RapidMV: Leveraging Spatio-Angular Latent Space for Efficient and Consistent Text-to-Multi-View Synthesis

Supplementary Material

A. Details of reconstruction evaluation

We elaborate on our reconstruction evaluation scheme, which was briefly mentioned in Sec. 4.3. Given F generated images, where $F \in \{24, 32\}$ in our experiments, we use even-numbered frames (half of the generated images) as the train set to train a NeRF, and use the odd-numbered frames (remaining half of the generated images) as the test set. This splitting strategy ensures that the training and test sets are mutually exclusive and that the model is evaluated on unseen viewpoints. For the NeRF, we use the nerfacc [23] implementation of Instant-NGP [28] for fast training. We use multi-resolution hash grids with a resolution of 128 and 4 levels. We used two separate MLPs to predict volume density and view-dependent color, and we utilized sigmoid for output layers to ensure outputs are in valid ranges. The learning rate was initialized at 1e-2 with a warm-up period of 100 steps, where the learning rate linearly increased from 1% to 100% of its maximum value. We used a multi-step scheduler that decayed the learning rate by a factor of 0.33 at 50%, 75%, and 90% of the total training steps. We used weight decay to regularize the model and prevent overfitting, set to 5e-4. The smooth L1 was used between the predicted and ground-truth RGB values to optimize the NeRF without additional objectives *e.g.* LPIPs loss, at an aim to benchmark the actual photometric multi-view consistency.

One issue we encountered was that we could not pre-determine the camera intrinsics of the generated views, as we apply the latent-wise anchor-pose modulation using the camera extrinsics only. Our training data rendered from Objaverse [6] had varying intrinsics, as we followed MVDream [38] to use a random field of view between $[15°, 60°]$ for improved diversity in renderings. While VideoMV [53] also reported consistency metrics for MVDream and VideoMV, their reconstruction pipeline was not released for reproduction. To resolve this, we devise a scheme to optimize a NeRF for each FoV in $[15°, 60°]$ at $5°$ intervals, iterating for 1,000 steps per configuration. We then identify the FoV that minimizes the PSNR and select the corresponding NeRF for further optimization with an additional 1,000 steps. This ensures that each method is evaluated as fairly as possible, leveraging its optimal multi-view consistency for the final comparison.

B. Image-conditioned RapidMV

In this section, we show that RapidMV can be optimized to generate multi-view images conditioned on not only text, but also image. The idea is simple; we concatenate the latent of the image prompt to the noisy multi-view latents, so that the multi-view latents can attend to the image prompt during the denoising process for an explicit image guidance. The overall implementation is motivated by the pixel controller of ImageDream [17, 46]. The frame-wise camera conditioning takes as input a zero vector as the camera extrinsics of the image prompt. The image prompt latent is not added with noise, and is not denoised during the diffusion process as well. For image-conditioned RapidMV, the image prompt latent is obtained by simply passing four copies of the image prompt through our spatio-angular VAE to obtain a single spatio-angular latent. The overall pipeline for image-conditioned RapidMV is illustrated in Fig. A1, and we provide qualitative examples in Fig. A2.

C. Filtering high-quality data from Objaverse

We mentioned in Sec. 3.4 that we decompose the training strategy, where we finally finetune our model on the high-quality subset of Objaverse [6]. It was explained in Sec. 4.1 that we filter out objects with less than 10 ‘likes’ in the metadata to collect our high-quality subset, which leaves around 70K objects. The efficacy of high-quality finetuning was demonstrated in Tab. 2 and Fig. 7.

In this section, we visualize some examples of our high-quality subset of Objaverse, in contrast to the objects which are not included in our high-quality subset, in Fig. A3. can be seen that our high-quality subset contains objects with more sophisticated and detailed geometry and texture. While it is not always the case that objects with lower than 10 likes counts have simple geometry and texture, the like count serves as a reliable metric to yield high-quality objects from the full dataset.

