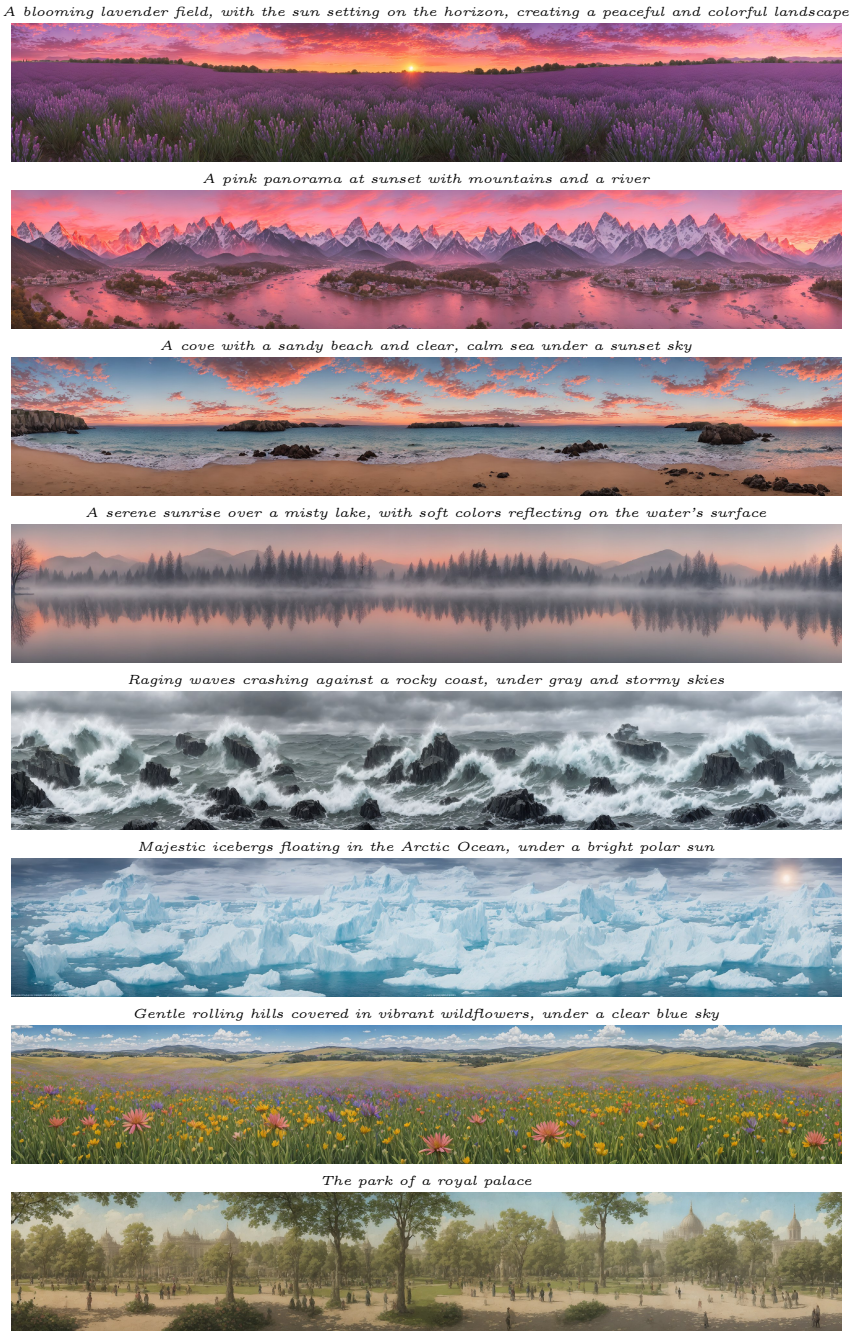

Fig. 28: Horizontal panorama images with various prompts and the LCM model.

Fig. 29: Vertical panorama images with various prompts and the LDM model.

