Cache Me if You Can:
Accelerating Diffusion Models through Block Caching

Supplementary Material

In this supplementary material, we provide

1. thoughts on future work in Sec. A
2. an overview of the limitations of our method in Sec. B
3. thoughts on ethical considerations and safety in Sec. C
4. additional figures for **qualitative results**, **change metric plots**, and **caching schedules** in Sec. D

A. Future Work

There are several directions for future work. First, we believe that the use of step-to-step change metrics is not limited to caching, but could / should also benefit e.g. finding a better network architecture or a better noise schedule. Secondly, we find that the effect of scale-shift adjustment can be quite significant on the overall structure and visual appeal of the image. It could be possible to use a similar technique for finetuning with human in the loop to make the model adhere more to the preference of the user without having to change the training data. Finally, it would be interesting if caching could be integrated into a network architecture even before training. This could not only improve the results of the final model, but also speed up training.

B. Limitations

While our method achieves good results, some noteworthy weaknesses remain. We observe that while the scale-shift adjustment improves results and reduces artifacts, it sometimes changes the identity of the image more than reducing the number of steps or using naive caching would. Furthermore, finding the perfect threshold for auto configuration can take time, as the model is sensitive to certain changes in the caching schedule. We recommend playing around with small variations of the desired threshold to obtain the perfect schedule.

C. Ethical Considerations & Safety

We do not introduce new image data to these model and the optimization scheme for scale-shift adjustment only requires prompts. Therefore, we believe that our technique does not introduce ethical or legal challenges beyond the model on which we apply our technique.

For safety considerations, it should be noted that scale-shift adjustment, while still following the prompt, can change the identities in the image slightly. This aspect might make an additional safety check necessary when deploying models with block caching.

D. Additional Figures

Additional Qualitative Results. We show additional results for all configurations mentioned in the main paper. For all configurations, we show our caching technique with and without scale-shift adjustment, a slower baseline with the same number of steps, and a baseline with the same latency as ours (by reducing the number of steps).

Additional Change Plots. For all above mentioned configurations, we show the step-to-step change per layer block averaged over 32 forward passes and two random seeds each measured via the $\mathrm{L1}_{\mathrm{rel}}$ metric. This corresponds to Fig. 2 b) in the main paper.

Additional Caching Schedules, Finally, we also show all the caching schedules, which are automatically derived from the change measurements mentioned above.

An overview of the figures is provided by Tab. 1

Model	*Steps*	*Solver*	*Quali.*	*Change*	*Schedule*
EMU-768	20 vs 14	DPM	Fig. 1	Fig. 9	Fig. 17
		DDIM	Fig. 2	Fig. 10	Fig. 18
	50 vs 30	DPM	Fig. 3	Fig. 11	Fig. 19
		DDIM	Fig. 4	Fig. 12	Fig. 20
LDM-512	20 vs 14	DPM	Fig. 5	Fig. 13	Fig. 21
		DDIM	Fig. 6	Fig. 14	Fig. 22
	50 vs 30	DPM	Fig. 7	Fig. 15	Fig. 23
		DDIM	Fig. 8	Fig. 16	Fig. 24

Table 1. **Additional Figures Overview.** *Quali.*: Qualitative results, *Change*: Change metric plots, *Schedule*: Chaching schedule

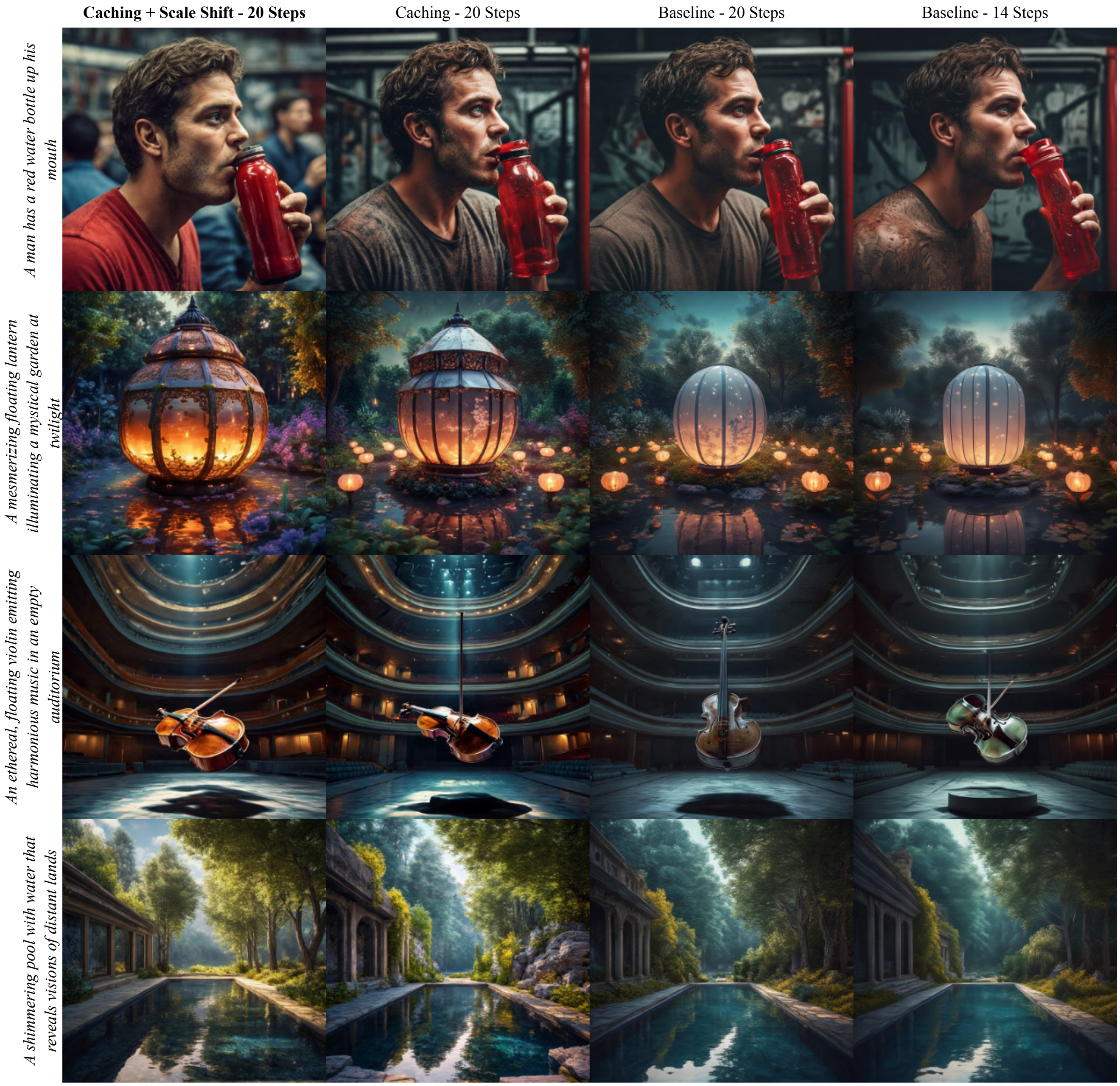
Caching + Scale Shift - 20 Steps
Caching - 20 Steps
Baseline - 20 Steps
Baseline - 14 Steps
A man has a red water bottle up his mouth
A mesmerizing floating lantern illuminating a mystical garden at twilight
An ethereal, floating violin emitting harmonious music in an empty auditorium
A shimmering pool with water that reveals visions of distant lands

