Accelerating Diffusion-based Video Editing via Heterogeneous Caching: Beyond Full Computing at Sampled Denoising Timestep
- Supplementary Materials -

Tianyi Liu[1] Ye Lu[1] Linfeng Zhang[2] Chen Cai[1] Jianjun Gao[1]
Yi Wang[3] Kim-Hui Yap[1] Lap-Pui Chau[3]

[1]Nanyang Technological University [2]Shanghai Jiao Tong University [3]The Hong Kong Polytechnic University

{liut0038, lu0001ye, e190210, gaoj0018}@e.ntu.edu.sg zhanglinfeng@sjtu.edu.cn

ekhyap@ntu.edu.sg {yi-eie.wang, lap-pui.chau}@polyu.edu.hk

1. Supplementary Heterogeneity Analysis

In this section, we provide additional visualizations related to the DiT [4] Heterogeneity Analysis in the main paper Sec. 3.2, offering a more direct illustration of the heterogeneity across layers and token types.

For the heterogeneity investigation discussed in Section 3.2 of the main paper, we analyze the behavior of the DiT backbone under the masked video-to-video generation (MV2V) paradigm for versatile video editing [2]. As illustrated in Fig. 1, we take the editing process of a real driving video with a rectangular mask in the Wan2.1 [5] framework as an example. At each denoising timestep t, we visualize five sampled frames from the intermediate outputs, and extract several attention maps from the last DiT layer to show the attention weights between unmasked context tokens in each frame and the masked generative tokens in frame 0. In addition, we annotate the timesteps that would be skipped when applying existing caching strategies based on difference-accumulation, such as TeaCache [3] with a cache threshold of 0.05.

For Argument 1 (Spatiotemporal Heterogeneity), the visualizations clearly indicate that different ranges of denoising timesteps exhibit distinct behaviors: early steps mainly focus on establishing the global structure and layout, whereas later steps are responsible for refining local details. This staged behavior explains why purely difference-accumulation–based caching tends to overlook high-SNR, perceptually critical refinement steps. As shown in the example, many of these late refinement timesteps are still frequently skipped by TeaCache-like strategies, which can negatively affect the final editing quality.

For Argument 2 (RoI characteristics), Fig. 1 shows that, for the generative tokens inside the rectangular masked region, the context tokens with the strongest attention across denoising timesteps are exactly the margin tokens defined in the main paper. Consistent with intuition, tokens corresponding to the car-related context in the surrounding area typically receive the second-highest attention weights.

Taken together, these visual findings reinforce that relying solely on the accumulated difference over denoising timesteps for pure timestep-level caching is inherently limited. To achieve a better efficiency–performance trade-off, it is necessary to further explore partial computation at different types of timestep, as well as selective context-involved attention that explicitly controls which context tokens participate in DiT updates.