D. Comparison against Bootstrap3D

In this section, we evaluate RapidMV against Bootstrap3D [41], a concurrent 4-view genereation model that proposes to use (1) densified captions, (2) large-scale synthetic multi-view dataset and (3) Training-time step Reschedule (TTR) to better leverage the synthetic dataset. Their model and pretrained weights were not open-source at the time of submission, and we try to evaluate as fairly as possible by using the same evaluation dataset from GPTEval3D [49]. We do not have Bootstrap3D’s generated image set from PlayGround2.5 and PixArt-α for FID calculation, and therefore omit the FID value comparisons. The results are shown in Tab. A1, where it can be seen that RapidMV outperforms Bootstrap3D in terms of CLIP-Recall, while being competitive in terms of CLIP-score.

E. Drawbacks and future directions.

A drawback in the current version of RapidMV is that it generates multi-view images within a static orbit at fixed elevation. However, it has been shown in SV3D [44] that having a dynamic orbit, *i.e.*, varying elevation of camera poses covering more various viewpoints, is definitely beneficial in 3D reconstruction. This could be achieved by rendering views from the Objaverse [6] dataset at dynamic orbits for training, as the camera conditioning scheme would still be applicable to cameras in a dynamic orbit, and spatio-temporal compression would still be effective.

Another shortcoming of RapidMV is that even after finetuning, the VAE still is not perfect at alleviating blurry textures or motion

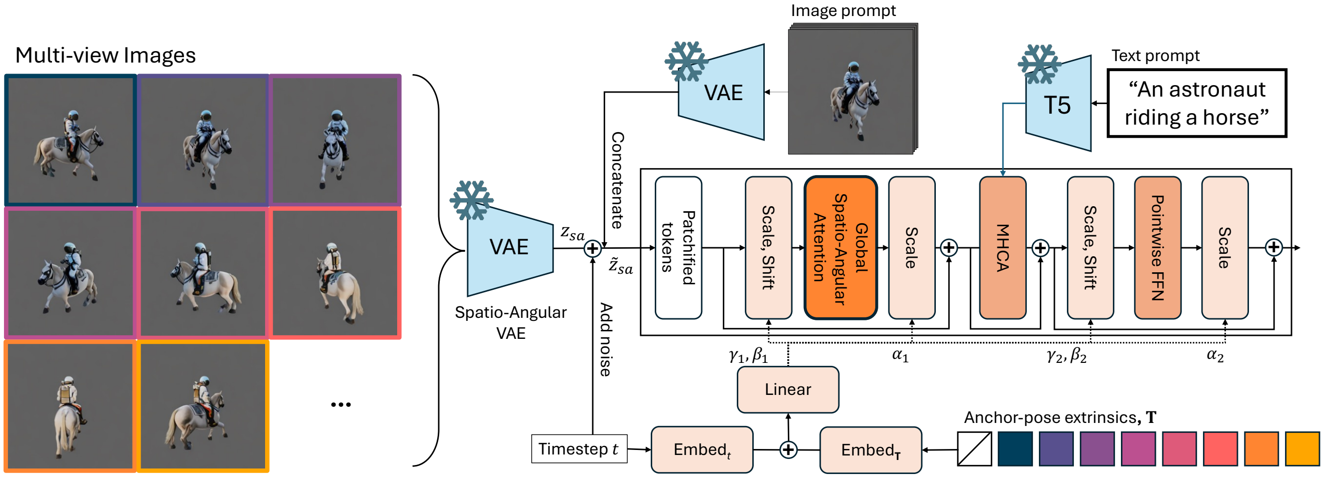


Figure A1. **Overview of Image-conditioned RapidMV**. Zero vectors are provided for the latent-wise anchor-pose modulation for the image prompt. The latent of the image prompt is concatenated to the *noisy* spatio-angular latents, as the image prompt latent should not be noisy, and is not denoised during the diffusion steps.