Figure 1. **Qualitative Results for EMU-768 - DPM 20 Steps.**

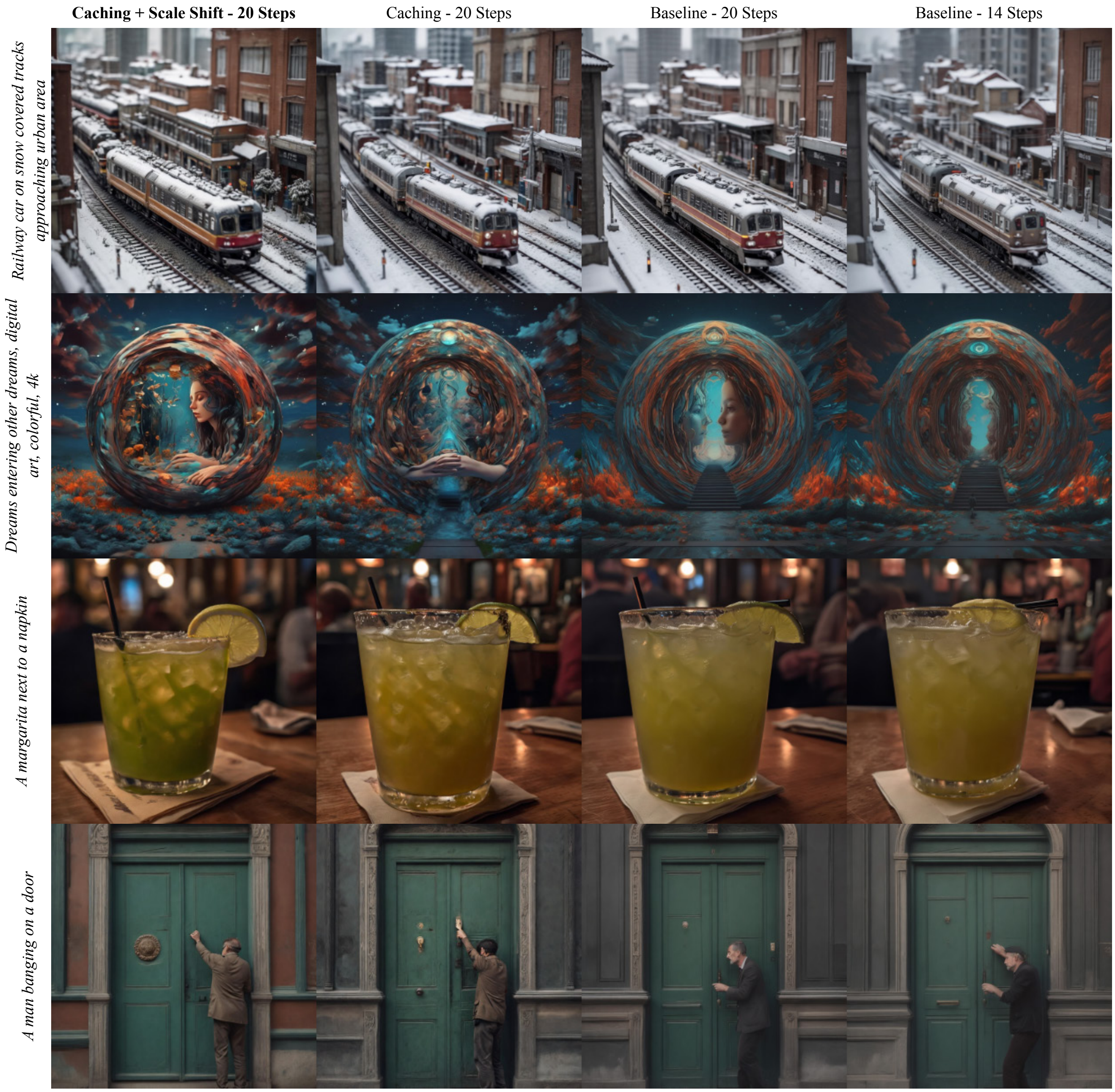
Caching + Scale Shift - 20 Steps
Caching - 20 Steps
Baseline - 20 Steps
Baseline - 14 Steps
Railway car on snow covered tracks approaching urban area
Dreams entering other dreams, digital art, colorful, 4k
A margarita next to a napkin
A man banging on a door

Figure 2. **Qualitative Results for EMU-768 - DDIM 20 Steps.**

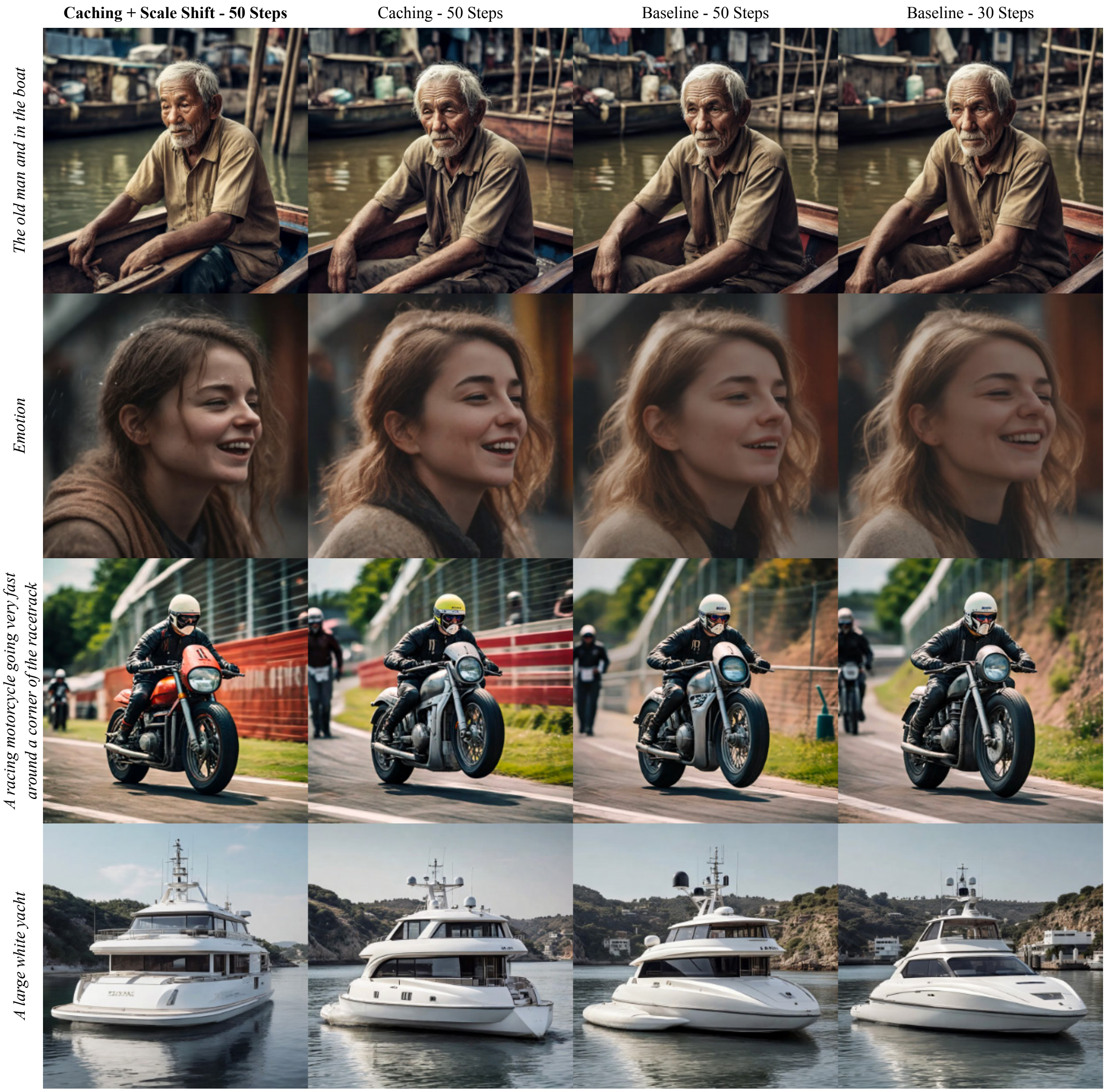
Caching + Scale Shift - 50 Steps
Caching - 50 Steps
Baseline - 50 Steps
Baseline - 30 Steps
The old man and in the boat
Emotion
A racing motorcycle going very fast around a corner of the racetrack
A large white yacht