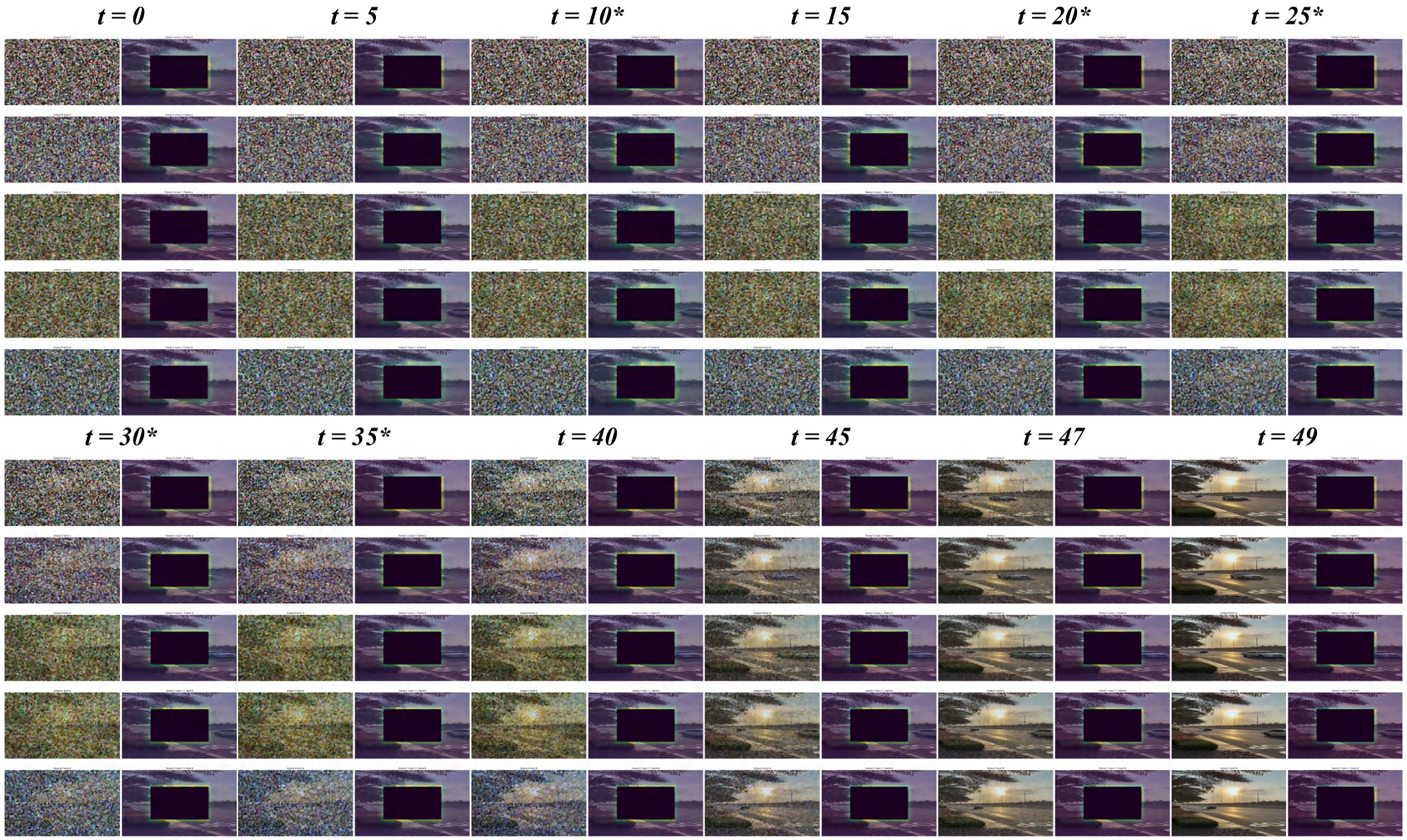
2. Hardware Setups and Model Configurations

All experiments are conducted on a single workstation equipped with a NVIDIA RTX 3090 GPU (24 GB VRAM), an AMD Ryzen Threadripper 3990X 64-core CPU, and 192 GB of system memory.

To match the masked video-to-video (MV2V) editing setting and ensure reliability and reproducibility, we adopt the open-source Wan2.1-VACE-1.3B model as the unified generative backbone. Unless otherwise specified, all experiments are performed at 480p resolution. For the video completion task on the VACE-Benchmark [2], we evaluate the completion quality of the first 33 frames of each video; for the text-guided video editing task on VPBench [1], we evaluate 25-frame video clips. It should be noted that the 1.3B variant is already an officially released lightweight model, so the exploitable redundancy for acceleration is inherently limited, which may partially compress the observable speedup of our method. We therefore encourage future work to reproduce and extend our approach on larger video generation models for a more comprehensive assessment.

3. Additional Evaluation Results

We include additional visualization results for evaluation and ablation. These examples are intended to qualitatively



* If TeaCache is enabled, denoising timesteps that are skipped, or whose nearby steps are skipped multiple times.

Figure 1. The visualized generation process to investigate the heterogeneity in DiT-based video editing.

examine the temporal dynamics and visual fidelity of the generated and edited videos. By comparing visual results under different methods and configurations, the proposed approach can be better assessed in terms of motion consistency, structural stability, and fine-detail reconstruction.

