# Vid3D: Synthesis of Dynamic 3D Scenes using 2D Video Diffusion

**Rishab Parthasarathy** [*1] **Zachary Ankner** [*1 2] **Aaron Gokaslan** [2 3]

## Abstract

A recent frontier in computer vision has been the task of 3D video generation, which consists of generating a time-varying 3D representation of a scene. To generate dynamic 3D scenes, current methods explicitly model 3D temporal dynamics by jointly optimizing for consistency across both time and views of the scene. In this paper, we instead investigate whether it is necessary to explicitly enforce multiview consistency over time, as current approaches do, or if it is sufficient for a model to generate 3D representations of each timestep independently. We hence propose a model, Vid3D, that leverages 2D video diffusion to generate 3D videos by first generating a 2D "seed" of the video's temporal dynamics and then independently generating a 3D representation for each timestep in the seed video. We evaluate Vid3D against two state-of-the-art 3D video generation methods and find that Vid3D is achieves comparable results despite not explicitly modeling 3D temporal dynamics. We further ablate how the quality of Vid3D depends on the number of views generated per frame. While we observe some degradation with fewer views, performance degradation remains minor. Our results thus suggest that 3D temporal knowledge may not be necessary to generate high-quality dynamic 3D scenes, potentially enabling simpler generative algorithms for this task.

## 1. Introduction

A key theme throughout the evolution of computer vision has been the modeling of new modalities that better capture the complexity of the real world (Brooks et al., 2024; Kerbl et al., 2023b; Mildenhall et al., 2020). A less explored and currently emerging frontier in computer vision is modelling dynamic 3D scenes instead of static 3D scenes. In this work, we tackle the challenge of efficiently generating high-quality 3D videos.

Recent works have proposed methods for 3D video generation that fall into one of two paradigms. In the first paradigm, classifier guidance from 2D video models and static 3D scene models is leveraged to optimize a dynamic 3D scene representation along multiple axes (Zhao et al., 2023; Bahmani et al., 2023). In the second paradigm, a 3D representation is constructed for the first frame and then deformed over time to be consistent with a 2D rendering of the scene (Ren et al., 2023). While these results achieve impressive results, they still have significant limitations.

Methods that rely on classifier guidance are extremely computationally intensive, requiring hours to generate a single 3D video. In contrast, methods that deform an initial 3D scene representation are more efficient, but they instead require rigid temporal structure. Furthermore, both paradigms are very technically complex, introducing many hyperparameters that must be tuned.

In this work, we thus investigate the common assumption that 3D temporal dynamics need to be explicitly modelled to perform 3D video generation. Specifically, we consider whether a model can generate high-quality 3D videos without explicitly enforcing consistency between multiple views over time. We propose Vid3D, a simple end-to-end pipeline for synthesizing dynamic 3D scenes from a single reference image. We generate 3D videos in three steps: (1) generating a 2D video outline for the 3D video, (2) independently generating multiple views for each timestep in the seed video, and (3) using the multiple views to generate 3D representations of each frame. We evaluate Vid3D against two state-of-the-art 3D video generation methods, finding that it achieves comparable quality even without explicitly modelling the 3D temporal dynamics of the scene. We further ablate key parameters of Vid3D to gain a deeper understanding of the method. Overall, we challenge the current assumption that methods require explicit temporal consistency to achieve SOTA results in the important task of 3D video generation (Ren et al., 2023; Zhao et al., 2023).

[*]Equal contribution [1]Department of EECS, MIT, Cambridge, MA, USA [2]Databricks Mosaic AI, San Francisco, CA, USA [3]Department of Computer Science, Cornell, Ithaca, NY, USA. Correspondence to: Rishab Parthasarathy <rpartha@mit.edu>.

