
Supplementary Material for Imagine360: Immersive 360 Video Generation from Perspective Anchor

Jing Tan^{1*} Shuai Yang^{2, 5*} Tong Wu^{3†} Jingwen He¹ Yuwei Guo¹
Ziwei Liu⁴ Dahua Lin^{1, 5†}

¹The Chinese University of Hong Kong ²Shanghai Jiao Tong University

³Stanford University ⁴S-Lab, Nanyang Technological University

⁵Shanghai Artificial Intelligence Laboratory

A More Experiment Settings

A.1 User Study

For human evaluation, we ask the users to evaluate the 360 videos across three dimensions: **Graphics Quality (GQ)**, **Structure Plausibility (SP)**, and **Temporal Coherence (TC)**. We provide both the 360 videos and their four perspective projection videos with $\phi = 0, \theta = [0, 90^\circ, 180^\circ, 270^\circ]$. Graphics quality refers to the clarity and detail richness of the panorama and perspective video frames. Structure plausibility refers to the level of distortion in each perspective projections. Temporal coherence refers to the motion consistency and subject consistency: whether there're objects that suddenly appear or disappear, etc. The users need to look at both the 360 videos and the perspective projections to evaluate across the three metrics. They are asked to score each method from 1 to 4, where a higher score indicates better performance. The average user score is reported in Table 1 of the main paper.

A.2 Ablation Settings

For ablations, in the ablation on antipodal mask, we compare the model variant using both the antipodal mask and the directly-mapped mask with the model variant using only the directly-mapped mask. Note that we do not ablate on the directly mapped mask because its effectiveness in dual-design structure was already addressed in PanFusion [32]. In the ablation on rotation-aware design ablation, we compare our full model using both the rotation sampling and rotation estimation with a variant that does not use either of the designs to see the effect of rotation handling.

A.3 Testing Benchmark

To validate on both generated videos and in-the-wild videos, we randomly sample videos from 360-1M, Realestate-10K, and use CogVideoX to generate videos to form this testing benchmark. The 360-1M videos are panoramic videos, and we also sample different camera trajectories to crop perspective videos from these panoramic videos as inference inputs. The original 360 videos serve as the ground truth for video optical flow EPE calculations. We ask LLM to generate 200 scene description prompts that cover indoor, outdoor, urban, and landscape scenarios for CogVideoX. In addition, we ask the LLM to make 100 out of the 200 prompts not contain camera translation but only camera rotation. Amongst these generated videos, we manually pick 100 videos with only camera rotation and ablate the antipodal masking on this subset to see if there is negative impact on the visual quality from antipodal masking.



Figure A: Failure cases.

A.4 Stress Testing Analysis

To probe the limitations of our model, we select several categories of representative corner cases for stress testing. As illustrated in Fig. C, in the left case, when the input video features a monotonous background with cluttered foreground objects and an ambiguous spatial layout, the generated panoramic video is prone to a loss of coherent panoramic structure. In the right case, when the distance between consecutive camera poses is excessively large or the motion trajectory contains overly abrupt rotational changes, the pose estimation in the preprocess stage is likely to fail, consequently leading to a degradation in the quality of the final panoramic video. The above analysis shows that the main challenges of the model lie in two aspects: first, when the foreground is cluttered and the spatial cues are blurred, it is difficult to ensure the structure of the generated content; second, when the camera moves violently, resulting in inaccurate pose estimation, it will directly lead to a decline in generation quality.

B More Implementation Details

B.1 Data Collection

We begin by collecting a large-scale, coarse 360° video dataset from YouTube using a combination of keywords such as “360,” “video,” “trip,” “tour,” “wildlife,” “animal,” and “vehicle,” among others. Then we extract a random frame from each video as a reference to manually filter out videos with large logos, missing polar views, or that do not exhibit 360° equirectangular property. After this step, a smaller, high-quality subset is obtained from the original coarse dataset. We employ a shot segmentation model [24] to divide the long videos into a number of short segments with a smooth camera trajectory. These short segments still contain low-quality content, as the initial filtering is only based on random frames. We filter out static videos based on extracted optical flow. The flow values are first normalized to the range of $[0, 1]$, with an average flow value calculated for each frame. Videos that contain less than 10% of frames with > 0.1 average flow value are considered static and removed from the dataset. We also conduct another round of random frame sampling from these segments to find which segment contains irrelevant logos. Finally, we divide these segments uniformly into 5-second clips as the standardized training data.