References

[1] Yuxuan Bian, Zhaoyang Zhang, Xuan Ju, Mingdeng Cao, Liangbin Xie, Ying Shan, and Qiang Xu. Videopainter: Any-length video inpainting and editing with plug-and-play context control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–12, 2025. 1

[2] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025. 1

[3] Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7353–7363, 2025. 1

[4] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 1

[5] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1

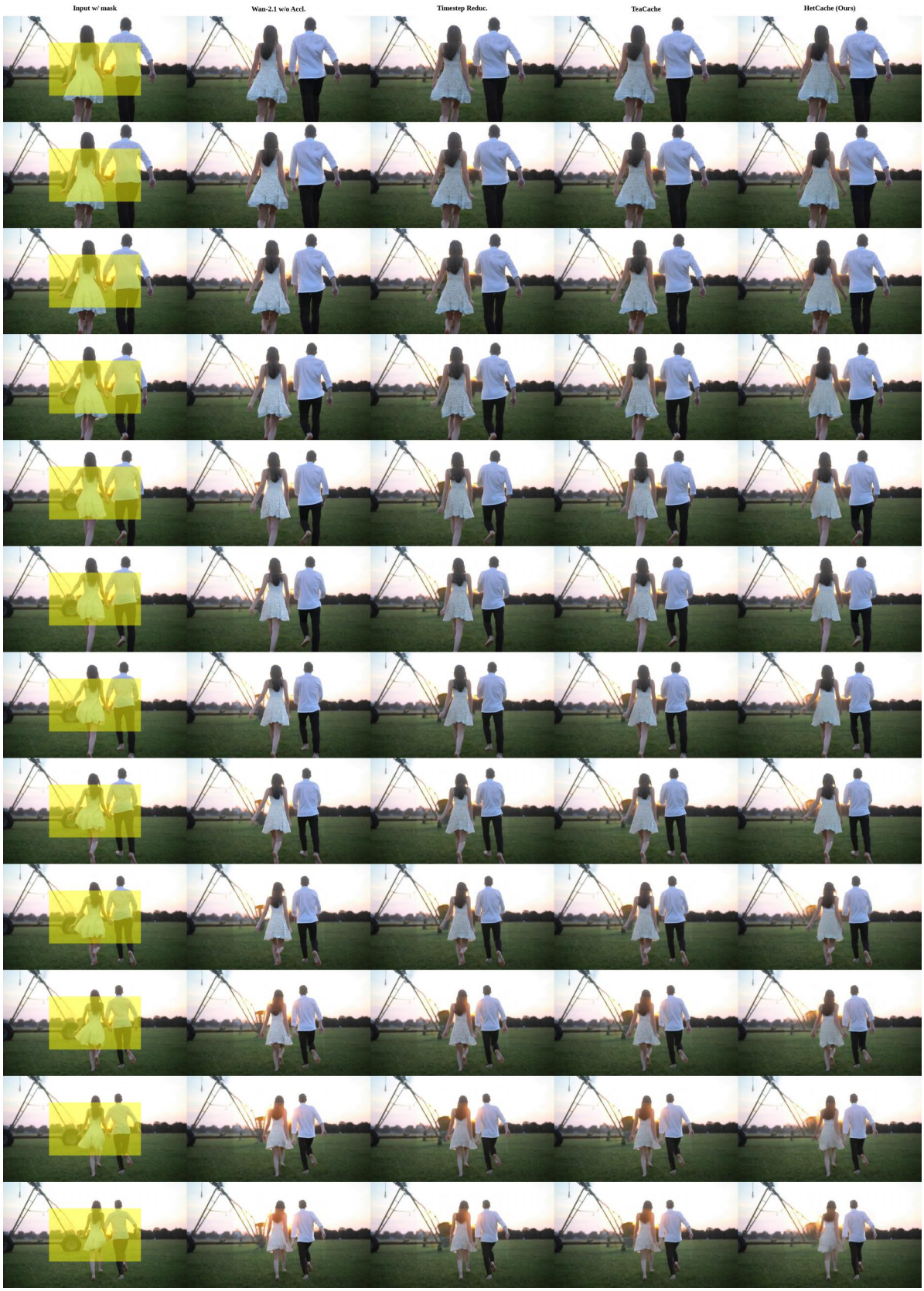


Figure 2. Supplementary qualitative evaluation for VACE inpainting results 1.

