

Figure A. **User study interface.** Participants were shown an input prompt and two generated videos from different methods. They were asked to compare the results based on *Consistency*, *Dynamics* and *Aesthetics*. Each question allowed skipping if the difference was hard to judge.

## Appendix

### A. Details of User Study

**User Study: Comparison with Existing Methods.** To evaluate the effectiveness of our method, we conducted a user study comparing it against several existing approaches. We collected a total of 74 video pairs, each generated from the same input image or text prompt to ensure fair comparisons. Competing methods included Free4D [4], 4Real [7], GenXD [9], and Animate124 [8]. All comparison videos were obtained from their official project pages. The study was conducted online, and a screenshot of the evaluation interface is shown in Fig. A. Participants were asked to assess each video pair across three criteria: Consistency, Dynamics, and Aesthetics. For each criterion, they were instructed to choose the video they perceived as better. If a comparison was too difficult to judge, they could skip to the next example without selecting an answer. All responses were collected anonymously, and no personal data were recorded during the study.

### B. Details of VBench Metrics

To comprehensively evaluate the quality of our synthesized novel-view videos, we adopt a suite of metrics introduced in VBench [2], covering three key aspects: *Consistency* (for both subject and background), *Degree of Motion*, and *Aesthetic Quality*.

**Subject / Background Consistency.** This metric assesses how consistently both the main subject (e.g., human, vehicle, animal) and the surrounding background are maintained

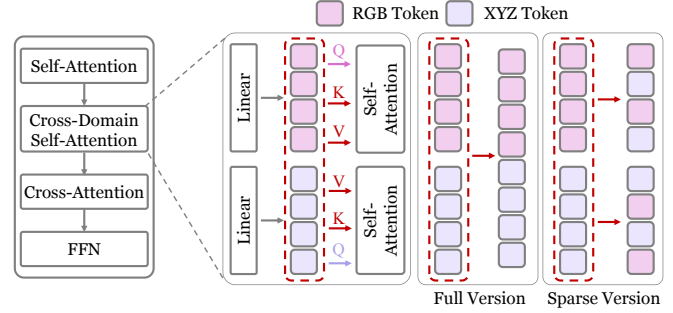


Figure B. **Architecture of the Cross-Domain Self-Attention (CDSA) module.** The CDSA block is inserted between self-attention and cross-attention layers to facilitate bidirectional interaction between RGB and XYZ modalities. We explore two variants: the *Full Version*, where all tokens interact densely, and the *Sparse Version*, where attention is restricted to spatially corresponding token pairs. This design enables effective cross-modal alignment with different trade-offs in efficiency and performance.

throughout the video. It leverages feature similarity across frames using DINO [1] for the foreground and CLIP [5] for the background. DINO focuses on preserving subject identity by comparing learned visual representations, while CLIP captures broader scene coherence. The average of both consistency metrics provides a balanced view of overall temporal consistency.

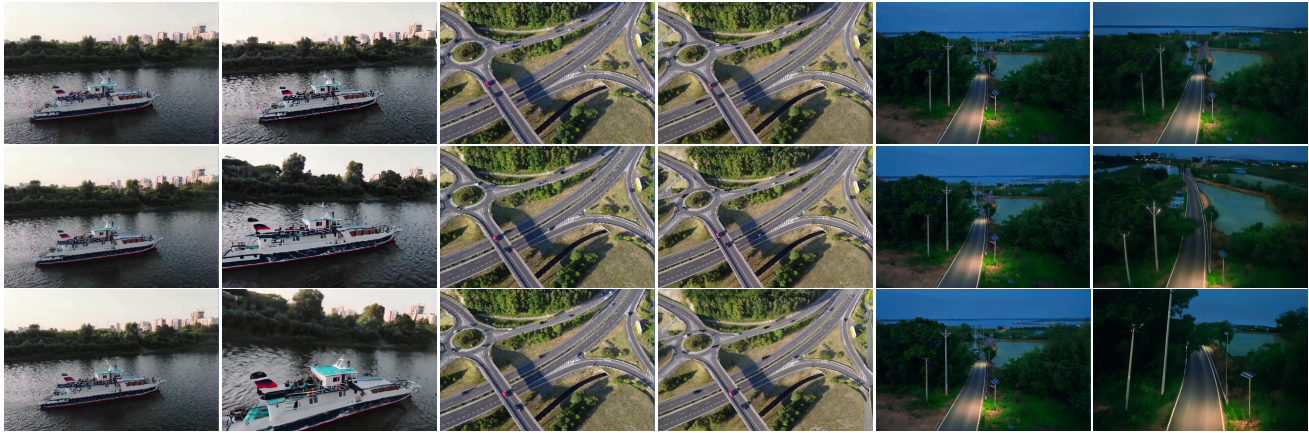
**Degree of Motion.** To avoid favoring overly static videos that may perform well on consistency metrics, we include a motion-aware measure. Specifically, RAFT [6] is applied to estimate optical flow, and the *Dynamic Degree* is computed by averaging the top 5% of largest flow magnitudes. This helps emphasize prominent movements, such as object actions or camera shifts, while de-emphasizing negligible or noisy motions, ensuring a more meaningful evaluation of dynamics.

