Free4D: Tuning-free 4D Scene Generation with Spatial-Temporal Consistency

Supplementary Material

A. More Implementation Details

4D-GS Network. 4D Gaussian Splatting (4D-GS) [12] lies in extending static 3D Gaussian primitives [4] to dynamically model temporal-spatial scenes. In 3D-GS [4], a scene is represented by a set of anisotropic Gaussians $\mathcal{G} = \{g_i\}_{i=1}^N$, where each Gaussian g_i is parameterized by its position $\mu_i \in \mathbb{R}^3$, rotation (quaternion $q_i \in \mathbb{R}^4$), scale $s_i \in \mathbb{R}^3$, and opacity $\alpha_i \in [0, 1]$. The covariance matrix Σ_i is derived from q_i and s_i, enabling differentiable rendering via splatting.

To model 4D dynamics, each Gaussian is further augmented with time-varying parameters. For temporal coherence, we parameterize the trajectory of g_i over time t through a deformation function $\Delta : \mathbb{R}^4 \to \mathbb{R}^9$:

$$[\Delta\mu_i(t), \Delta q_i(t), \Delta s_i(t)] = \Delta(\mu_i, q_i, s_i, t), \tag{1}$$

where Δ can be implemented via MLPs or explicit keyframe interpolation. The interpolated Gaussian $g_i(t)$ at time t is then rendered following the 3D-GS rendering pipeline, but with all parameters conditioned on t. Optimization typically requires multi-view RGB videos with camera poses. While achieving real-time dynamic rendering (30+ FPS), 4D-GS depends heavily on consistent multi-view video supervision.

Training Setup. We adopt the 4D representation proposed in [12]. Our hyperparameter settings mainly follow those in [12]. The learning rate is initialized at 1.6×10^{-3} and gradually decays to 1.6×10^{-4} by the end of training. The Gaussian deformation decoder, implemented as a tiny MLP, starts with a learning rate of 1.6×10^{-4}, which is reduced to 1.6×10^{-5} over time. The training batch size is set to 1. During the coarse stage, we train for 9k iterations, followed by an additional 1k iterations in the fine stage. The λ used in the fine-stage loss is 0.1. In modulation-based refinement, $\bar{T}$ is set to 5 to improve efficiency, and w_i linearly decreases from 0.5 to 0. Viewcrafter [15] uses its default denoising steps, which is 50. The guidance scale s used in CFG is the default value 7.5. For multi-view image generation at the first timestamp $t = 1$, we use adaptive CFG. For $t > 1$, CFG is disabled because the reference information from the multi-view generation at $t = 1$ has already been introduced into the missing regions. All experiments are conducted on a single NVIDIA A100 (40GB) GPU.

B. Details of User Study

User Study I: Comparison with Other Methods. We conducted the first user study to compare our method with other existing methods. Since the source codes of these methods were not publicly available, we compared our method with the videos provided on their respective project pages. A total of 32 pairs of videos were used in this study. Each pair was generated from the same input images or text prompts to ensure a fair comparison. The methods included in this study were 4Real [14], 4Dfy [1], Dream-in-4D [19], DimensionX [8], GenXD [18], and Animate124 [17]. The user study was conducted online, and a screenshot of the interface is shown in Fig. A. Participants were asked to evaluate the generated videos based on four criteria: Consistency, Dynamic, Aesthetic, and Overall. For each pair of videos, they were required to select which method performed better for each criterion. They could skip to the next example without selecting if they found it difficult to judge. The user study was conducted anonymously, and no personally identifiable data were collected.

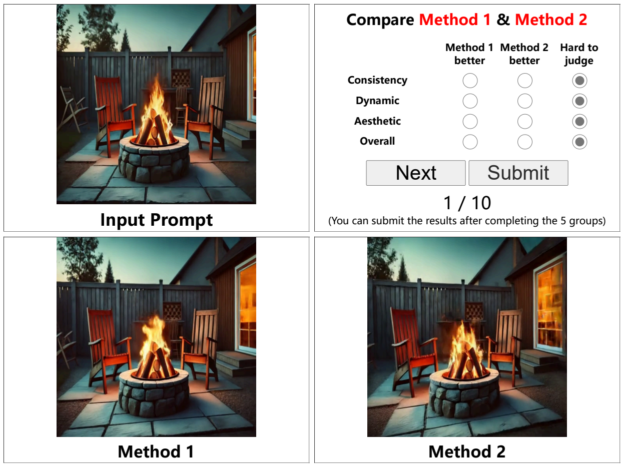
User Study II: Ablation Study. The second user study evaluated the impact of our method's different components through an ablation experiment. The components included in this study were *Monst3R*, *Adaptive CFG*, *Point Cloud Guided Denoising*, *Reference Latent Replacement*, *Reference Latent Replacement*, *Coarse-to-fine optimization*, and *Modulation-based Refinement*. For each component, we randomly sampled 10 different scenes and generated video pairs using the full version of our method and a variant with the specific component removed or modified. Participants were asked to evaluate the generated video pairs based on the same four criteria as in User Study I: consistency, aesthetics, motion dynamic, and overall quality. The ablation study was also conducted anonymously, without collecting any personally identifiable data.