*First Workshop on Controllable Video Generation at ICML 2024*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

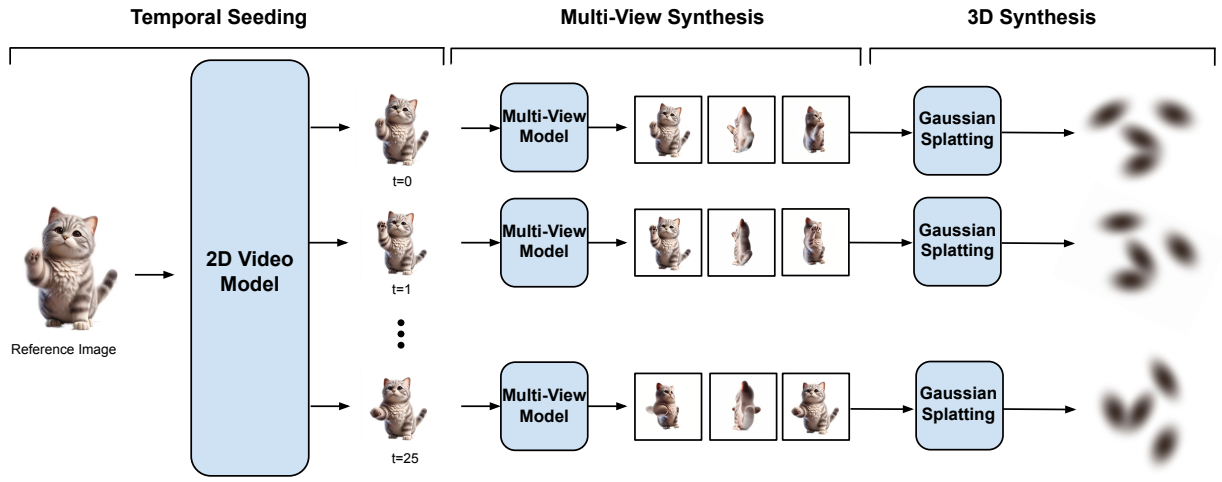


*Figure 1.* An overview of the Vid3D method. In stage 1, we generate a 2D video from a reference image to seed the dynamics of the scene. In stage 2, we generate multiple views for each timestep in the 2D video. In stage 3, we train a Gaussian splat on the collection of views from each timestep. Ultimately, each trained Gaussian splat represents a timestep in the 3D video.

## 2. Our Approach

We hypothesize that it is possible to generate dynamic 3D scenes from a source image without explicitly modeling 3D temporal dynamics, meaning that it is not necessary to enforce multi-view consistency across time. To create a model capable of generating 3D videos without modeling 3D temporal dynamics, we factorize the task into generating the 2D temporal dynamics of the scene and then generating 3D representations of each timestep in the 2D scene. Specifically, we propose a three-step process for generating 3D scenes from a source image: (1) seeding the scene dynamics with a 2D video, (2) synthesizing multiple views of the scene for each time-step (also known as a frame) in the seed video, and (3) generating a 3D representation of each time step based on the generated views. We present a diagram of our Vid3D method in Figure 1.

To construct the dynamics of the scene, which we refer to as *temporal seeding*, we query a 2D video model with the reference image. This seeding provides us with a dynamic rendering of the object from a singular view. We next transform the obtained 2D representation of the dynamic scene into a 3D representation. As we do not enforce multi-view consistency across time, we generate the 3D representation of each frame in our 2D seed independently of all other frames. To generate 3D representations of each 2D frame we follow the methodology proposed by Chen et al. (2024). For each seed frame, we generate multiple views of the target object by querying a 2D video model finetuned for multi-view generation on a curated subset of the Objaverse dataset (Deitke et al., 2023). For each frame, we then train a Gaussian Splat (Kerbl et al., 2023a) on the collection of generated views. The sequence of Gaussian Splats ultimately defines a 3D video.

### 2.1. Model Choices

We use the Stable Video Diffusion model (Blattmann et al., 2023) to generate 2D temporal seeds. The model generates seeds that consist of 25 frames. The seed generations are conditioned on a "motion score" that controls the degree of motion in the video; our main results use a motion score of 120. To generate multiple views for each timestep in the seed video, we use a version of Stable Video Diffusion fine-tuned for the multiview generation task on Objaverse (Chen et al., 2024; Voleti et al., 2024; Deitke et al., 2023). We select this model over Zero 1-to-3 as it generates higher quality multi-views over a number of quantitative and qualitative metrics including human evaluation (Liu et al., 2023; Chen et al., 2024; Voleti et al., 2024). By default, this model generates 18 views that are uniformly sampled from a full orbit with zero elevation along the azimuth. To obtain 3D representations from the collection of views, we train Gaussian splats (Ren et al., 2023) consisting of 100K splats that are optimized for 4K steps.