Figure 3. **Qualitative Results for EMU-768 - DPM 50 Steps.**

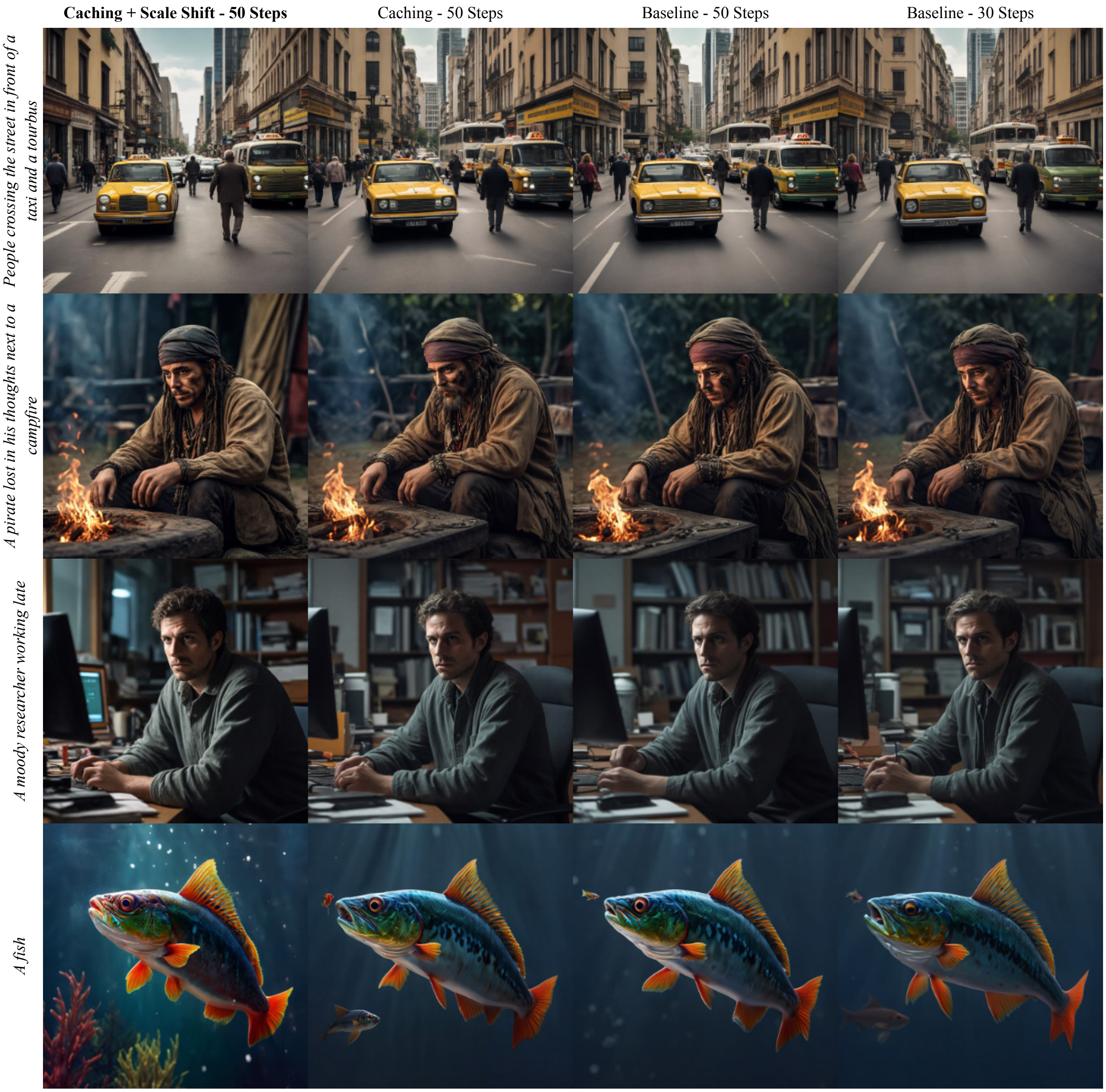
Caching + Scale Shift - 50 Steps
Caching - 50 Steps
Baseline - 50 Steps
Baseline - 30 Steps
People crossing the street in front of a taxi and a tourbus
A pirate lost in his thoughts next to a campfire
A moody researcher working late
A fish