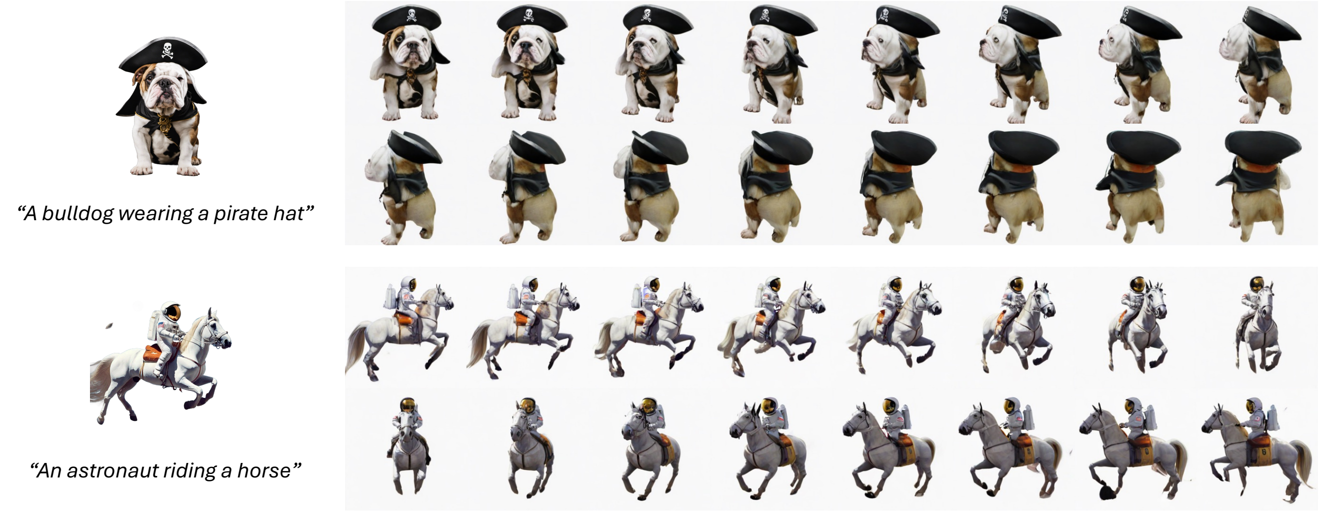

"A bulldog wearing a pirate hat"
"An astronaut riding a horse"

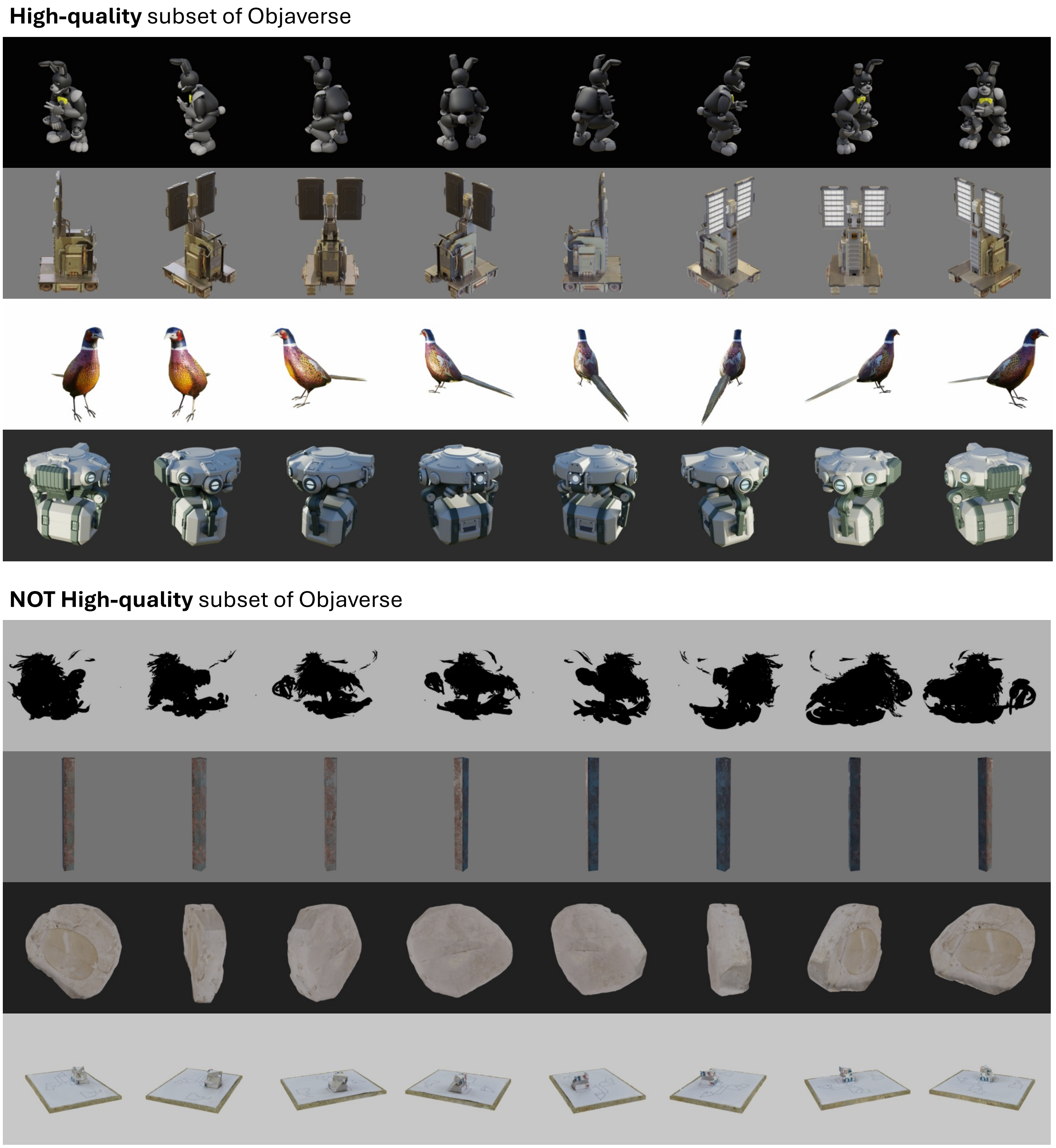
Figure A2. **Qualitative results of image-conditioned RapidMV**. We show that RapidMV can flexibly handle image prompts to generate 32 consistent views. We visualize 16 contiguous frames in this figure for better visibility.