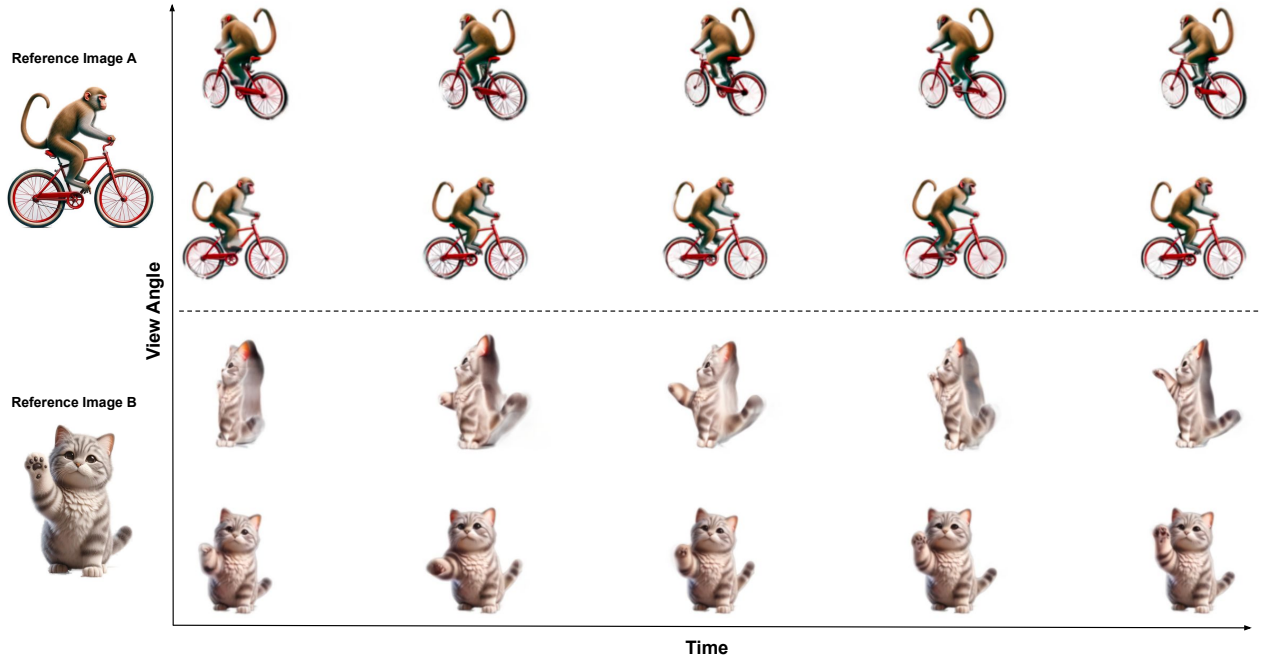


*Figure 2.* An example of multiple 2D renderings of 3D videos generated by our method. The 3D videos are rendered from two different camera views (y-axis) through time (x-axis). We observe consistency between different camera views for the same timestep as well as plausible dynamics across time for the same view.

### 2.2. Evaluation Setup

To evaluate Vid3D we use the benchmark proposed by Zhao et al. (2023), which is composed of 24 high-quality reference images. This benchmark has been adopted by recent 3D video papers as the standard evaluation framework (Zhao et al., 2023; Ren et al., 2023). Following their evaluation procedure, to measure the quality of each 3D video, we render 10 2D videos from ten different uniformly sampled camera angles. As a quantitative evaluation metric, we report the CLIP-I score, which is defined as the average cosine similarity between the CLIP-features (Radford et al., 2021) of the reference image and each frame in each 2D video rendering. Specifically, we use the CLIP-ViT-Base with a patch size of 32 as the image encoder. As CLIP features represent an embedding of an image in a semantically meaningful latent space, this metric captures the average semantic similarity (and hence model fidelity) across both time and camera angle. We compare Vid3D to both Animate124 and DreamGaussian4D which represent the state-of-the-art for 3D video generation.