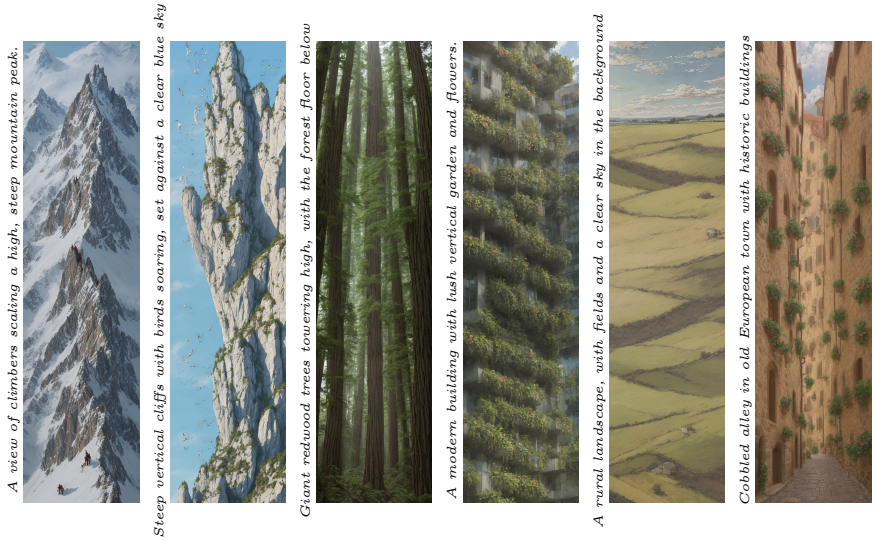

Fig. 30: Vertical panorama images with various prompts and the LCM model.

Fig. 31: Plug&Play applications of MAD.

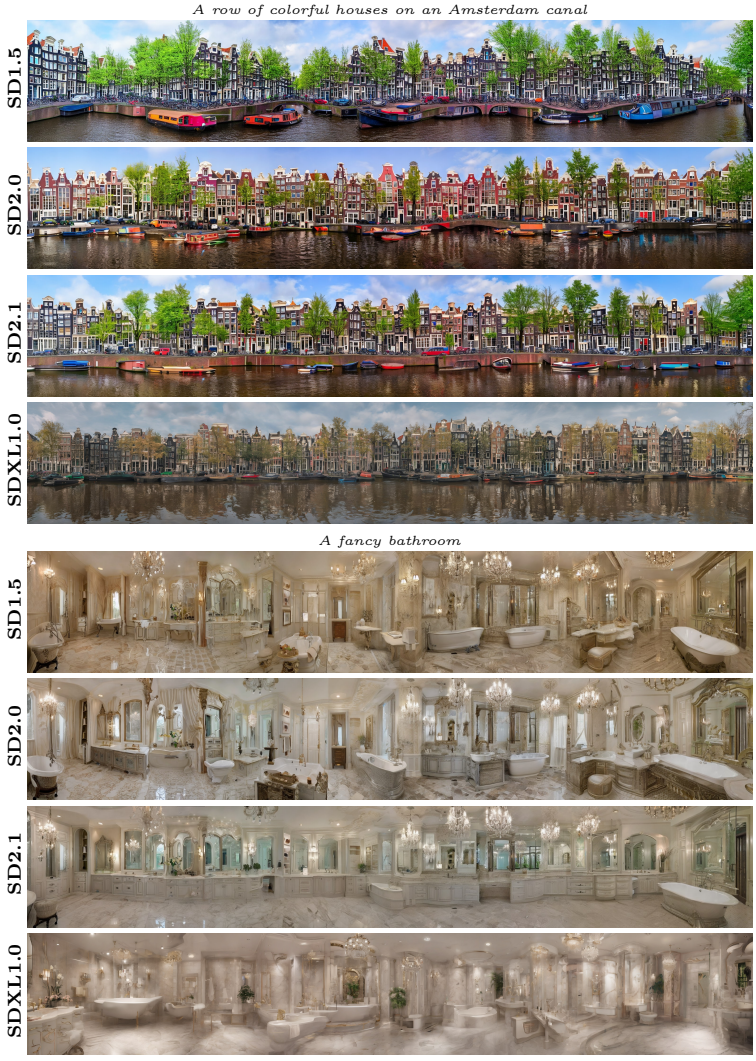

Fig. 32: Images generated with MAD applied to different LDM backbones and resolution 512×3072. For SDXL1.0 [6], the resolution is 1024×6144.

Fig. 33: Qualitative comparison between SD-L, SD-L+AttnS, MD, SyncD, and MAD on the generation of images with a prompt that does not fit well into all the specified aspect ratios (top: 512×3072, bottom-left: 512×2048, bottom-right: 512×1024). We use the LDM as backbone, the same seed, and the prompt *A fancy living room*.