blurs, as a compromise for the efficiency of spatio-angular latents. We hypothesize this is because each latent has to encode not only the appearance of the original image, but also the *angular viewpoint deviations* across 4 frames, which makes it challenging to seamlessly reconstruct the fine details. While our current spatio-angular VAE yields a 4-channel spatio-angular latent, more recent spatial VAEs [1, 7] and spatio-temporal VAEs [50, 51] pproduces 16-channel latents, which we conjecture would be more effective at capturing both the appearance and motion information accurately.

The blurring effects are particularly pronounced in the first frame of generation, which we conjecture is due to the *causal* 3D convolution layers within the spatio-angular VAE. We conjecture this can be solved if we propose to encode $1 + 4N$ frames, where the first frame is encoded separately to better preserve the details and to be usable for individual images, following recent spatio-temporal VAE structures [50].

F. Additional qualitative results.

In this section, we provide additional qualitative results of RapidMV on full 32 generated views. The results are shown in Fig. A4 to Fig. A6. RapidMV shows promising quality and high multi-view consistency and camera coherency, despite generating 32 images in just around 5 seconds. As mentioned in Appendix E, the first image of the generated multi-view images is more prone to blurs, which is strongly visible in the results of the prompt "Dragon armor, 3D asset".



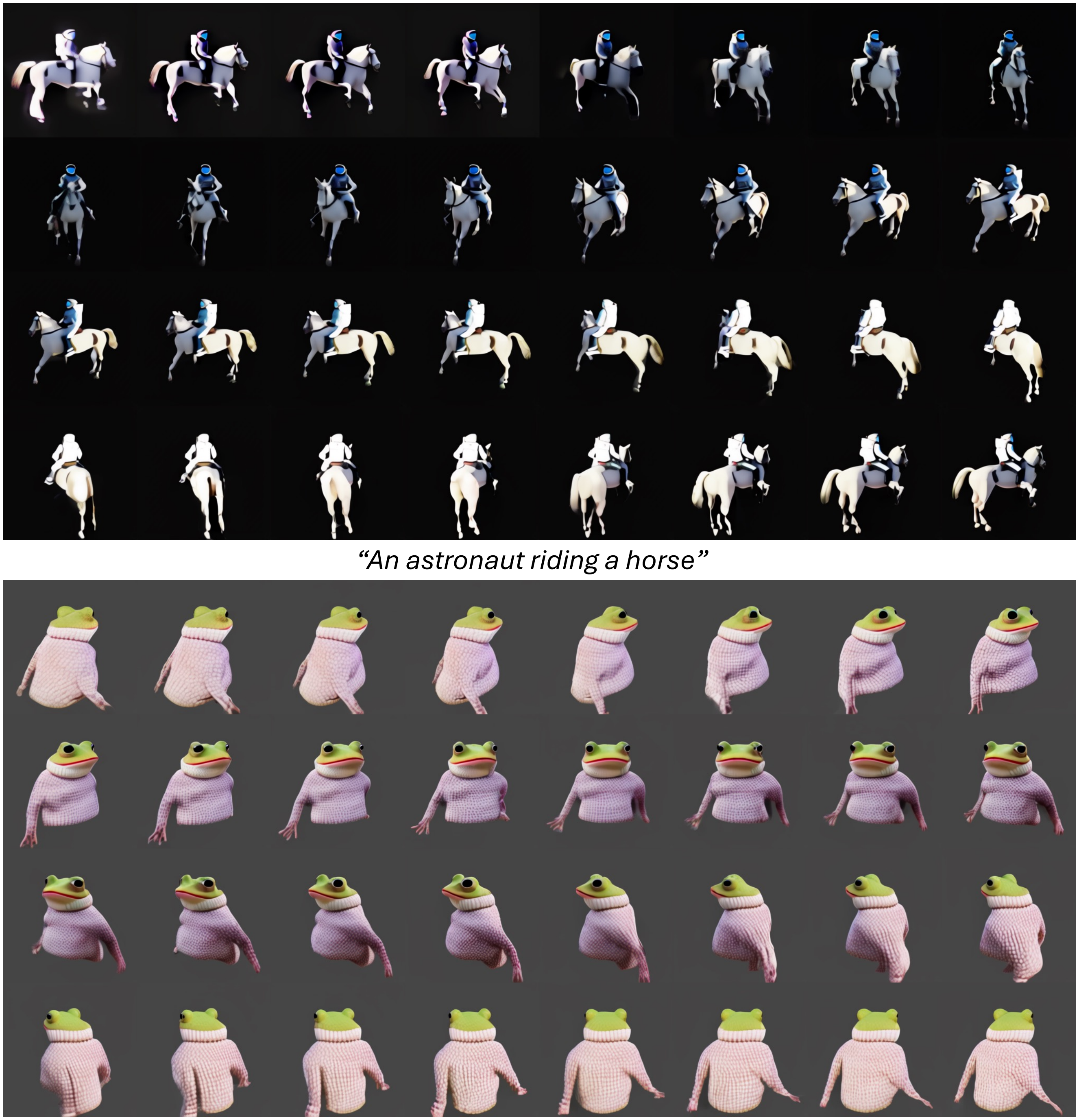
High-quality subset of Objaverse

NOT High-quality subset of Objaverse

Figure A3. **Visualization of high-quality subset of Objaverse [6].** It can be seen that our high-quality subset contains objects with more sophisticated and detailed geometry and texture. We filter out objects from the Objaverse data whose 'like' counts in the metadata is less than 10. While it is not always the case that objects with lower than 10 likes counts have simple geometry and texture, the like count serves as a reliable metric to yield high-quality objects from the full dataset.

Method	CLIP-R ↑		CLIP Score ↑	
	CLIP-L/14	CLIP-bigG	CLIP-L/14	CLIP-bigG
Instant3D [22]*	83.6	91.1	25.6	39.2
MVDream [38]	84.8	89.3	25.5	38.4
Bootstrap3D [41]	88.8	92.5	25.8	40.1
RapidMV$_s$ (ours)	**90.0**	**93.4**	**26.3**	39.5