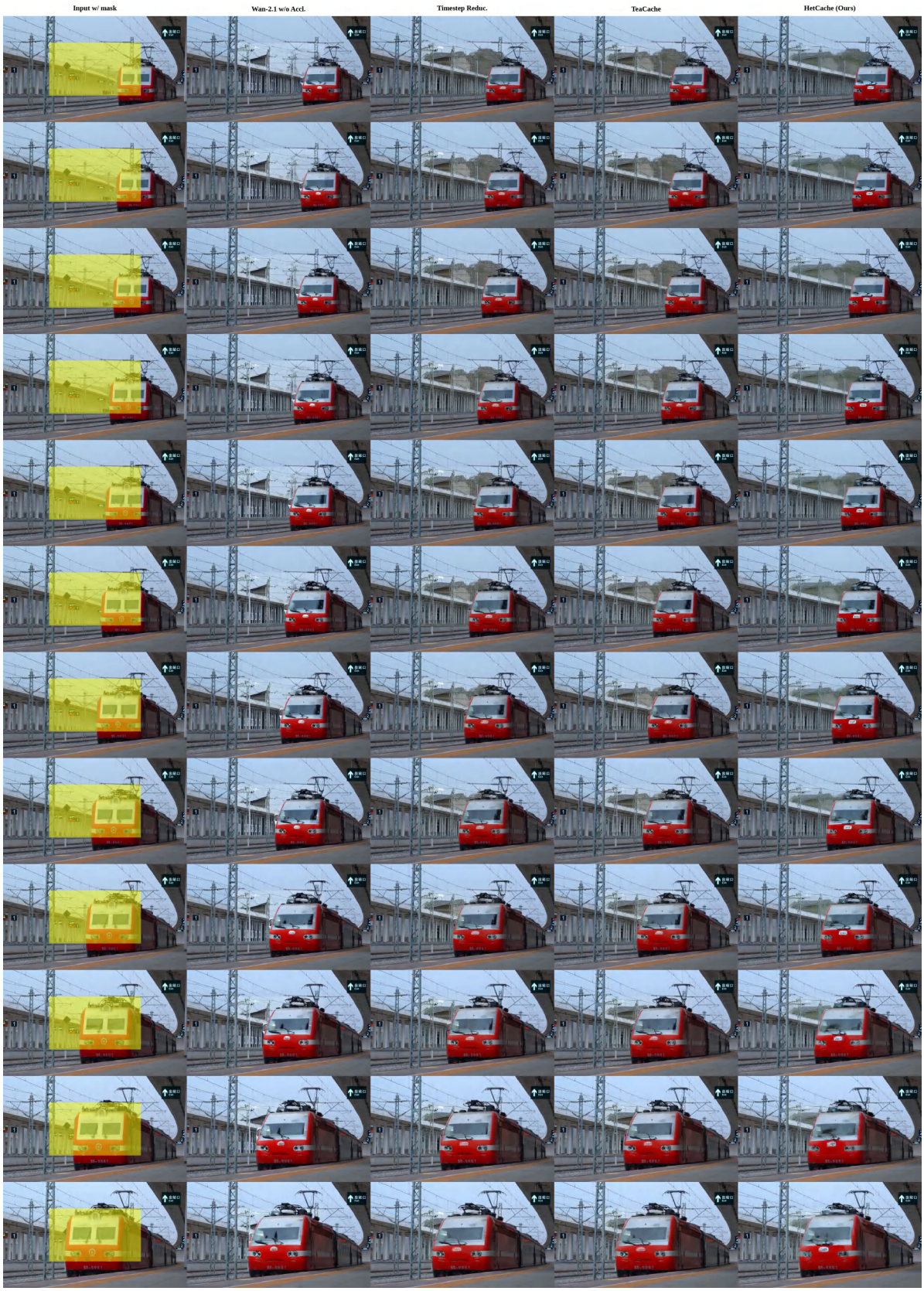


Figure 3. Supplementary qualitative evaluation for VACE inpainting results 2.