## 3. Results and Discussion

### 3.1. Main Results

We provide 2D video renderings from two 3D videos generated by Vid3D in Figure 2. For each 3D video, we render 2D videos from two distinct camera views. For both reference images, the renderings demonstrate subject consistency with the original frame as well as interesting and plausible dynamic details such as the cat raising and lowering its paws or the monkey pedaling the bike. In addition to temporal consistency across time for the same view, there is spatial consistency for different views from the same timestep; the subject and the content agree across multiple views. While the capacity of Vid3D for general 3D scene rendering is impressive, there are still weaknesses to the model. As can be seen in the cat rendering, rendering quality degrades moving from views aligned with the reference image to alternative camera views. We believe this decrease in quality is not due to multi-view information being dropped between timesteps, suggesting that the quality could be improved with higher quality multi-view synthesis models.

*Table 1.* CLIP-I score for Vid3D compared to Animate124 and DreamGaussian4D, showing that our model does not need 3D temporal dynamics to yield competitive results.

| Model | CLIP-I |
|---|---|
| Animate124 | 0.8544 |
| DreamGaussian4D | 0.9227 |
| Vid3D (Ours) | 0.8946 |

We also quantitatively compare our proposed Vid3D method to both Animate124 (Zhao et al., 2023) and DreamGaussian4D (Ren et al., 2023) in Table 1. While Vid3D achieves

a higher CLIP-I score than Animate124, it underperforms the CLIP-I score of DreamGaussian4D. Although Vid3D does not achieve a new state-of-the-art score, that Vid3D's quantitative performance is competitive with state-of-the-art 3D video baselines implies that 3D temporal dynamics may not be fully necessary for generating dynamic 3D scenes. Instead, it is possible to rely solely on 2D video model priors to ensure multi-view consistency through time, which with further tuning, could achieve even greater performance.

We present further qualitative results and an analysis of failure modes in Appendix D.

### 3.2. Varying Number of Views

*Table 2.* CLIP-I score values for Vid3D for different numbers of views. This result shows that reducing the number of views from 18 to 9 does not significantly degrade performance, while further reduction does.

| Number of views | CLIP-I |
|---|---|
| 3 | 0.8532 |
| 9 | 0.8879 |
| 18 (Baseline) | 0.8946 |

We ablate how the number of views produced per timestep during the multi-view synthesis process affects the quality of the resulting 3D generations. The number of views generated determines the information content available when training Gaussian splats and more views provides denser coverage of the scene. While for capturing views of real objects, i.e., object views captured by sensors, more views corresponds to higher quality 3D representations, our work synthesizes views from a 2D video model and as such we investigate the effect of view number on quality. We report the CLIP-I score for varying view numbers in Table 2. We find that the CLIP-I score strictly decreases as the number of views decreases. Furthermore, the decrease in CLIP-I score grows larger at lower view numbers; there is only a drop of 0.0067 from 18 to 9 frames, but a drop of 0.0347 from 9 to 3 frames. As compared to the baselines, even at 3 frames Vid3D is comparable to Animate124, only performing worse by 0.0012.

We provide a qualitative example of single timestep renderings from Gaussian splats trained from varying number of views in Figure 3. There is no perceivable degradation in quality when comparing 18 views to 9 views, but there is a significant loss of quality for the model trained with 3 views. Specifically, at 3 views all high frequency detail is lost and the boundary of the object is poorly defined.

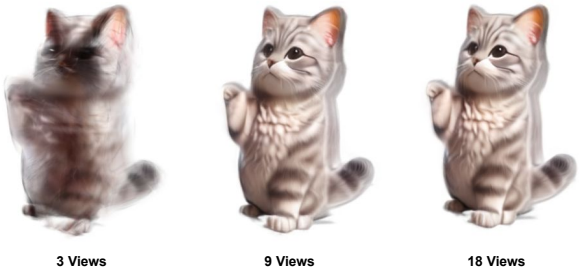


*Figure 3.* Rendering of singular frame from 3D videos generated using the same reference image but trained with a varying number of synthesized views. There are no perceivable differences between 18 and 9 views, but there is significant degradation and noise using 3 views.