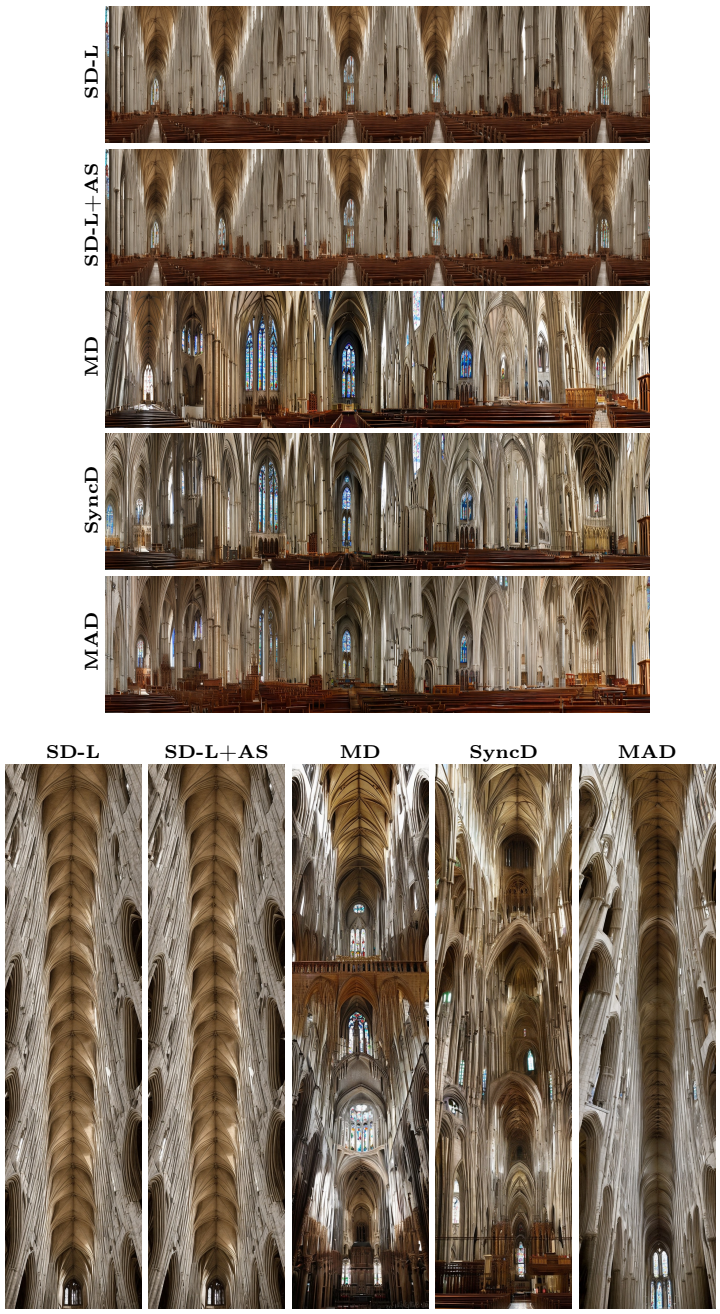

Fig. 34: Horizontal and vertical images generated with the same seed and for the prompt *A gothic cathedral nave* using different methods applied on the LDM model: direct inference (SD-L), Attention Scaling (SD-L+AS), MD, SyncD, and MAD.

References

1. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. ICML (2023)
2. Jin, Z., Shen, X., Li, B., Xue, X.: Training-free Diffusion Model Adaptation for Variable-Sized Text-to-Image Synthesis. NeurIPS (2023)
3. Lee, Y., Kim, K., Kim, H., Sung, M.: Syncdiffusion: Coherent montage via Synchronized Joint Diffusions. In: NeurIPS (2023)
4. Linderman, G.C., Rachh, M., Hoskins, J.G., Steinerberger, S., Kluger, Y.: Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. Nat. Methods **16**(3), 243–245 (2019)
5. Van der Maaten, L., Hinton, G.: Visualizing Data using t-SNE. Journal of Machine Learning Research **9**(11) (2008)
6. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In: ICLR (2024)
7. Poličar, P.G., Stražar, M., Zupan, B.: openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding. BioRxiv (2019)