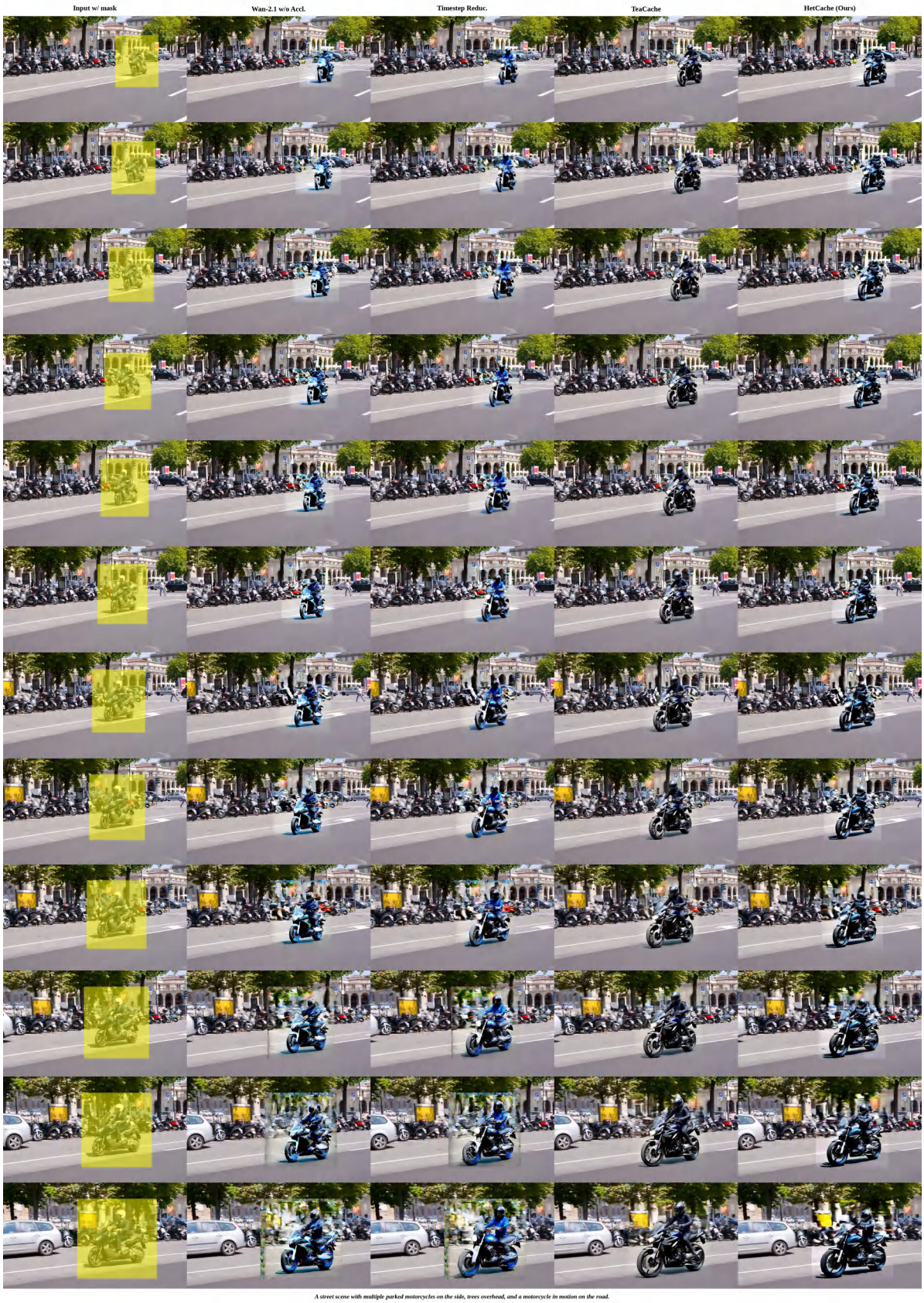


Figure 4. Supplementary qualitative evaluation for VPBench editing results 1.