Figure 4. **Qualitative Results for EMU-768 - DDIM 50 Steps.**

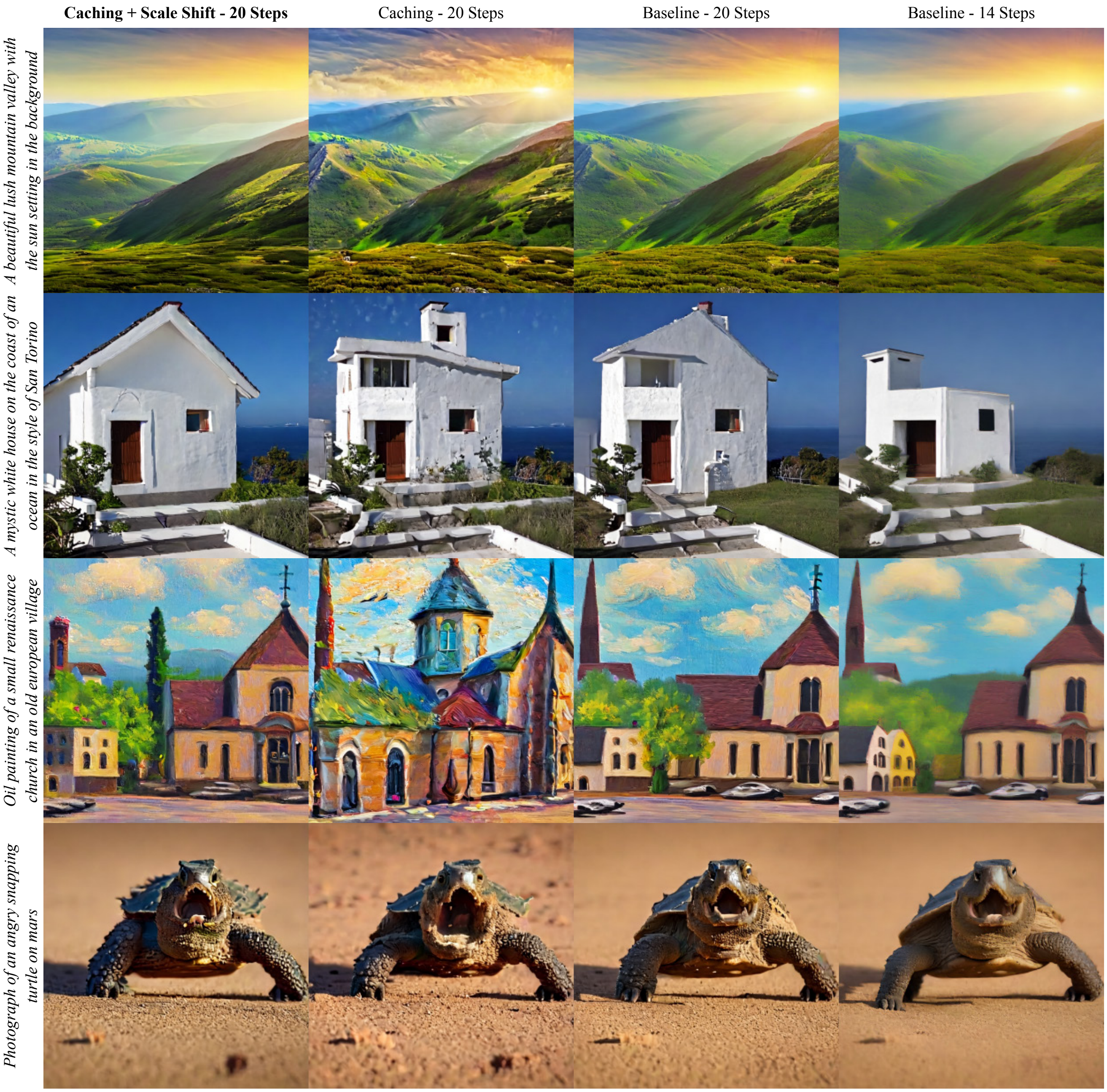
Caching + Scale Shift - 20 Steps
Caching - 20 Steps
Baseline - 20 Steps
Baseline - 14 Steps
A beautiful lush mountain valley with the sun setting in the background
A mystic white house on the coast of an ocean in the style of San Torino
Oil painting of a small renaissance church in an old european village
Photograph of an angry snapping turtle on mars

Figure 5. **Qualitative Results for LDM-512 - DPM 20 Steps.**

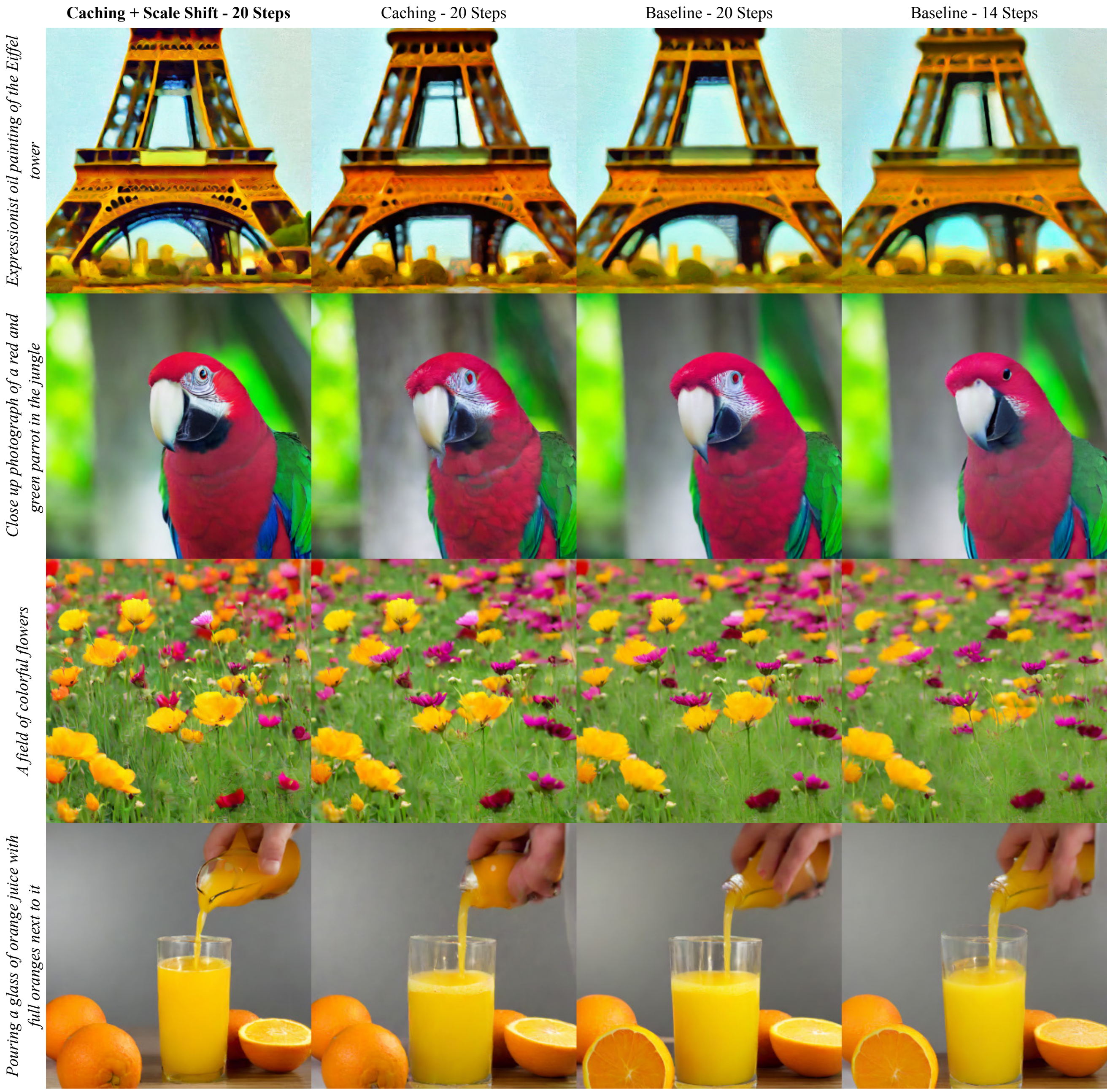
Caching + Scale Shift - 20 Steps
Caching - 20 Steps
Baseline - 20 Steps
Baseline - 14 Steps
Expressionist oil painting of the Eiffel tower
Close up photograph of a red and green parrot in the jungle
A field of colorful flowers
Pouring a glass of orange juice with full oranges next to it

