DIMO: Diverse 3D Motion Generation for Arbitrary Objects

Supplementary Material

This document contains more implementation details and more results not elaborated in our paper as follows:

- Demo Video (Appendix A)
- Experiments on Challenging Objects (Appendix B)
- Motion-Rich Video Generation Details (Appendix C)
- Implementation Details (Appendix D)
- Experiment Details (Appendix E)
- Limitations and Future Directions (Appendix F)

We open-sourced our codes in https://github.com/Friedrich-M/DIMO to benefit future research in 4D generation.

A. Demo Video

To better visualize our diverse 3D motion generation results and to compare them with baseline methods beyond 2D images, we provide a demo video in the supplementary folder. This video also includes a brief visualization to enhance understanding of our overall pipeline, illustrating *where* the diverse motions come from and *how* to jointly model these motion patterns in a unified motion latent space.

B. Experiments on Challenging Objects

Multi-Object and Large Motions. To show the effectiveness of DIMO to deal with real-world scenarios with multi-object or large motions, we conduct experiments on the DAVIS [12] dataset. DIMO can produce noticeable and accurate motions in real-world videos with multi-objects and large motions as shown in Fig. 1.

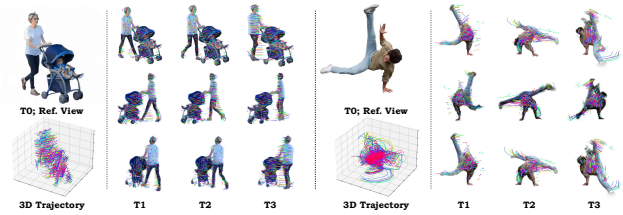

Figure 1. **Qualitative results on real-world DAVIS dataset.**

To further evaluate our robustness in real-world scenes, we quantitatively compare our method with the recent 4D scene generation method DreamScene4D [4] on the DAVIS [12] dataset in Tab. 1. Following DreamScene4D, we adopt CLIP [13] and LPIPS [25] to evaluate the generation quality and conduct a user study to evaluate the overall visual quality. The results show that DIMO can achieve comparable or even better 4D generation quality against the task-specific scene-level method DreamScene4D.

Deformable and Fluid Objects. As mentioned in our main paper, our DIMO can handle a wide range of objects, including deformable and fluid objects. Here we demonstrate our effectiveness in generating deformable soft flag and splatting liquid water from the Objaverse [5] dataset in Fig. 2, which further supports our claim.

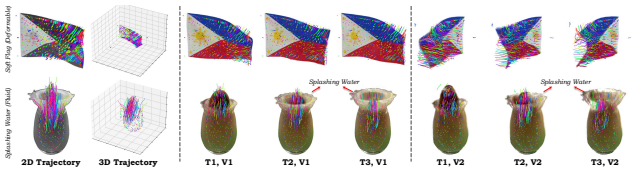

Figure 2. **Qualitative results on deformable and fluid objects.**

C. Motion-Rich Video Generation

We distill motion priors by generating motion-rich video clips for various object categories, including synthetic and real-world objects. Specifically, given a single-view image, we first segment the foreground object with SAM2 [14] and then crop and rescale the image to 512×512. We generate diverse motion prompts using fine-tuned Llama3 [17] and GPT-4o [1] and employ the open-sourced, text-conditioned image-to-video model CogVideoX5B-I2V [22] to perform single-view video generation. To further capture the object geometry, we employ multi-view video models SV3D [19] and SV4D [21] to generate multiple views of the object. We provide the system prompts used for GPT-4o and more details about both the single-view and multi-view video generation processes below.

C.1. Motion Prompts Preparation

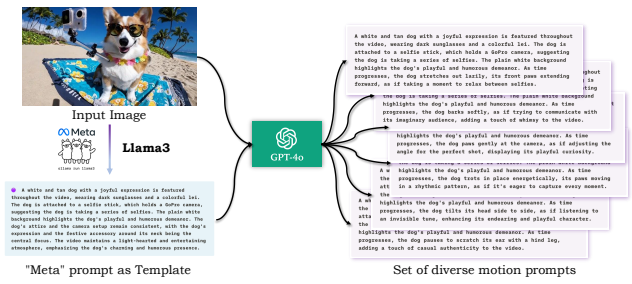

Figure 3. **Overview of Motion Prompts Generation.**

As shown in Fig. 3, given a reference image and a "meta" prompt generated by Llama3 [17] (blue box) as input, we employ the multi-modal large language model GPT-4o [1] to generate diverse motion prompts (pink box) for subsequent video generation. The GPT-4o system prompts are provided below.