Figure 5. Supplementary qualitative evaluation for VPBench editing results 2.

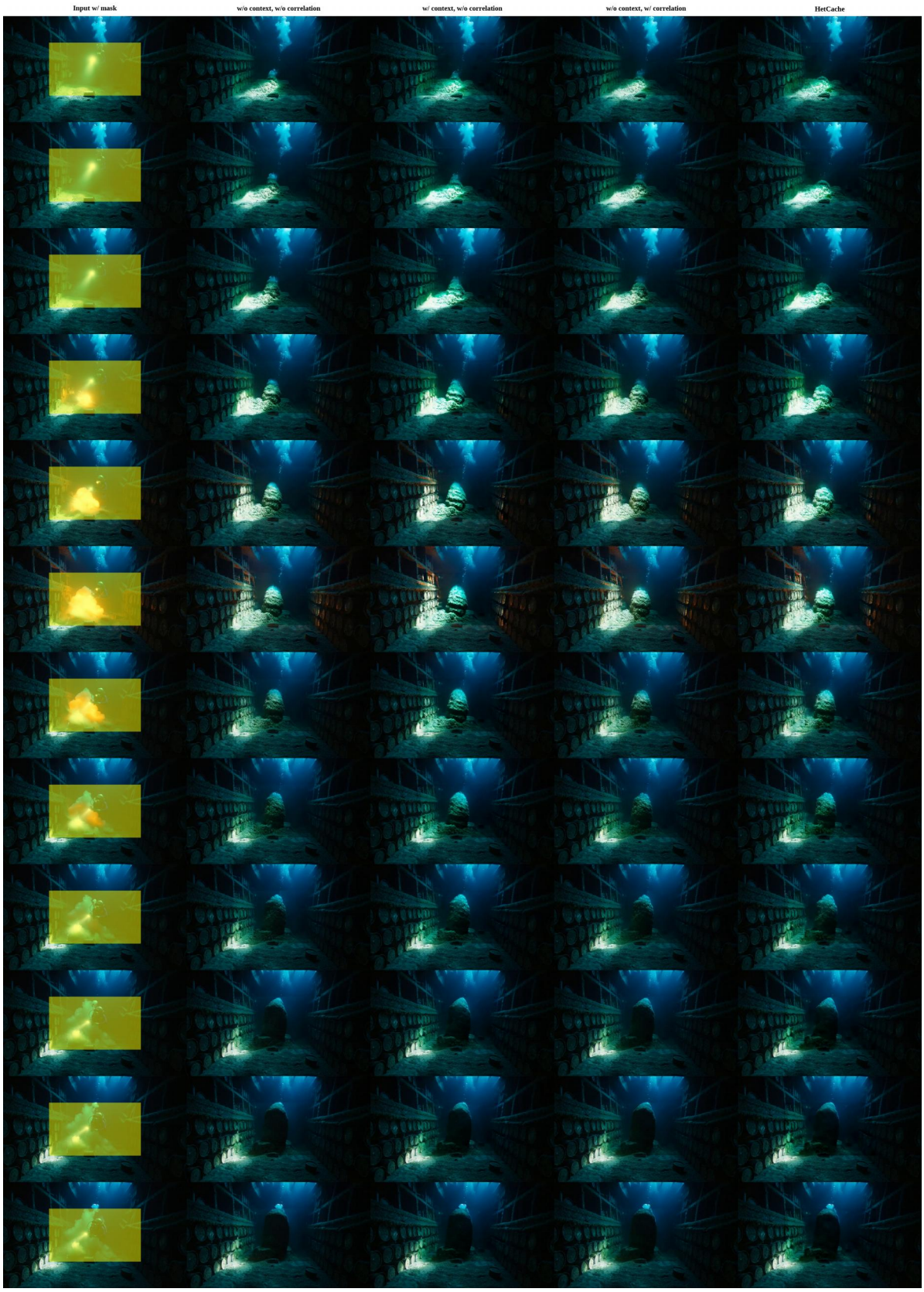


Figure 6. Supplementary visualized ablation study results 1.

Figure 7. Supplementary visualized ablation study resultss 2.