B.2 Inference Settings.

We modified the VEnhancer [14] to keep the 360° close-loop property when interpolating the output panorama video for a better 360 VR immersive experience in the webpage. Note that we do not use video SR of any kind in our comparison and ablation experiments.

C Discussion with PanFusion

Building upon the dual-branch denoising architecture introduced in PanFusion [33], which was successful in text-to-panorama image generation, we adapt this design to a significantly different and more complex task: video-conditioned 360° video generation. Although both methods utilize a dual-branch structure, our work differs from PanFusion in several critical aspects.

First, the target domain and task are fundamentally different. PanFusion addresses static image generation guided by text, producing panoramic images on a spatial canvas. In contrast, our approach generates 360° videos conditioned on video inputs, requiring temporal modeling in spacetime and the ability to process video tokens in both the denoising and conditioning branches.

Second, the overall pipeline diverges substantially. Our model accepts general video inputs and incorporates rotation-aware mechanisms, such as rotation-aware mask sampling during training and a rotation estimation module at inference, to handle diverse real-world camera trajectories. PanFusion, by comparison, operates on static, upright panoramas from text prompts and lacks any camera-aware design.

Third, temporal modeling is a core component of our method, but is entirely absent in PanFusion. We integrate motion modules into the Stable Diffusion architecture and introduce an antipodal mask in the cross-domain attention module. This extends the receptive field beyond local neighborhoods to antipodal regions on the 360° sphere, enabling more coherent motion across frames.

In general, our task presents greater challenges than text-to-panorama generation due to the necessity of handling temporal dynamics, complex video conditions, and varying camera movements. Overall, PanFusion’s architecture is infeasible to solve our setting due to the different inputs. Even with adaptations for video generation, as we show in the ablations of our newly introduced modules, this variant remains insufficient to generate robust 360 videos for general video inputs and diverse camera trajectories. Thanks to our customized designs in antipodal masking and rotation-aware modules, Imagine360 achieves high-quality, video-conditioned 360° generation with strong generalization to in-the-wild inputs.

D Discussion on Panorama Image Outpaint

A bonus advantage of Imagine360 is that apart from panorama video outpainting, we also achieve superior performance for panorama image outpainting. We compared our method with state-of-the-art panorama image outpainting approaches, including Diffusion360 [7], PanoDiffusion [29] and SIG-SS [13]. We use the first frame of a video as the input image and extract the first frame of our outpainted video as our result for panorama image outpainting. For quantitative comparison, we adopt **Intra-Style** [9; 19] metric to evaluate the panorama style coherence; **CLIP** [15] to measure the alignment between the panorama and the input text prompts; **IQA** [28] measures the overall image quality. Tab. A shows our method achieves the best performance among the compared methods across all metrics.

We also show the qualitative comparison in Fig. B, with a red dashed box indicating the input image. The results of Diffusion360 [7] show less consistency as its newly generated pixels sharing different style with the know pixels. SIG-SS [13] is a GAN-based methods and its generations exhibits over-smoothness compared to diffusion-based approaches. PanoDiffusion [29] focuses on indoor scenes and does not generalize well to outdoor scenes. In contrast, Imagine360 outpaints more consistent, high-quality, and aesthetic panorama compared to other approaches.

Table A: **Quantitative comparison with the state-of-the-art Panorama Outpainting methods.**

Method	Intra-Style($\times 10^{-3}$) ↓	CLIP ↑	IQA ↑
Diffusion360 [7]	3.40	27.49	0.77
PanoDiffusion [29]	3.46	23.19	0.72
SIG-SS [13]	2.06	26.26	0.48
Ours	0.99	29.12	0.78

E Extended Ablations

E.1 Ablation on rotation-aware data sampling.