Figure 6. **Qualitative Results for LDM-512 - DDIM 20 Steps.**

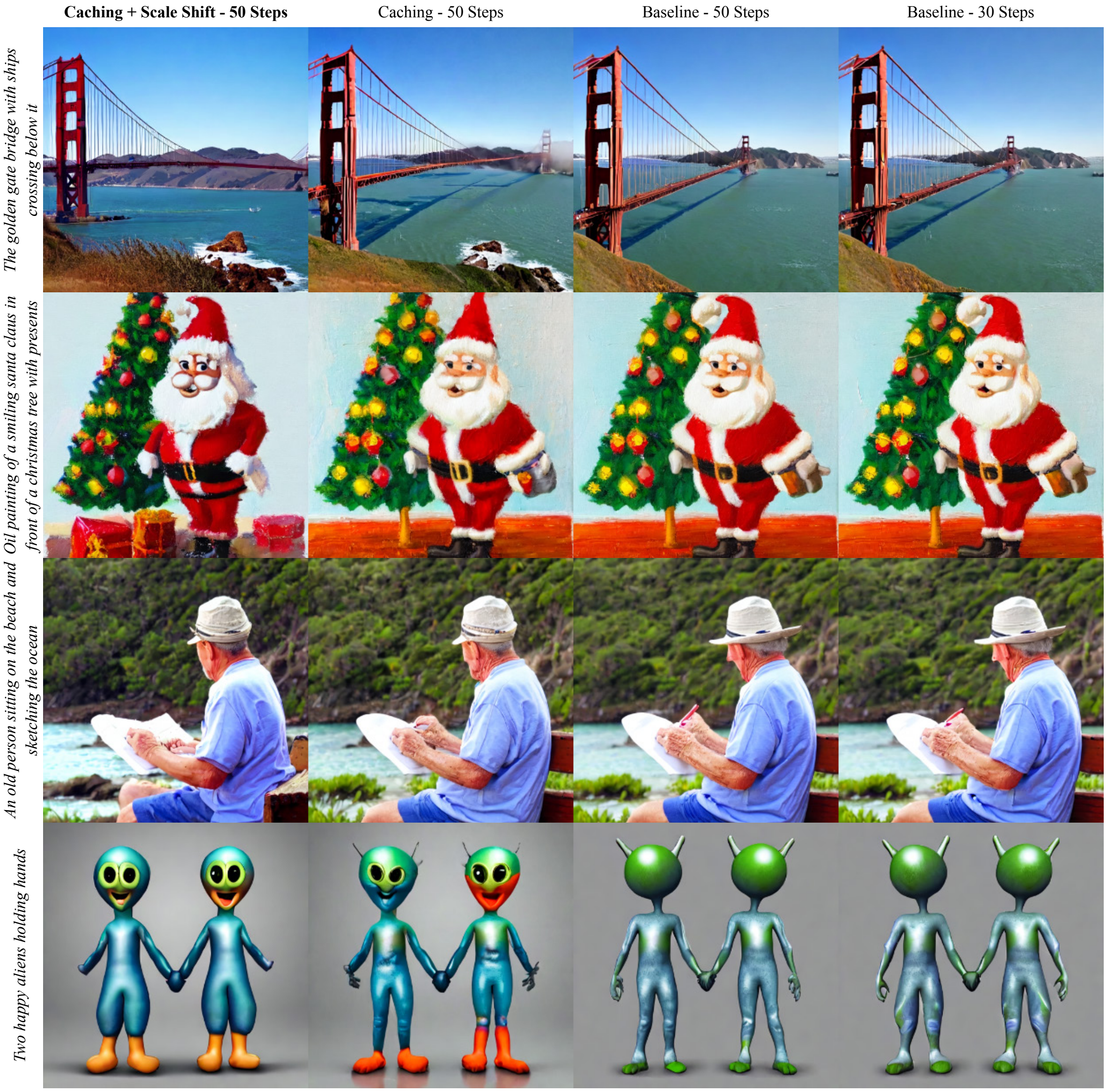
Caching + Scale Shift - 50 Steps
Caching - 50 Steps
Baseline - 50 Steps
Baseline - 30 Steps
The golden gate bridge with ships crossing below it
Oil painting of a smiling santa claus in front of a christmas tree with presents
An old person sitting on the beach and sketching the ocean
Two happy aliens holding hands

Figure 7. **Qualitative Results for LDM-512 - DPM 50 Steps.**

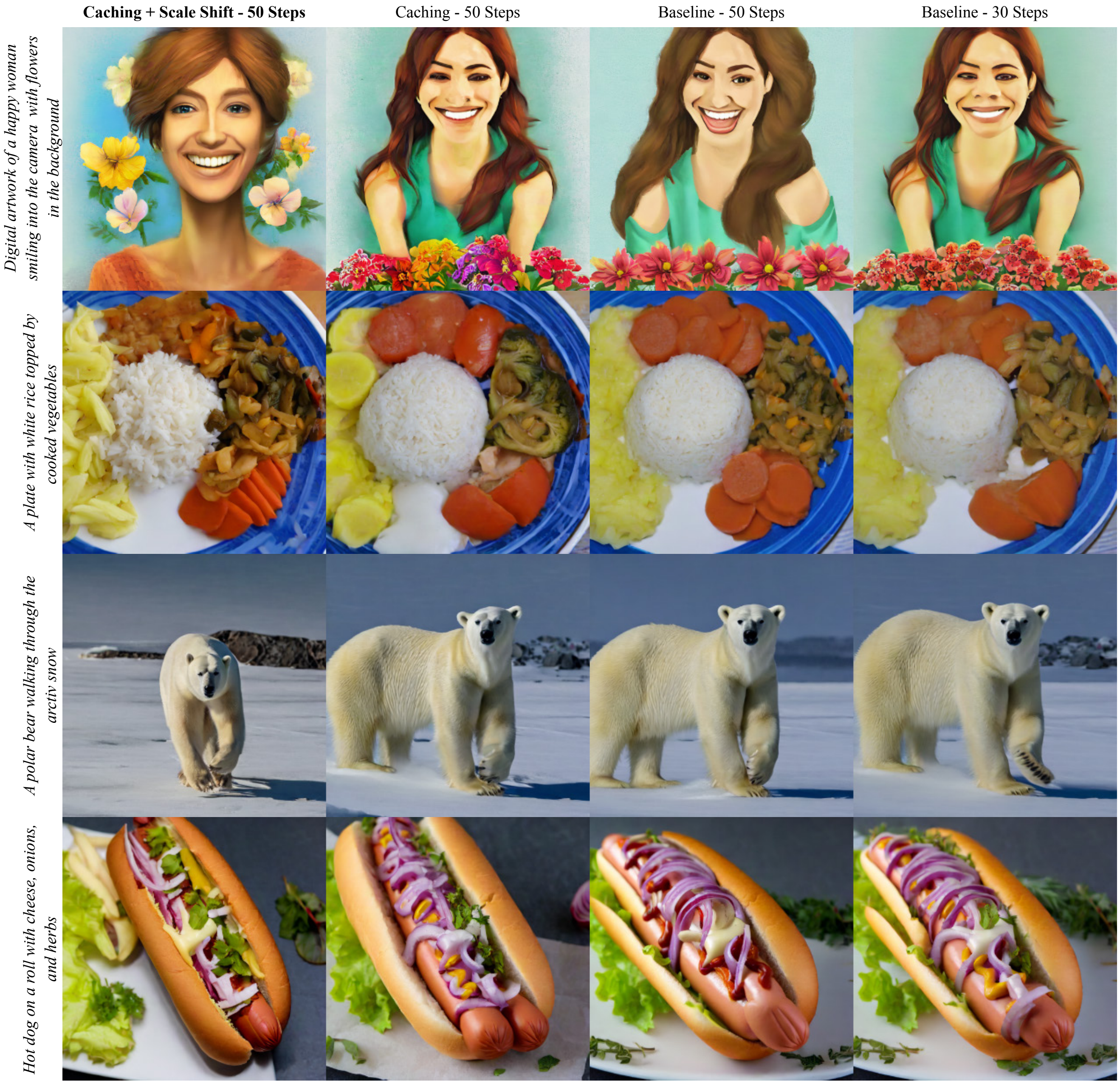
Caching + Scale Shift - 50 Steps
Caching - 50 Steps
Baseline - 50 Steps
Baseline - 30 Steps
Digital artwork of a happy woman smiling into the camera with flowers in the background
A plate with white rice topped by cooked vegetables
A polar bear walking through the arctiv snow
Hot dog on a roll with cheese, onions, and herbs