**Aesthetic Quality.** To reflect the perceived visual appeal of the generated videos, we utilize the LAION Aesthetic Predictor [3], a lightweight regressor trained atop CLIP features to score image aesthetics on a scale from 1 to 10. It considers multiple factors, including color composition, realism, layout, and overall artistic impression. We apply this predictor to each frame and report the average score as the final *Aesthetic Quality* metric.

### C. Cross-Domain Self-Attention (CDSA)

As introduced in Sec. ??, we introduce a Cross-Domain Self-Attention (CDSA) module to enhance the alignment between RGB and XYZ modalities, particularly under the batch-wise fusion strategy. Figure B illustrates the architecture of this module.

As shown in the left part of Fig. B, the CDSA block is inserted between the standard self-attention and cross-



Generated Video Novel-view Video Generated Video Novel-view Video Generated Video Novel-view Video

Figure C. Novel-view video results on in-the-wild data.

attention layers within a transformer block. It explicitly enables bidirectional interaction between RGB and XYZ tokens through attention mechanisms—allowing RGB tokens to attend to XYZ tokens and vice versa—thus facilitating cross-modal information exchange.

To balance performance and efficiency, we implement and compare two versions of CDSA:

- Full Version: All RGB and XYZ tokens participate in dense cross-domain attention. This version achieves stronger modality interaction at the cost of higher memory and computation.
- Sparse Version: Token interactions are restricted to spatially corresponding positions between RGB and XYZ sequences. This reduces overhead while retaining most of the alignment benefits.

While both versions aim to bridge the modality gap by promoting fine-grained token-level communication, our experiments reveal that under the batch-wise fusion setting, even with CDSA, the overall cross-modal alignment remains limited. This is primarily due to the spatial separation of RGB and XYZ tokens, which contrasts with the more effective width-wise fusion strategy where the interaction distance is inherently shorter.

## D. More Results

We present novel-view video generation results on in-the-wild data. As shown in Fig. C, our method generates high-quality multi-view videos, demonstrating the model’s effectiveness in complex real-world environments.

## References

- [1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In

2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, pages 9630–9640. IEEE, 2021. 1

- [2] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 21807–21818. IEEE, 2024. 1
- [3] LAION-AI. aesthetic-predictor, 2022. 1
- [4] Tianqi Liu, Zihao Huang, Zhaoxi Chen, Guangcong Wang, Shoukang Hu, Liao Shen, Huiqiang Sun, Zhiguo Cao, Wei Li, and Ziwei Liu. Free4d: Tuning-free 4d scene generation with spatial-temporal consistency. *arXiv preprint arXiv:2503.20785*, 2025. 1
- [5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 1
- [6] Zachary Teed and Jia Deng. RAFT: recurrent all-pairs field transforms for optical flow (extended abstract). In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4839–4843. ijcai.org, 2021. 1
- [7] Heng Yu, Chaoyang Wang, Peiye Zhuang, Willi Menapace, Aliaksandr Siarohin, Junli Cao, László A. Jeni, Sergey Tulyakov, and Hsin-Ying Lee. 4real: Towards photorealistic 4d scene generation via video diffusion models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024*,

- 130 *NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15,*  
 131 *2024, 2024. [1](#)*
- 132 [8] Yuyang Zhao, Zhiwen Yan, Enze Xie, Lanqing Hong, Zhen-  
 133 guo Li, and Gim Hee Lee. Animate124: Animating one image  
 134 to 4d dynamic scene. *CoRR*, abs/2311.14603, 2023. [1](#)
- 135 [9] Yuyang Zhao, Chung-Ching Lin, Kevin Lin, Zhiwen Yan,  
 136 Linjie Li, Zhengyuan Yang, Jianfeng Wang, Gim Hee Lee,  
 137 and Lijuan Wang. Genxd: Generating any 3d and 4d scenes.  
 138 *CoRR*, abs/2411.02319, 2024. [1](#)