C. Details of VBench Metrics

To evaluate the quality of multi-view videos rendered from 4D representations, we report common VBench [3] metrics: Consistency (average for subject/background), Dynamic Degree, Aesthetic Score, and Text Alignment (only for text-to-4D).

Subject / Background Consistency. To evaluate the consistency of both subjects (*e.g.*, a person, car, or cat) and background scenes in the video, VBench uses DINO [2] and CLIP [6] feature similarities across frames. DINO captures subject consistency by comparing frame embeddings, while CLIP assesses background stability. Together, they provide a comprehensive measure of consistency.

Dynamic Degree. Since a static video can score well in the aforementioned consistency metrics, it is important to eval-



Compare Method 1 & Method 2

	Method 1 better	Method 2 better	Hard to judge
Consistency	○	○	●
Dynamic	○	○	●
Aesthetic	○	○	●
Overall	○	○	●

1 / 10

(You can submit the results after completing the 5 groups)

Figure A. **The web interface of our user studies.** The input prompt can be either a single image or a short text.

uate the degree of dynamics (*i.e.*, whether it contains large motions). To this end, the Dynamic Degree metric uses RAFT [9] to estimate the degree of dynamics in synthesized videos. Specifically, this metric takes the average of the largest 5% optical flows (considering the movement of small objects in the video). This approach ensures that minor movements (*e.g.*, small objects or slight camera shakes) do not disproportionately influence the overall dynamic assessment.

Aesthetic Score. We evaluate the artistic and beauty value perceived by humans towards each video frame using the LAION Aesthetic Predictor [5]. This predictor is a linear model built on top of CLIP embeddings, trained to assess the aesthetic quality of images on a scale from 1 to 10. It reflects various aesthetic aspects, including the layout, richness and harmony of colors, photo-realism, naturalness, and overall artistic quality of the video frames. The Aesthetic Score metric obtains a normalized aesthetic score by applying this predictor to each frame.

Text Alignment. This metric uses overall video-text consistency computed by ViCLIP [11] on general text prompts as an aiding metric to reflect text semantics consistency. ViCLIP is a video-text contrastive learning model that leverages a large-scale video-text dataset to learn robust and transferable representations.

D. Runtime Analysis

The runtime comparison is shown in Table A. We compare our approach with object-level methods [1] and the text-to-4D scene generation method [14]. Since [8] and [18] have not reported runtime details (including feed-forward inference time and 4D representation optimization time) or released their code, they are excluded from the comparison. Notably, compared to previous methods, our approach supports higher resolutions while efficiently handling more frames and viewpoints, achieving the fastest optimization. Our total runtime is composed of three main steps: running MonST3R (1 min), generating multi-view videos with ViewCrafter (25 min), and optimizing 4D-GS (35 min).

Method	Time	Resolution	Frames	Views
4Dfy [1]	10h+	256×256	-	-
4Real [14]	1.5h	256×144	8	16
Ours	1h	1024×576	16	25

Table A. **Comparison of runtime with other methods.** Frames and Views represent the number of video frames and the number of viewpoints, respectively. The running time of Structure from Motion (SfM), such as colmap [7], is not included due to significant variations across different scenes.

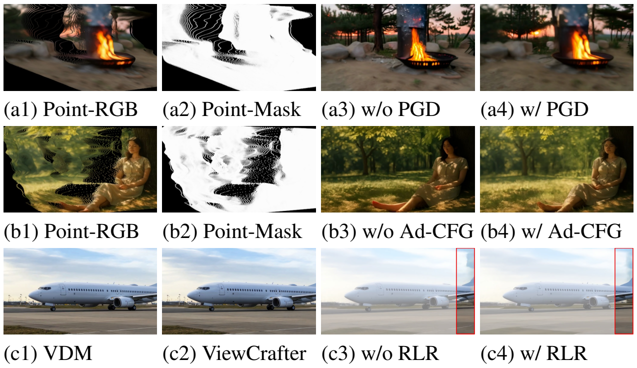

(a1) Point-RGB (a2) Point-Mask (a3) w/o PGD (a4) w/ PGD
(b1) Point-RGB (b2) Point-Mask (b3) w/o Ad-CFG (b4) w/ Ad-CFG
(c1) VDM (c2) ViewCrafter (c3) w/o RLR (c4) w/ RLR

Figure B. **Failure cases** of our proposed PGD / Ad-CFG / RLR.