Figure 8. **Qualitative Results for LDM-512 - DDIM 50 Steps.**

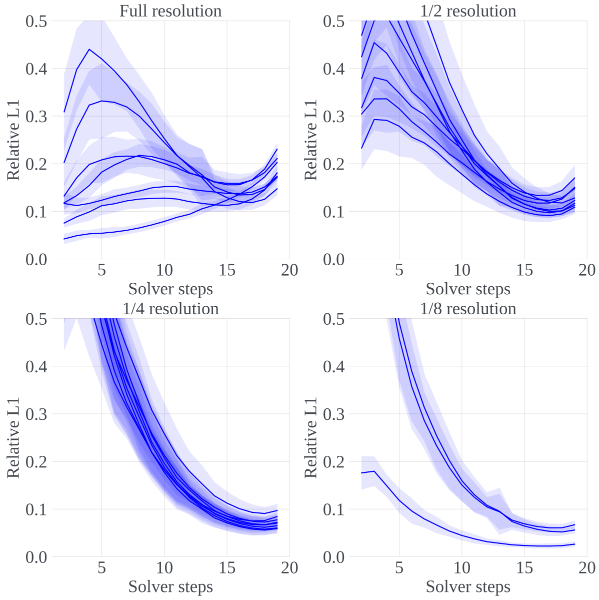

Figure 9. **Change Metrics for EMU-768 - DPM 20 Steps.**

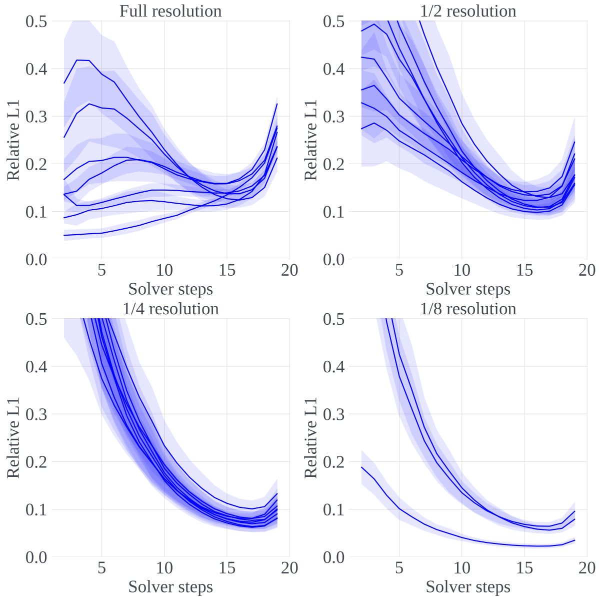

Figure 10. **Change Metrics for EMU-768 - DDIM 20 Steps.**

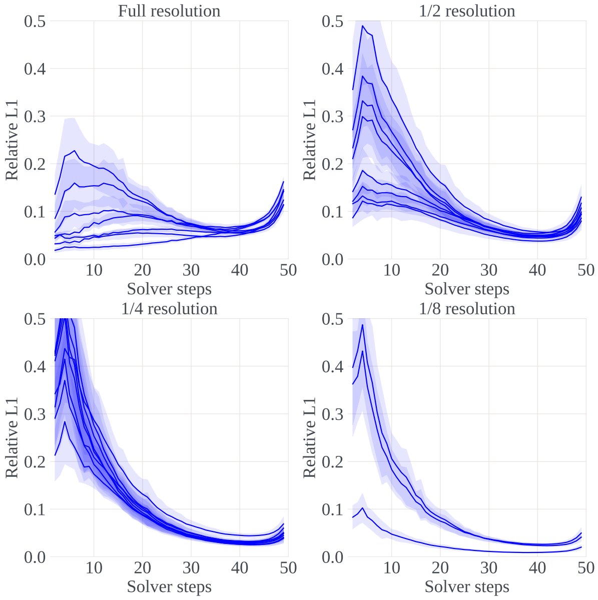

Figure 11. **Change Metrics for EMU-768 - DPM 50 Steps.**

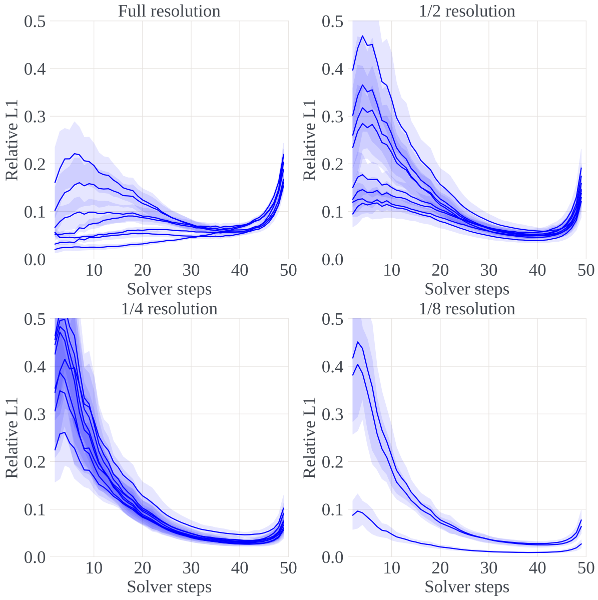

Figure 12. **Change Metrics for EMU-768 - DDIM 50 Steps.**

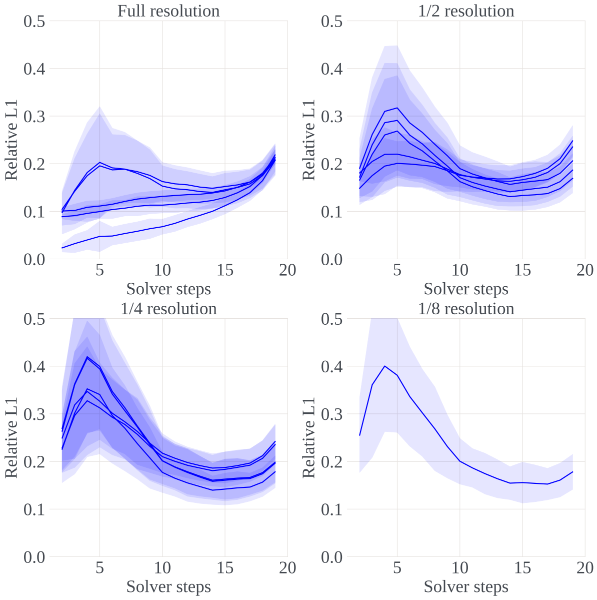

Figure 13. **Change Metrics for LDM-512 - DPM 20 Steps.**