Table 1. **Comparison with DreamScene4D [4] on DAVIS [12] subset.** The quantitative results demonstrate the effectiveness of our methods in real-world scenarios with large and multi-object motions.

Method	CLIP ↑			LPIPS ↓			User Preference
	stroller	rollerblade	breakdance	stroller	rollerblade	breakdance	
Ours	**86.68**	85.87	**86.62**	**0.2345**	**0.1359**	**0.1522**	**60.47%**
DreamScene4D [4]	85.29	**89.22**	78.58	0.2755	0.1379	0.1698	39.53%

Motion Prompts Generation

```
"""
**Objective**: **Give fifty different and highly descriptive video captions with diverse motions based on the input image and user input. **. As an expert, delve deep into the image with a discerning eye, leveraging rich creativity, and meticulous thought. When describing the details of an image, include appropriate dynamic information to ensure that the video caption contains reasonable actions and plots. The caption should be modified according to the user's input.

**Note**: The input image is the first frame of the video, and it's a single-centered object with a white background, and the output video caption should describe motion starting from the current image. User input is a template of one possible video caption with a specific motion, and the model should generate different video captions with diverse motions according to the user input.

**Note**: Don't contain camera transitions!!! Don't contain screen switching!!! Don't contain perspective shifts !!!

**Answering Style**:
Answers should be comprehensive, conversational, and use complete sentences. Provide context where necessary and maintain a certain tone. The motion should be diverse, reasonable, simple, and not too exaggerated.

**Output Format**: a comprehensive, conversational list of video captions that describe the motion in the following JSON format:

{
    "motion_type": "Short description of the motion",
    "video_caption": "Detailed description of the video caption"
}

user input:
"""
```

(Optional) Given the generated "meta" prompt from Llama3, we refine it to a standard template for GPT-4o following a unified structure as follows.

Meta Prompt Refinement

```
"""
**Objective**: Provide a highly descriptive video caption based on the input image and user input. As an expert, delve deeply into the image with a discerning eye, leveraging rich creativity and meticulous thought. When describing the details of the image, include appropriate dynamic information to ensure that the video caption contains reasonable actions and plots. If the user input is not empty, refine the caption provided by the user.

**Note**: The input image is the first frame of the video, and the output video caption should describe the motion starting from the current image. The user input is a meta prompt used as a template.

**Important**: Do not include camera transitions, screen switching, or perspective shifts.
```

```
**Answering Style**: Answers should be comprehensive, conversational, and use complete sentences. The answer should be in English regardless of the user's input. Provide context where necessary and maintain a consistent tone. Begin directly without introductory phrases like "The image/video showcases" or "The photo captures." The generated prompt should include two parts separated by "As time progresses." The first part before "As time progresses" should include a detailed description of the object's appearance, expression, and posture in the initial state. The second part after "As time progresses" should be a detailed description of the object's motion.

**output Format**: "[Highly descriptive image caption here]"

user input:
"""
```

C.2. Single-view Video Generation

We employ the text-conditioned image-to-video model CogVideoX5B-I2V [22] for single-view video generation. The pre-processed image with a white background serves as the image condition, while we leverage a GPT-4o generated motion prompt as the text condition. We set the number of CogVideoX inference steps to 50, and the generated videos have a frame rate of 8 FPS. We then subsample the generated 50 frames to 21 frames. To ensure sufficient motion while filtering out excessive movement, we use RAFT [16] to estimate the optical flow within the instance mask, using the motion magnitude as a criterion. Videos exceeding a predefined flow threshold are discarded. And to further filter the low-quality videos, we use an MLLM VideoScore [7] to automatically assess the generated video in terms of visual quality, temporal consistency, dynamic degree, text-to-video alignment, and factual consistency. We filter out the low-quality videos based on the predefined score threshold. Notably, our framework is agnostic to the video model. We provide quantitative ablation results on different video diffusion models for motion quality and diversity in Tab. 2.

C.3. Multi-view Video Generation

We employ SV3D [19] to generate novel views for the input single-view image. To maintain the object identity, all videos share the same object multi-views. Then we use SV4D [21] to condition on the CogVideoX single-view video and SV3D multi-view images to generate novel multi-view videos for the target object motion.

