Turbo2K: Towards Ultra-Efficient and High-Quality 2K Video Synthesis Supplementary Material

This supplementary document provides dicussion of limitation and potenrial future work in Sec1, additional implementation details in Sec. 2, further analysis of LR guidance strategies and additional visual results in Sec. 3.

1. Limitation and Future Work

While Turbo2K demonstrates strong efficiency and quality in 2K video generation, its performance remains challenged in highly dynamic or visually complex scenes. In particular, the model occasionally produces unnatural or inconsistent hand poses, reflecting difficulties in synthesizing fine-grained motion details. These limitations are likely attributable to the constrained model capacity and limited diversity in high-quality training data. Future work will investigate scaling the model and expanding the dataset to enhance its generalization and fidelity in such challenging scenarios.

Moreover, the current VAE architecture does not support temporally partitioned decoding, which poses a bottleneck for generating long-duration videos due to substantial memory demands. Addressing this constraint through a more efficient, sequential VAE decoding scheme will be essential for enabling scalable and temporally consistent high-resolution video synthesis.

2. Implementation Details

To train our model, we curated a high-resolution video dataset with full copyright ownership. The dataset comprises approximately 410K high-quality videos covering a diverse range of scenes and categories, with the majority of samples available in 4K resolution. Each video is labeled using ShareGPT4Video [1], providing rich textual descriptions to facilitate text-to-video training. To enhance the diversity of training samples, we adopt a mixed training approach that combines videos and images at a 2:1 ratio. This strategy ensures that the model effectively learns both temporal dynamics from videos and high-quality spatial details from images, contributing to improved generative performance. All experiments are conducted using the Adam optimizer with a learning rate of 10^{-4}.

Progressive training strategy. To optimize training efficiency and stabilize convergence, we employ a progressive training strategy where the model is trained at incrementally increasing resolutions. During the heterogeneous model distillation stage, the student model is first trained at a resolution of 544×960, allowing it to effectively inherit knowledge from the teacher model while maintaining computational efficiency. For the two-stage synthesis, the HR model is initially trained at a resolution of $49 \times 1440 \times 2560$ for 5K iterations, enabling it to establish a coarse high-resolution structure. Subsequently, the model is further fine-tuned at a resolution of $121 \times 1440 \times 2560$ for an additional 8K iterations, allowing for enhanced detail refinement and temporal consistency. Both LR and HR models consist of 28 DiT blocks, with LR guidance extracted at block indices $0, 7, 14, 21$. Each extracted feature is fused with its corresponding HR feature through a fusion block.

3. Addition Results

This section first provides further explanations on the timestep configurations for extracting LR guidance in Sec. 3.0.1, followed by additional generated results of Turbo2K in Sec. 3.1. We also recommend watching the videos provided in the supplementary file for a more comprehensive evaluation.

Additionally, we present a visual comparison in video format for heterogeneous model distillation, including results from the LTX baseline, pure data fine-tuning, distillation with a fixed teacher timestep at the final step, and distillation where the teacher timestep is aligned with the student model. The comparison demonstrates that pure data fine-tuning yields limited improvements, while fixing the teacher's timestep at the final step provides insufficient supervision, as the teacher's features at this stage closely resemble clean data. In contrast, aligning teacher and student timesteps during distillation better preserves the denoising trajectory, leading to superior generative quality and improved semantic coherence.

3.0.1. Analysis of LR Guidance Strategies

To further analyze the effect of LR guidance on HR generation, we provide visual comparisons of intermediate frames across different guidance configurations in Fig. 1.

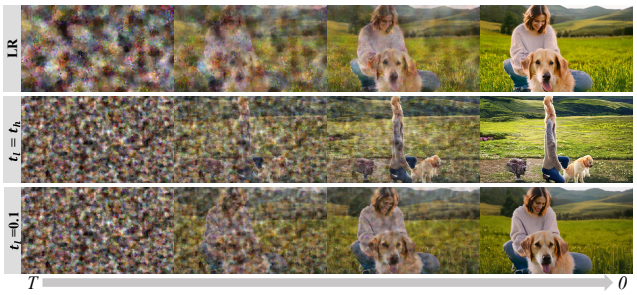

Figure 1. Comparison of LR results and HR results guided by LR-based feature guidance across timesteps.

We compare standard LR generation, synchronized LR-HR timesteps where LR features are extracted at the same timestep as HR denoising, and final-step LR guidance where LR features are taken from the last denoising step. The results indicate that synchronizing LR and HR timesteps leads to unstable HR structures, as early LR features are not yet well-formed, causing HR synthesis to inherit ambiguous details. By the time LR features stabilize, HR is already in its final refinement stage, limiting its ability to incorporate structural corrections. In contrast, using LR features from the final denoising step provides the most stable and informative guidance, ensuring coherent structural formation in HR synthesis.