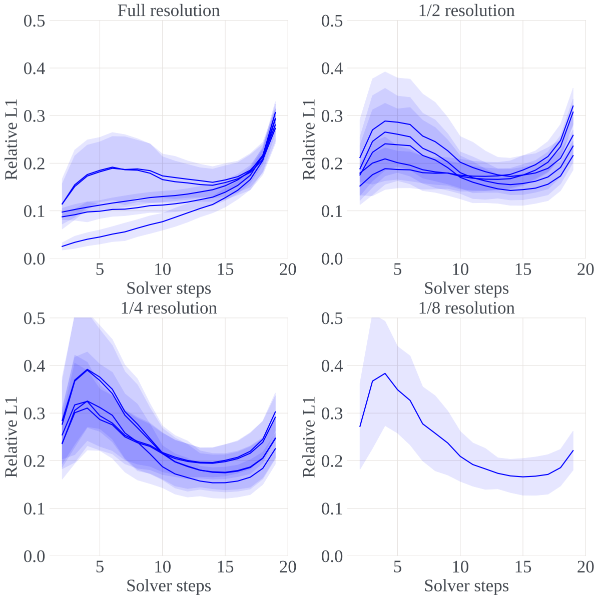

Figure 14. **Change Metrics for LDM-512 - DDIM 20 Steps.**

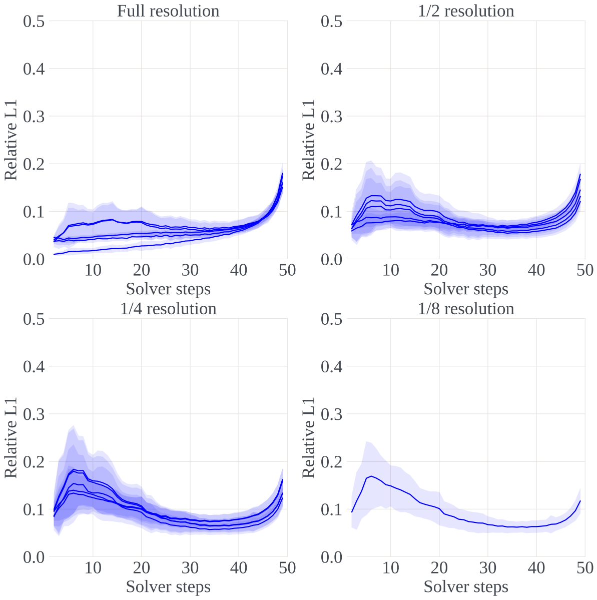

Figure 15. **Change Metrics for LDM-512 - DPM 50 Steps.**

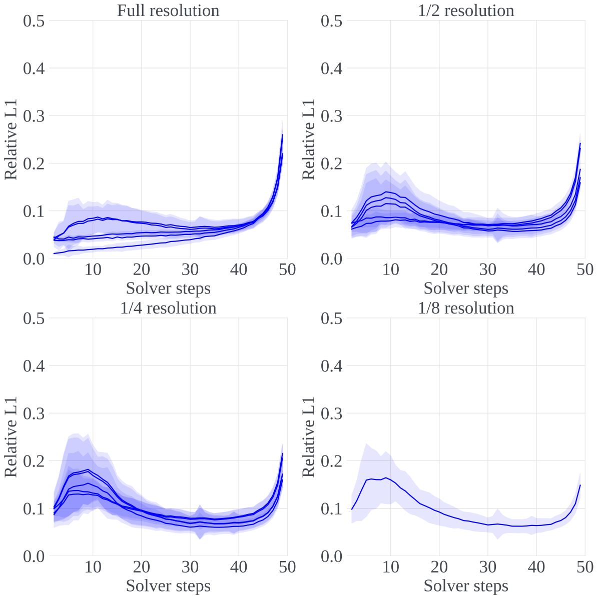

Figure 16. **Change Metrics for LDM-512 - DDIM 50 Steps.**

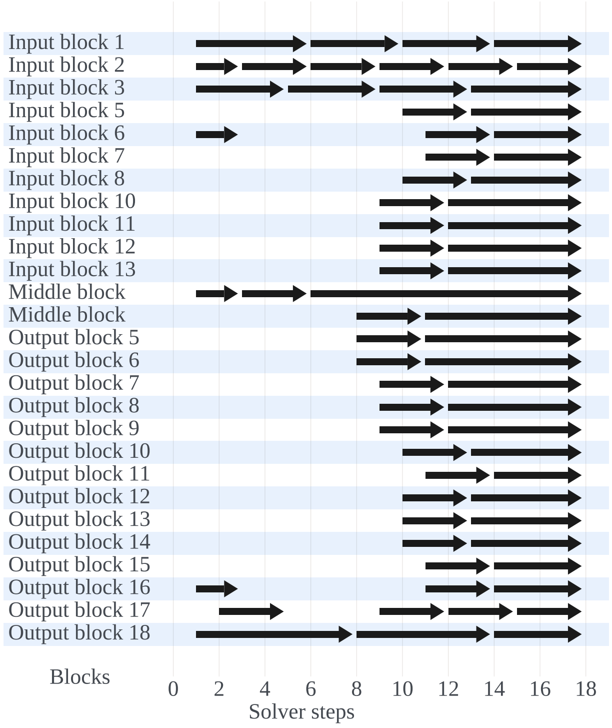

Figure 17. **Cache Schedules for EMU-768 - DPM 20 Steps.**

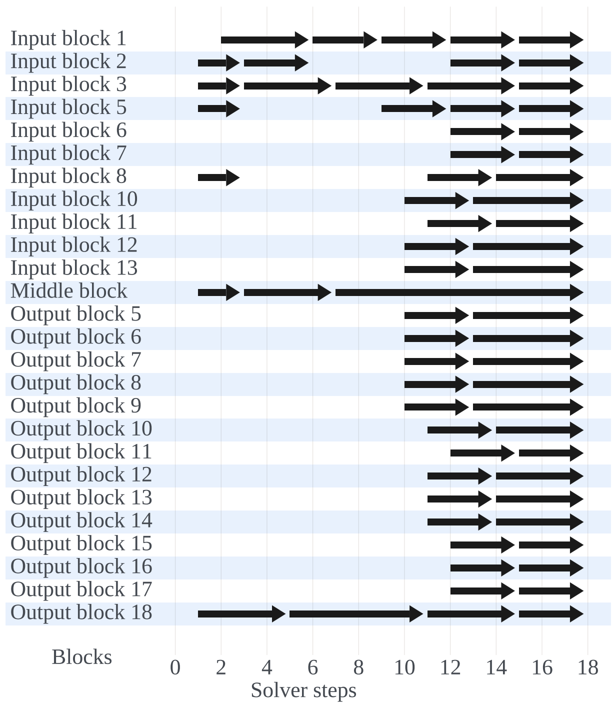

Figure 18. **Cache Schedules for EMU-768 - DDIM 20 Steps.**

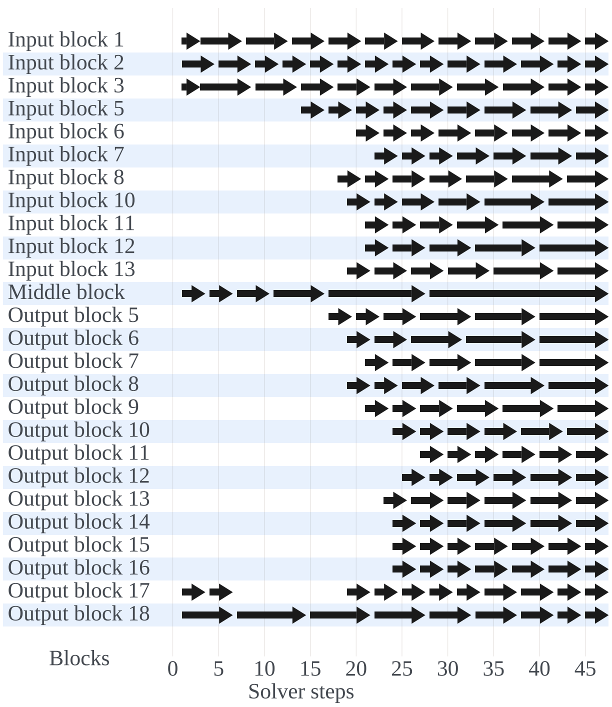