We currently employ random trajectory sampling to capture diverse camera motion conditions for training. Another intuitive strategy would be using camera pose estimation models on massive monocular videos or employ camera trajectory generation model [34] to collect a large camera trajectory library for training trajectory sampling. We collect a camera pose trajectory library based on Realestate10K [37] training samples, and randomly sample from these trajectories during training. The result are reported in Tab. B. We find the results less robust compared to our default random sampling. Realestate videos are mostly slow, and this data bias might downgrade inference performance on more rapid videos, leading to sub-optimal results.

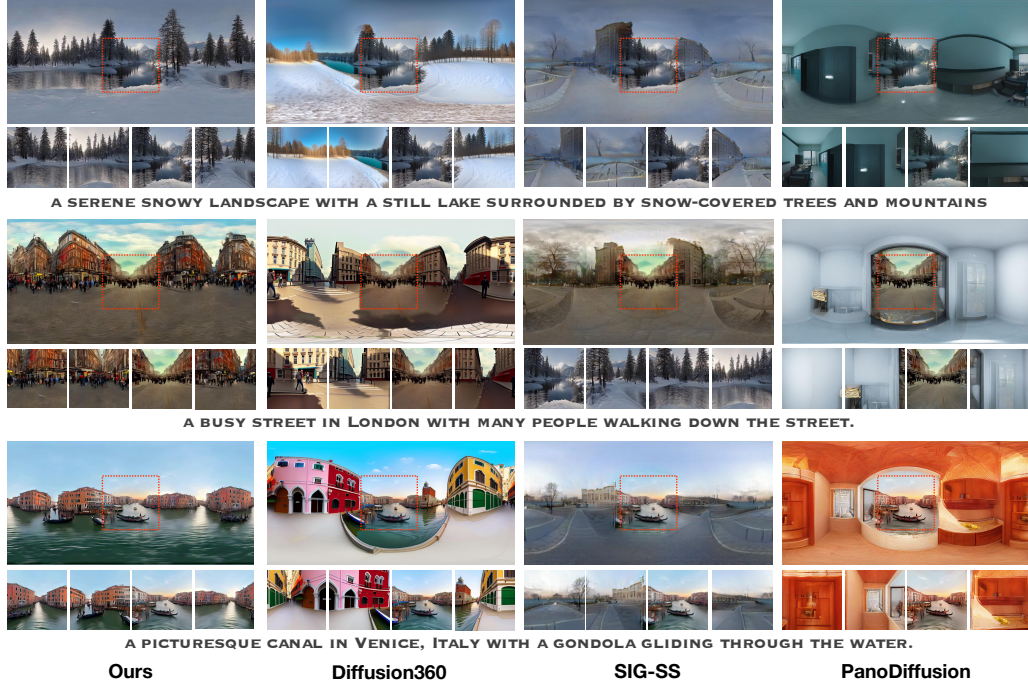


Figure B: Qualitative comparisons between Imagine360 and the state-of-the-art panorama outpainting methods.

Table B: Extensive ablation studies on training rotation sampling under Vbench, EPE, and OmniFVD metrics.

Method	Vbench				EPE ↓	OmniFVD ↓
	IQ ↑	AQ ↑	MS ↑	SC ↑		
Ours	0.7372	0.5722	0.9866	0.9649	2.5583	204.0
+ real rotation sampling	0.7314	0.5127	0.9773	0.9450	3.1049	272.9

E.2 Ablation on Extended Web Data

As we collect additional panorama video data from the web in our training, we also analyze the effect of the extended data. We train our baseline models on our training data and report the results of baselines+LoRA trained on our data (marked with *) in Tab. C. The results are tested on a subset of benchmark due to limited computation resource. Results show that most of the metrics improve with our training data, especially EPE, which reflects 360 motion correctness, showing the effectiveness of our curated 360 video data.