D. Implementation Details

D.1. Architecture Details

We adopt an auto-decoder architecture for learning the latent space. Specifically, we parameterize a Gaussian distribution using learnable mean μ and variance σ with a dimension of 32, initialized as a standard Gaussian distribution. Using the reparameterization trick, we sample a latent code $\boldsymbol{z} \sim \mathcal{N}(\mu, \sigma)$, which is then decoded by the latent code-conditioned motion decoder $\mathcal{D}_c$ into key point 6DoF transformations $\mathcal{E} \in SE(3)$. For motion auto-decoder, we adopt an MLP comprising 8 fully connected layers each of which is applied with weight-normalization, and each intermediate vectors with the dimension 256 are processed with RELU activation. A skip connection is included at the fourth layer. The network takes as input the latent vector, the positional embedding [11] of the key points' time and canonical position, with respective frequencies of 6 and 10.

D.2. Training and Testing Details

All experiments were conducted using a single 40GB A100 GPU for both training and testing. The learning rates for the latent vectors, decoder parameters, canonical key point positions, and global control radius were set to 5e-3, 2e-4, 2e-6, and 1e-2, respectively. For the single-motion overfitting setting, we trained the first stage for 1,000 steps and the second stage for 3,000 steps. For the 50-motion joint training setting, we trained the first stage for 2,800 steps and the second stage for 8,000 steps.

Chamfer Distance Regularization. To improve the training stability and efficiency, we leverage the canonical key point positions obtained in the motion pre-training stage (*Stage 1*) to guide the motion decoder's predictions in the motion geometry joint training stage (*Stage 2*) using the Chamfer distance loss. Given the key point poses $\mathcal{E}_1$ from *Stage 1* and $\mathcal{E}_2$ decoded during *Stage 2*, the Chamfer distance [6] is computed as the sum of the nearest-neighbor distances between each point in one set and the closest point in the other set, formed as:

$$d_{CD}\left(\mathcal{E}_1, \mathcal{E}_2\right) = \sum_{x \in \mathcal{E}_1} \min_{y \in \mathcal{E}_2} \|x - y\|_2^2 + \sum_{y \in \mathcal{E}_2} \min_{x \in \mathcal{E}_1} \|x - y\|_2^2 \tag{1}$$

Stability. As shown in Fig. 4, jointly modeling diverse motion patterns into a single generative model can effec-

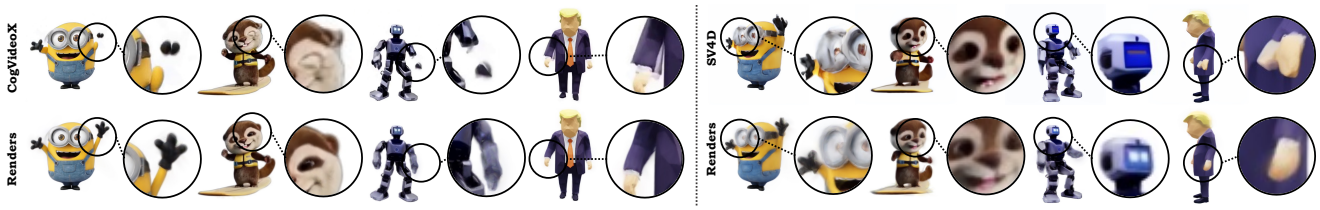

Figure 4. **Robustness against inconsistent video supervision.** By jointly modeling diverse motion patterns into a single generative model, we can extract the common 4D structure and a continuous motion space; thus, the generation results will be less sensitive to the low-quality or inconsistency of single motion video supervision generated at the motion prior distillation stage.

tively enhance the robustness of motion reconstruction over high-frequency errors (e.g, artifacts, missing parts) or appearance inconsistency of video supervision. To ensure robust and stable training of different modules, we propose a *coarse-to-fine* motion pre-training schedule by *disentangling* 3D motion and geometry. In the first stage, we pre-train the latent codes and canonical keypoint trajectories to obtain a reasonable motion space. During the second stage, we jointly model motion and geometry with Chamfer Distance and ARAP regularization to improve the training stability. Both multi-motion joint training and motion pretraining lead to satisfactory stability and robustness, as shown in Fig. 4, Fig. 5, and Fig. 6.

E. Experiment Details

E.1. User Study

We conducted human evaluations (user studies) through an anonymous Google Form to assess the diversity and quality of our generated 3D motion and 4D assets.