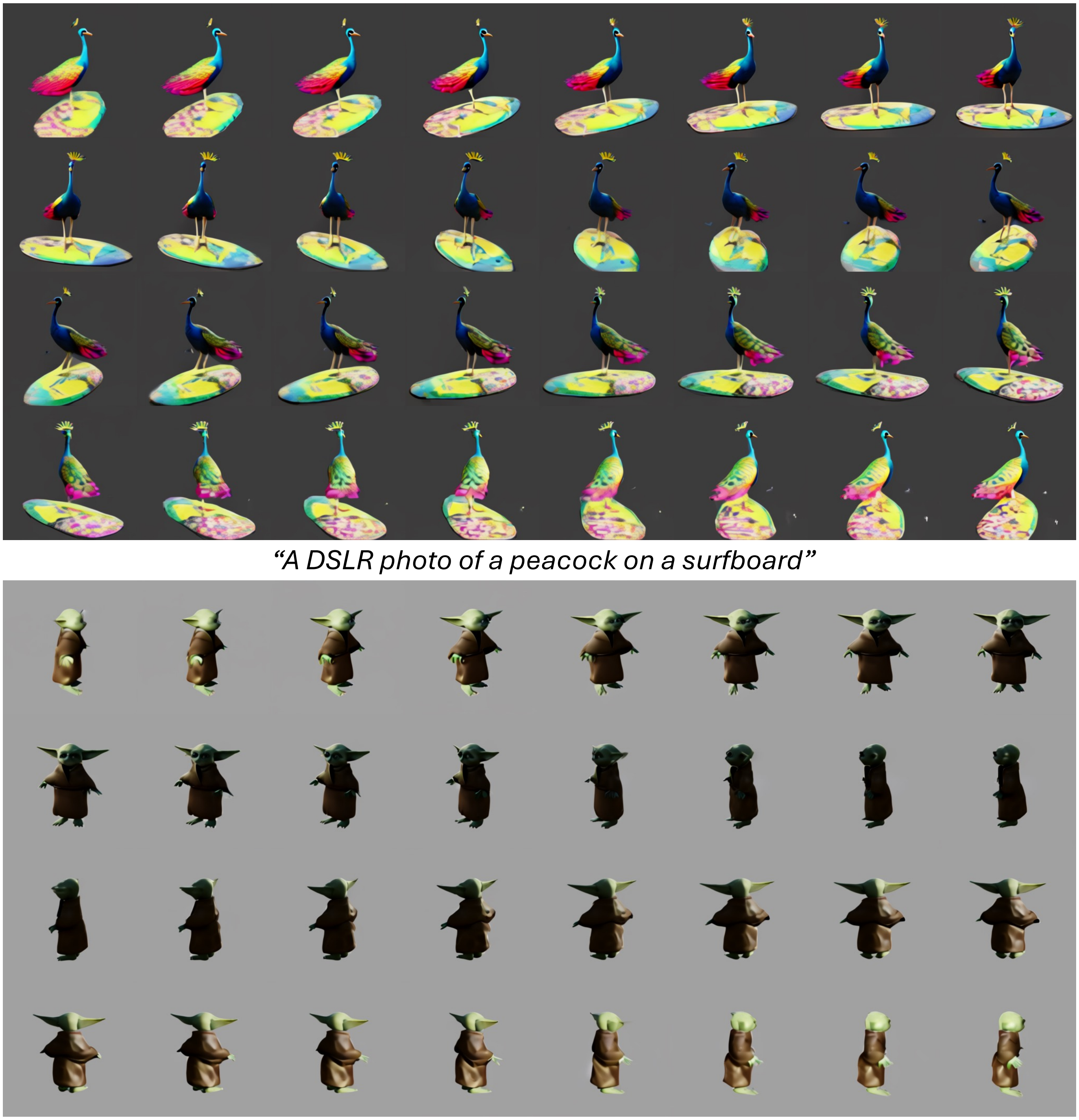
Table A1. **Quantitative comparison against Bootstrap3D [41] on 4 generated views.** The evaluation was performed on the 110 prompts from GPTEval3D [49]. Instant3D* [22] are results from an unofficial implementation by the authors of Bootstrap3D. All resolutions are at 256×256. The results show that our proposed RapidMV exhibits the best CLIP-R score overall, and the best CLIP-Score when using the CLIP-L/14 model [34] and the second-best when using the CLIP-bigG model [14].



“An astronaut riding a horse”

“A DSLR photo of a frog wearing a sweater”