F Limitations and Future Work

During inference, Imagine360 leverages MonST3R to estimate the input camera poses. While MonST3R generally performs well on dynamic scenes, it can underestimate large rotations (e.g., 30° roll or 60° pitch). Hence, the derived Euler angles for mask generation may fail to fully compensate for the actual video rotation, leading to noticeable distortions that limit the video fidelity in certain scenarios, as seen in Fig. C. As geometry estimation research advances rapidly, this issue can be mitigated

Figure C: MonST3R can underestimate large camera rotations, leading to distorted input video in the canvas and hence inconsistent geometry in the generated results.

Table C: Ablative study on our extended panorama video dataset.

Method	Vbench				EPE ↓
	IQ ↑	AQ ↑	MS ↑	SC ↑	
FYC	0.6248	0.5368	0.9856	0.8770	3.1767
Animatediff	0.6257	0.5211	0.9799	0.9122	3.5393
360dvd	0.5501	0.4359	0.9856	0.9356	3.1904
FYC+LoRA*	0.6582	0.5565	0.9864	0.9633	2.9330
Animatediff+LoRA*	0.5707	0.4874	0.9861	0.8706	3.0068
360DVD*	0.6948	0.5245	0.9783	0.9635	3.1466
Ours	0.7372	0.5722	0.9866	0.9649	2.5583

by integrating more advanced models in future work.

G Societal Impacts

This paper proposes Imagine360, a video-conditioned 360 video generation framework to create immersive video content. On the positive side, our framework helps creators to produce VR-ready experiences from standard videos, lowering the production cost and expanding access to immersive media for general users without prior expertise in 360 content creation. However, like other generative models, this technology synthesizes new realistic pixels from limited viewpoints, which may lead to misinformation or potential privacy concerns. To mitigate such risks, we employ manual efforts to carefully remove training data with sensitive content and private information. Our codes will be made public under CC BY 4.0 license to prevent misuse.

H Dataset Availability Statement and Clarification

Our study uses only publicly available data, following precedents like InternVid, Panda-70M, and 360-1M. The data is solely for research, aligning with YouTube’s privacy and fair use policies. No user data or privacy rights are violated during the data collection process. We will only supply YouTube IDs and start, end timestamps for downloading the respective content. The dataset is made available under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

I Additional Related Work

I.1 Diffusion Models

Diffusion models [17; 22; 23] have achieved remarkable success in image generation [21; 20; 6], leading to advancements in video diffusion models [31; 12; 2; 26; 3; 4; 16]. The first video diffusion model (VDM) [18] adopts a space-time factorized U-Net in pixel space to model low-resolution videos. Imagen-Video [16] proposes to use cascaded DMs for generating high-definition videos. Subsequent research [2; 12; 3; 4; 26] adapts existing text-to-image (T2I) models to text-to-video (T2V) models by incorporating temporal layers, including both convolution and attention layers. More recently, several works [8; 31] directly use 3D full attention to model space-time information for more unified video representation. On the other hand, image-to-video (I2V) [35; 30; 11; 1; 36; 10] has arisen great attention as it enables more precise control on video generation. Some works achieve I2V by incorporating the image condition into the pretrained T2V models and finetuning newly added modules [30; 11] or the inherited weights [1; 35], while plug-and-play methods [36; 10] aims to turn any text-to-image models into image animators.

J Code and License

Our codebase is mainly built upon PanFusion¹ [33] protected by MIT License and Follow-Your-Canvas² [5] (does not specify an open-source license, therefore treated as "all rights reserved"). WEB360 [27] dataset is publicly accessible but does not specify an explicit license. Therefore, it is treated as "all rights reserved" and used only for non-commercial research purposes under fair use principles. Realestate10K [37] is licensed by Google LLC under a Creative Commons Attribution 4.0 International License. 360-1M [25] is protected by MIT license. CogVideoX [31] is protected by Apache-2.0 license.