E. Discussion

Clarification on *Tuning-free*. Our "tuning-free" means no fine-tuning is applied to foundation models (*e.g.*, ViewCrafter [15]) during multi-view video generation. This efficient design avoids the expensive cost of tuning (*e.g.*, GenXD [18] uses 32 A100s for several days). Instead, the final 4DGS optimization is lightweight, scene-specific, and runs on a single GPU within 30 minutes.

Necessity of 4DGS. **1) Task:** Our goal is to generate explicit 4D representations, following 4Real [14]. **2) Consistency:** The 4DGS stage is essential for spatial-temporal consistency (see our project page). **3) Application:** Beyond discrete multi-view videos from diffusion models, 4DGS offers a continuous spatial-temporal representation, enabling real-time rendering and interactivity.

4D Evaluation. For a thorough 4D assessment, we disentangle and separately evaluate temporal and spatial dimensions in Table B. **1) C-T/V↑:** VBench consistency applied to time-/view-variant videos, following SV4D [13]. **2) #PC↑:** number of 3D points reconstructed by COLMAP from view-variant frames, indicating spatial consistency. **3)**

Method	C-T	C-V	#PC	FWE	Method	C-T	C-V	#PC	FWE
4Real	98%	96%	7k	3.3	D-in-4D	98%	91%	5k	4.6
Ours	98%	97%	53k	2.4	Ours	98%	97%	47k	2.3
4Dfy	99%	91%	4k	4.0	Ani124	95%	93%	0.1k	4.9
Ours	99%	97%	45k	2.2	Ours	98%	98%	109k	2.9
DimX	97%	92%	3k	4.1	GenXD	97%	90%	1k	5.3
Ours	97%	96%	4k	2.6	Ours	98%	97%	47k	4.2

Table B. Disentangled evaluation on temporal / spatial dimension.

Flow Warping Error (FWE)↓: measures inter-frame difference after flow warping, reflecting temporal consistency. All methods are evaluated under the same settings (*e.g.*, resolution and video length). Our method shows better spatial-temporal consistency despite greater dynamics.

F. Limitations and Future Work

Limitations. We show failure cases of our proposed PGD / Adaptive CFG / RLR in Fig. B. **1)** Point-Guided Denoising (PGD): While PGD (a4) reduces unintended dynamics in frozen-time videos compared to (a3), the fire (a4) appears unnatural due to **inaccurate depths** (a1) under blurry input. **2)** Adaptive CFG: It alleviates abrupt facial and lighting changes (b3) yet minor **facial** inconsistencies remain (b4). **3)** Reference Latent Replacement (RLR): When dynamic regions are not fully captured in input view, VDM (c1) and ViewCrafter (c2) may generate **inconsistent content** (*e.g.*, airplane tail). In such cases, RLR causes discontinuities (c4), while disabling it results in temporal flickering (c3). Though imperfect, the proposed heuristics perform better.

Future Work. We recognize that the accuracy of MonST3R [16]'s estimation of dynamic videos is crucial. We observed that Dust3R [10] demonstrates better robustness than MonST3R in some static scenes. Therefore, a potential approach is to use Dust3R to estimate the geometry of the first frame, and employ optical flow to link different views during the subsequent 4DGS optimization.

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas J. Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B. Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 7996–8006. IEEE, 2024. 1, 2

[2] 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

[3] 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

[4] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, 2023. 1

[5] LAION-AI. aesthetic-predictor, 2022. 2

[6] 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

[7] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2

[8] Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *CoRR*, abs/2411.04928, 2024. 1, 2

[9] 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. 2

[10] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 3

[11] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 2

[12] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20310–20320, 2024. 1

[13] Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. SV4D: dynamic 3d content generation with multi-frame and multi-view consistency. In *The*

Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net, 2025. 2

[14] 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, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 1, 2

[15] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 1, 2

[16] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024. 3

[17] Yuyang Zhao, Zhiwen Yan, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Animate124: Animating one image to 4d dynamic scene. *CoRR*, abs/2311.14603, 2023. 1

[18] Yuyang Zhao, Chung-Ching Lin, Kevin Lin, Zhiwen Yan, Linjie Li, Zhengyuan Yang, Jianfeng Wang, Gim Hee Lee, and Lijuan Wang. Genxd: Generating any 3d and 4d scenes. *CoRR*, abs/2411.02319, 2024. 1, 2

[19] Yufeng Zheng, Xueting Li, Koki Nagano, Sifei Liu, Otmar Hilliges, and Shalini De Mello. A unified approach for text- and image-guided 4d scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7300–7309, 2024. 1