3.1. More Visual Results of Turbo2K

We present additional frames generated by Turbo2K in Fig. 2, Fig. 3, and Fig. 4, demonstrating rich details, high aesthetic quality, and strong semantic coherence. Additionally, we provide video results in this supplementary file and recommend viewing them for a more comprehensive evaluation.

References

[1] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024. 1

Figure 2. Our Turbo2K generated results

Figure 3. Our Turbo2K generated results

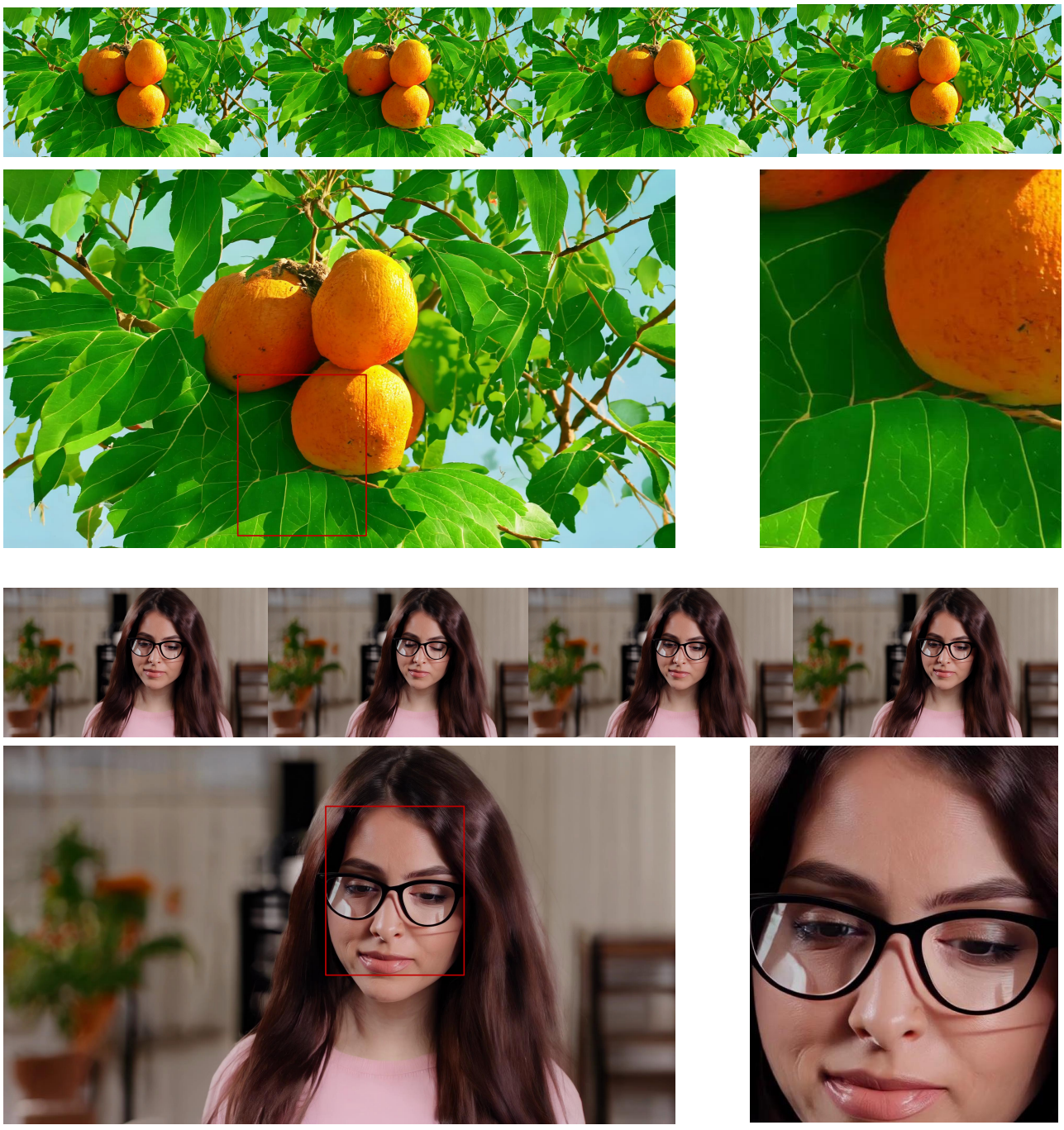

Figure 4. Our Turbo2K generated results