References

- [1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.
- [2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.
- [3] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.
- [4] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *arXiv preprint arXiv:2401.09047*, 2024.
- [5] Qihua Chen, Yue Ma, Hongfa Wang, Junkun Yuan, Wenzhe Zhao, Qi Tian, Hongmei Wang, Shaobo Min, Qifeng Chen, and Wei Liu. Follow-your-canvas: Higher-resolution video outpainting with extensive content generation. *arXiv preprint arXiv:2409.01055*, 2024.
- [6] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.
- [7] Mengyang Feng, Jinlin Liu, Miaomiao Cui, and Xuansong Xie. Diffusion360: Seamless 360 degree panoramic image generation based on diffusion models, 2023.
- [8] Peng Gao, Le Zhuo, Ziyi Lin, Chris Liu, Junsong Chen, Ruoyi Du, Enze Xie, Xu Luo, Longtian Qiu, Yuhang Zhang, et al. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945*, 2024.
- [9] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.
- [10] Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Chongyang Ma, Weiming Hu, Zhengjun Zha, Haibin Huang, Pengfei Wan, et al. I2v-adapter: A general image-to-video adapter for video diffusion models. *arXiv preprint arXiv:2312.16693*, 2023.
- [11] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pages 330–348. Springer, 2025.
- [12] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.
- [13] Takayuki Hara, Yusuke Mukuta, and Tatsuya Harada. Spherical image generation from a few normal-field-of-view images by considering scene symmetry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–15, 2022.
- [14] Jingwen He, Tianfan Xue, Dongyang Liu, Xinqi Lin, Peng Gao, Dahua Lin, Yu Qiao, Wanli Ouyang, and Ziwei Liu. Venhancer: Generative space-time enhancement for video generation. *arXiv preprint arXiv:2407.07667*, 2024.
- [15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP (1)*, pages 7514–7528. Association for Computational Linguistics, 2021.
- [16] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.
- [17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

¹<https://github.com/chengzhag/PanFusion>

²<https://github.com/mayuelala/FollowYourCanvas>

- [18] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.
- [19] Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via synchronized joint diffusions, 2023.
- [20] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. *CoRR*, 2023.
- [21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [22] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [23] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- [24] Tomáš Souček and Jakub Lokoč. Transnet v2: An effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020.
- [25] Matthew Wallingford, Anand Bhattad, Aditya Kusupati, Vivek Ramanujan, Matt Deitke, Aniruddha Kembhavi, Roozbeh Mottaghi, Wei-Chiu Ma, and Ali Farhadi. From an image to a scene: Learning to imagine the world from a million 360° videos. *Advances in Neural Information Processing Systems*, 37:17743–17760, 2024.
- [26] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.
- [27] Qian Wang, Weiqi Li, Chong Mou, Xinhua Cheng, and Jian Zhang. 360dvd: Controllable panorama video generation with 360-degree video diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6913–6923, 2024.
- [28] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching Imms for visual scoring via discrete text-defined levels. In *ICML*. OpenReview.net, 2024.
- [29] Tianhao Wu, Chuanxia Zheng, and Tat-Jen Cham. Panodiffusion: 360-degree panorama outpainting via diffusion, 2024.
- [30] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023.
- [31] 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.
- [32] Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. Taming stable diffusion for text to 360 panorama image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6347–6357, 2024.
- [33] Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. Taming stable diffusion for text to 360 panorama image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6347–6357, 2024.
- [34] Mengchen Zhang, Tong Wu, Jing Tan, Ziwei Liu, Gordon Wetzstein, and Dahua Lin. Gendop: Autoregressive camera trajectory generation as a director of photography. *arXiv preprint arXiv:2504.07083*, 2025.
- [35] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023.
- [36] Yiming Zhang, Zhening Xing, Yanhong Zeng, Youqing Fang, and Kai Chen. Pia: Your personalized image animator via plug-and-play modules in text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7747–7756, 2024.
- [37] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.