For 3D motion generation, we use 7 cases including synthetic Humans, Cats, and in-the-wild WALL-E Robots, Birds, and SORA [3] generated in-the-wild Kangaroos, Monsters, and Otters. For comparison, we generated 5 motion sequences for each species; thus, the total example number for each method is 35. For Image-to-4D generation, we use 12 images from Animate124 [26] benchmark and 4 images from Objaverse [5] dataset. For Text-to-4D generation, we use 6 examples from Animate124 [26] benchmark and each image with 10 text prompts generated by GPT [1].

E.2. Comparisons

3D Motion Generation. We evaluate the diversity and quality of our generated 3D motions by comparing them with the baseline methods DreamGaussian4D (DG4D) [15] and 4DGen [23]. Since the 4D generation and reconstruction code for the recent work Diffusion4D [10] has not yet been released, we are unable to include it in our comparisons. We provide qualitative comparison results in the demo video. For quantitative comparison, we conduct user studies on four metrics: motion diversity, image alignment, motion overall quality, and 3D appearance. For each case, we render the 4D outputs at three different timestamps from both the reference view and two novel views.

Image-to-4D. To further evaluate the visual quality of the generated 3D motions from a single image, we conduct comparisons in the Image-to-4D setting. We compare our method with the baseline methods Animate124 [26], 4DGen [23], STAG4D [24], DreamGaussian4D [15], and SV4D [24]. Following [15, 26], we evaluate the image-video alignment by calculating the cosine similarity between the CLIP visual features of each rendered frame and the reference image and evaluate the temporal consistency between frames by calculating the cosine similarity between CLIP visual features of every two consecutive rendered frames. Since SV4D has not released its 4D reconstruction code, we re-implemented it according to the official implementation. For other baselines, we use their official codes to generate the results. To ensure a fair comparison, we use the same reference video for all methods.

Video-to-4D. We evaluate the 4D reconstruction quality in the per-motion Video-to-4D setting using the widely-used Consistent4D [8] benchmark. We compare our method with the baseline methods Consistent4D [8], STAG4D [24], DreamGaussian4D [15], 4DGen [23] and SV4D [21]. Following [8, 21, 24], we use Learned Perceptual Similarity (LPIPS [25]) and CLIP-score (CLIP-S [13]) to evaluate the visual quality. We also evaluate the video temporal consistency by reporting FVD [18], a video-level metric commonly used in video generation tasks. Qualitative comparison results are also provided in the demo video.

Text-to-4D. We evaluate the text-guided 4D generation quality using the standard Animate124 [26] benchmark. We compare our method with the current SOTA text-to-4D models Animate124 [26] and 4D-fy [2]. Animate124 takes both the input image and the text description as input while 4D-fy only takes the text description as input. To ensure a fair comparison, we use the same text prompts generated from GPT [1] for all baselines and use the same input image for Animate124 comparison. We use RAFT [16] to estimate optical flow strengths between consecutive frames of a rendered video, and then compute the average of the largest 20% optical flows as the Motion Amplitude. Following [2],

we conduct user studies to evaluate the text alignment and motion diversity.

E.3. Ablations

Multi-Motion Joint Optimization. Incorporating multiple motion patterns into a single generative model can significantly enhance the robustness of motion reconstruction (Fig. 4). In Fig. 5, we compare the distributions of key points in the canonical space obtained from multi-motion joint training and per-motion optimization approaches. Our results indicate that multi-motion joint training yields a more uniform distribution of canonical key points across each rigid part, whereas per-motion optimization leads to a disorganized and inconsistent distribution. Since the canonical key points serve as the motion basis in our approach, these findings demonstrate that learning a variety of motion patterns contributes to a more coherent 4D representation.

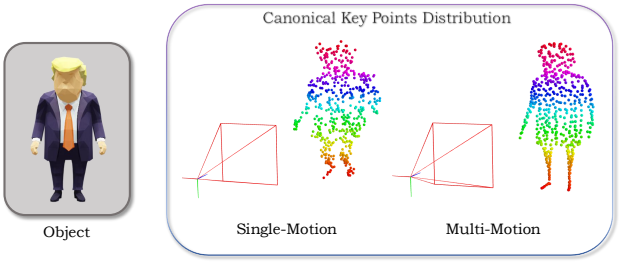

Figure 5. **Visualization of canonical key points distribution.**

Motion Pre-Training. A two-stage training schedule that incorporates motion pre-training (the first stage) is crucial for achieving robust and precise motion reconstruction. As shown in Fig. 6, the model without motion pre-training exhibits blurry results and unfaithful motion.