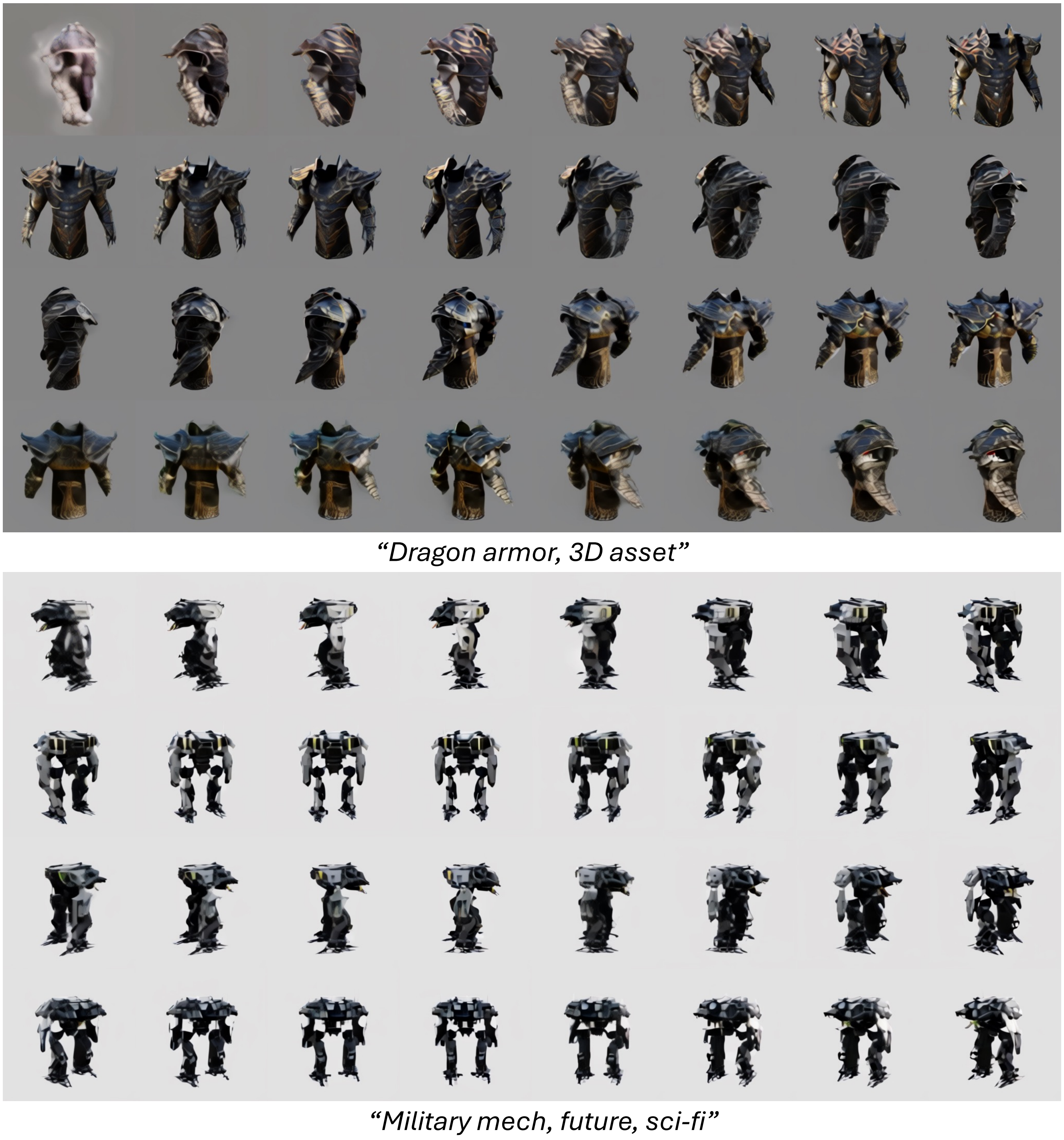
Figure A4. **Text-to-32-view results of RapidMV**. We visualize all 32 frames in this figure for better visibility. RapidMV shows promising quality and high multi-view consistency and camera coherency, despite generating 32 images in just around 5 seconds.



"A DSLR photo of a peacock on a surfboard"

"Baby Yoda in the style of a Mormookiee"

Figure A5. **Text-to-32-view results of RapidMV**. We visualize all 32 frames in this figure for better visibility. RapidMV shows promising quality and high multi-view consistency and camera coherency, despite generating 32 images in just around 5 seconds.



“Dragon armor, 3D asset”

“Military mech, future, sci-fi”

Figure A6. **Text-to-32-view results of RapidMV**. We visualize all 32 frames in this figure for better visibility. RapidMV shows promising quality and high multi-view consistency and camera coherency, despite generating 32 images in just around 5 seconds.