Figure 19. **Cache Schedules for EMU-768 - DPM 50 Steps.**

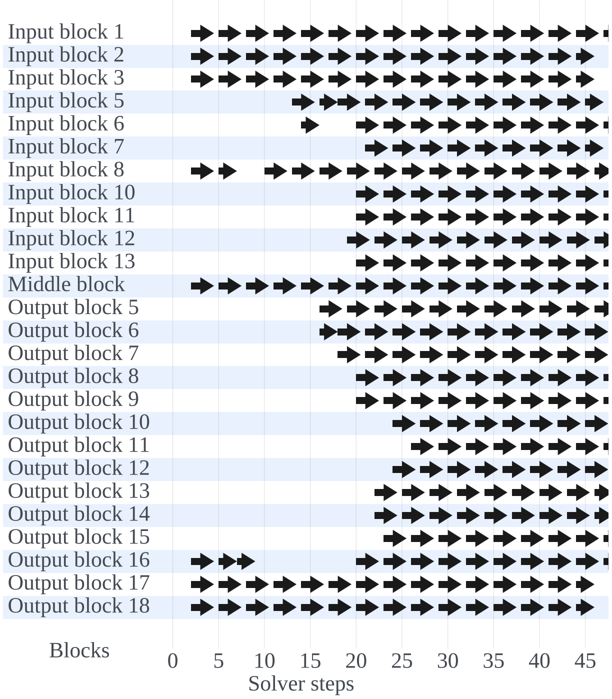

Figure 20. **Cache Schedules for EMU-768 - DDIM 50 Steps.**

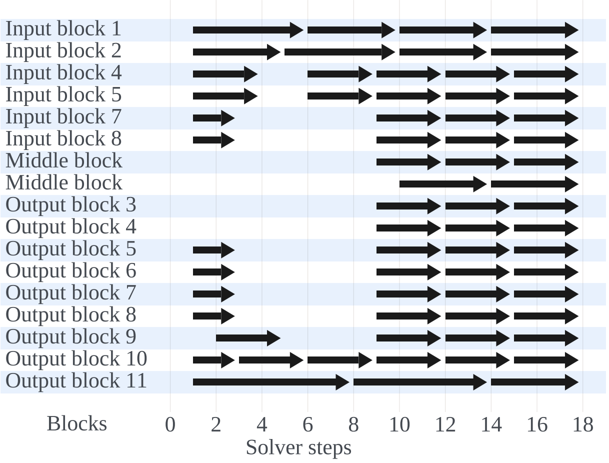

Figure 21. **Cache Schedules for LDM-512 - DPM 20 Steps.**

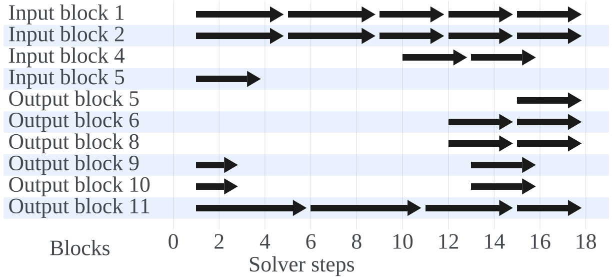

Figure 22. **Cache Schedules for LDM-512 - DDIM 20 Steps.**

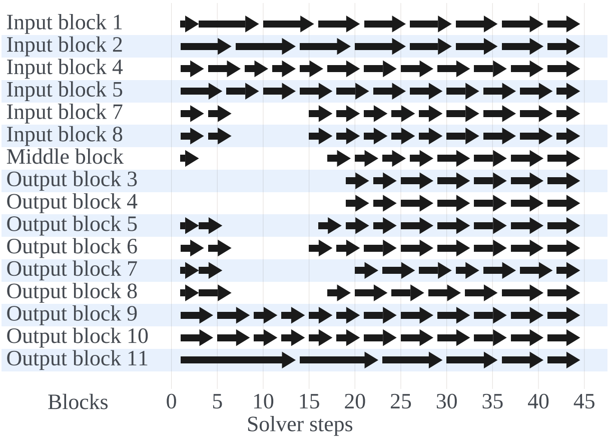

Figure 23. **Cache Schedules for LDM-512 - DPM 50 Steps.**

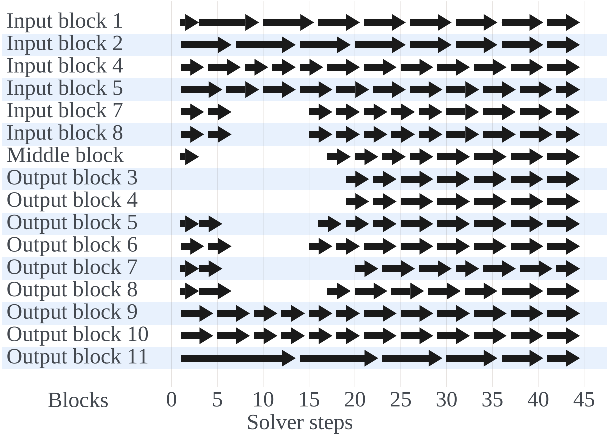

Figure 24. **Cache Schedules for LDM-512 - DDIM 50 Steps.**