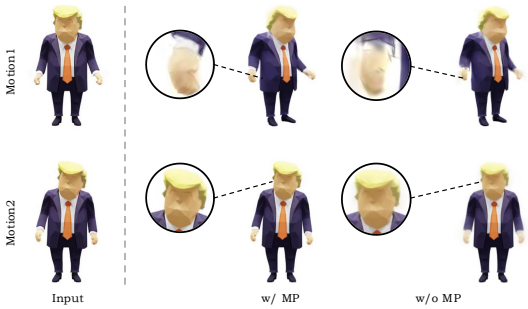

Figure 6. **Qualitative evaluation of Motion Pre-Training.**

Video Models. Our framework is *agnostic* to the video model, which is the current *bottleneck* of motion quality. To make the evidence more explicit, we now add an ablation in Tab. 2, in which we attach advanced video models to our pipeline and drive each with the *same 50 motion prompts*. The results indicate that motion quality (complex, non-repetitive aspects) and diversity can improve greatly with stronger video models, indicating that improvements of these models are critical for enhancing our performance.

Table 2. **Ablation on video models.** The quantitative results demonstrate that advanced video model can improve our motion quality and diversity.

Video Model	Motion Diversity ↑	Motion Quality ↑
CogVideoX1.5-5B	18.92%	21.62%
Wan2.1-I2V-14B [20]	29.73%	35.14%
Kling1.6-I2V-10B [9]	51.35%	43.24%

F. Limitations and Future Directions

While our DIMO demonstrates promising results in diverse 3D motion generation, several areas remain for future improvement. The major bottleneck of our work is its reliance on video generation models to distill motion and geometry priors, which currently have limited capabilities when handling complex real-world objects or those with extremely large-scale motions. Advancements in these video generation models would enhance our performance greatly.

At present, we use the shared structured keypoint motion graph with constrained and diverse latent vectors to represent 3D motions. Incorporating more physics-aware models, such as articulation models, could improve downstream applications like cross-species motion transfer. A potential avenue for future research involves the joint discovery of articulation structures in conjunction with video training.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 4

[2] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7996–8006, 2024. 4

[3] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators, 2024. 4

[4] Wen-Hsuan Chu, Lei Ke, and Katerina Fragkiadaki. Dreamscene4d: Dynamic multi-object scene generation from monocular videos. *arXiv preprint arXiv:2405.02280*, 2024. 1, 2

[5] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 1, 4

[6] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 3

[7] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang,

Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *arXiv preprint arXiv:2406.15252*, 2024. 3

[8] Yanqin Jiang, Li Zhang, Jin Gao, Weimin Hu, and Yao Yao. Consistent4d: Consistent 360 {\deg} dynamic object generation from monocular video. *arXiv preprint arXiv:2311.02848*, 2023. 4

[9] Kuaishou. Kling ai. https://klingai.kuaishou.com/, 2024. 5

[10] Hanwen Liang, Yuyang Yin, Dejia Xu, Hanxue Liang, Zhangyang Wang, Konstantinos N Plataniotis, Yao Zhao, and Yunchao Wei. Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models. *arXiv preprint arXiv:2405.16645*, 2024. 4

[11] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[12] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 1, 2

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 4

[14] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1

[15] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023. 4

[16] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 3, 4

[17] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[18] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 4

[19] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*, 2024. 1, 3

[20] 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. 5

[21] Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. *arXiv preprint arXiv:2407.17470*, 2024. 1, 3, 4

[22] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1, 3

[23] Yuyang Yin, Dejia Xu, Zhangyang Wang, Yao Zhao, and Yunchao Wei. 4dgen: Grounded 4d content generation with spatial-temporal consistency. *arXiv preprint arXiv:2312.17225*, 2023. 4

[24] Yifei Zeng, Yanqin Jiang, Siyu Zhu, Yuanxun Lu, Youtian Lin, Hao Zhu, Weiming Hu, Xun Cao, and Yao Yao. Stag4d: Spatial-temporal anchored generative 4d gaussians. *arXiv preprint arXiv:2403.14939*, 2024. 4

[25] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1, 4

[26] Yuyang Zhao, Zhiwen Yan, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Animate124: Animating one image to 4d dynamic scene. *arXiv preprint arXiv:2311.14603*, 2023. 4