Acknowledgement. M. Cho and S. Kim were supported by the IITP grants (RS-2022-II220290: Visual Intelligence for Space-Time Understanding and Generation based on Multi-layered Visual Common Sense (50%), RS-2021-II212068: AI Innovation Hub (45%), RS-2019-II191906: AI Graduate School Program at POSTECH (5%)) funded by the Korea government (MSIT).

References

[1] Announcing FLUX1.1 [pro] and the BFL API. https://blackforestlabs.ai/announcing-flux-1-1-pro-and-the-bfl-api/. 2

[2] 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. 3, 4

[3] 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

[4] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-\alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 4, 5

[5] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-\sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*, 2024. 3, 4, 5, 6

[6] 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. 5, 6, 1, 3

[7] 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. 2

[8] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[9] Yuze He, Yushi Bai, Matthieu Lin, Wang Zhao, Yubin Hu, Jenny Sheng, Ran Yi, Juanzi Li, and Yong-Jin Liu. T3bench: Benchmarking current progress in text-to-3d generation. *arXiv preprint arXiv:2310.02977*, 2023. 5

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[11] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 3, 4

[12] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 2

[13] Zehuan Huang, Yuan-Chen Guo, Haoran Wang, Ran Yi, Lizhuang Ma, Yan-Pei Cao, and Lu Sheng. Mv-adapter: Multi-view consistent image generation made easy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16377–16387, 2025. 2

[14] Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. *Advances in Neural Information Processing Systems*, 35:29262–29277, 2022. 4

[15] Ryota Kaji and Keiji Yanai. Vq-vdm: Video diffusion models with 3d vqgan. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia*, pages 1–5, 2023. 3

[16] Seungwook Kim, Kejie Li, Xueqing Deng, Yichun Shi, Minsu Cho, and Peng Wang. Enhancing 3d fidelity of text-to-3d using cross-view correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10649–10658, 2024. 2

[17] Seungwook Kim, Yichun Shi, Kejie Li, Minsu Cho, and Peng Wang. Multi-view image prompted multi-view diffusion for improved 3d generation. *arXiv preprint arXiv:2404.17419*, 2024. 1

[18] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[19] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 5

[21] PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan, 2024. 3

[22] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023. 2, 4

[23] Ruilong Li, Hang Gao, Matthew Tancik, and Angjoo Kanazawa. Nerfacc: Efficient sampling accelerates nerfs. *arXiv preprint arXiv:2305.04966*, 2023. 7, 1

[24] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 2

[25] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[26] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 5

[27] 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. 2

[28] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 7, 1

[29] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2

[30] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 2

[31] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3

[32] 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. *arXiv preprint arXiv:2307.01952*, 2023. 2, 6

[33] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[34] 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. 4

[35] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv e-prints*, pages arXiv–2204, 2022. 2

[36] 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. 2

[37] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

[38] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2, 3, 4, 5, 6, 7, 1

[39] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[41] Zeyi Sun, Tong Wu, Pan Zhang, Yuhang Zang, Xiaoyi Dong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Bootstrap3d: Improving 3d content creation with synthetic data. *arXiv preprint arXiv:2406.00093*, 2024. 2, 5, 1, 4

[42] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2025. 2

[43] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*, 2022. 3

[44] 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. In *European Conference on Computer Vision*, pages 439–457. Springer, 2025. 1

[45] Jacob Walker, Ali Razavi, and Aäron van den Oord. Predicting video with vqvae. *arXiv preprint arXiv:2103.01950*, 2021. 3

[46] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 1

[47] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 2

[48] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 2

[49] Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22227–22238, 2024. 5, 1, 4

[50] 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. 3, 2

[51] Sijie Zhao, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Muyao Niu, Xiaoyu Li, Wenbo Hu, and Ying Shan. Cv-vae: A compatible video vae for latent generative video models. *arXiv preprint arXiv:2405.20279*, 2024. 2

[52] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. 3, 4, 5, 6, 8

[53] Qi Zuo, Xiaodong Gu, Lingteng Qiu, Yuan Dong, Zhengyi Zhao, Weihao Yuan, Rui Peng, Siyu Zhu, Zilong Dong, Liefeng Bo, et al. Videomv: Consistent multi-view generation based on large video generative model. *arXiv preprint arXiv:2403.12010*, 2024. 2, 4, 5, 6, 7, 1