## 4. Conclusion

In this work, we systematically analyze an alternative to recent dynamic 3D scene generation algorithms and challenge the assumption that 3D video generation methods must explicitly model 3D temporal dynamic. We propose Vid3D, a simple method for 3D video generation that generates 3D representations for each timestep in the video independently of all other timesteps. We find that this simpler methodology still holds up to the state-of-the-art baselines of Animate124 and DreamGaussian4D. Vid3D achieves a CLIP-I score of 0.8946 as compared to baseline scores of 0.8544 and 0.9227 for Animate124 and DreamGaussian4D respectively. We ablate the effect that the number of views generated for each timestep has on 3D scene quality, finding that scene quality only significantly degrades for a very low number of views. Overall, Vid3D is a simple alternative to current 3D video generation methods and it achieves results consistent with the state-of-the-art even though it does not explicitly model 3D temporal dynamics.

## Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. With easier generation of dynamic 3D scenes, embodied intelligence models can be trained more easily. This could result in societal benefits such as better trained robotics and prosthetics. However, we also elect to also point out the potential effects of this research on the creation of malicious videos like deepfakes. We encourage further work along preventing this malicious behavior for greater alignment of future models.

## References

Armandpour, M., Sadeghian, A., Zheng, H., Sadeghian, A., and Zhou, M. Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. *CoRR*, abs/2304.04968, 2023. doi: 10.48550/ARXIV.2304.04968. URL https://doi.org/10.48550/arXiv.2304.04968.

Bahmani, S., Skorokhodov, I., Rong, V., Wetzstein, G., Guibas, L. J., Wonka, P., Tulyakov, S., Park, J. J., Tagliasacchi, A., and Lindell, D. B. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. *CoRR*, abs/2311.17984, 2023. doi: 10.48550/ARXIV.2311.17984. URL https://doi.org/10.48550/arXiv.2311.17984.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., and Rombach, R. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023.

Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., and Ramesh, A. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation/-models-as-world-simulators.

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., and Yu, F. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

Chatzitofis, A., Saroglou, L., Boutis, P., Drakoulis, P., Zioulis, N., Subramanyam, S., Kevelham, B., Charbonnier, C., Cesar, P., Zarpalas, D., et al. Human4d: A human-centric multimodal dataset for motions and immersive media. *IEEE Access*, 8:176241–176262, 2020.

Chen, Z., Wang, Y., Wang, F., Wang, Z., and Liu, H. V3d: Video diffusion models are effective 3d generators. *arXiv preprint arXiv:2403.06738*, 2024.

Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., and Farhadi, A. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13142–13153, 2023.

Du, Y., Watkins, O., Darrell, T., Abbeel, P., and Pathak, D. Auto-tuned sim-to-real transfer. In *IEEE International Conference on Robotics and Automation, ICRA 2021, Xi'an, China, May 30 - June 5, 2021*, pp. 1290–1296. IEEE, 2021. doi: 10.1109/ICRA48506.2021.9562091. URL https://doi.org/10.1109/ICRA48506.2021.9562091.

Han, J., Kokkinos, F., and Torr, P. Vfusion3d: Learning scalable 3d generative models from video diffusion models, 2024.

Kerbl, B., Kopanas, G., Leimkühler, T., and Drettakis, G. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023a.

Kerbl, B., Kopanas, G., Leimkühler, T., and Drettakis, G. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023b. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Liu, R., Wu, R., Hoorick, B. V., Tokmakov, P., Zakharov, S., and Vondrick, C. Zero-1-to-3: Zero-shot one image to 3d object. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 9264–9275. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00853. URL https://doi.org/10.1109/ICCV51070.2023.00853.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Pagés, R., Amplianitis, K., Ondrej, J., Zerman, E., and Smolic, A. Volograms & v-sense volumetric video dataset. *ISO/IEC JTC1/SC29/WG07 MPEG2021/m56767*, 2021.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Reimat, I., Alexiou, E., Jansen, J., Viola, I., Subramanyam, S., and Cesar, P. Cwipc-sxr: Point cloud dynamic human dataset for social xr. In *Proceedings of the 12th ACM Multimedia Systems Conference*, pp. 300–306, 2021.

Ren, J., Pan, L., Tang, J., Zhang, C., Cao, A., Zeng, G., and Liu, Z. Dreamgaussian4d: Generative 4d gaussian splatting. *CoRR*, abs/2312.17142, 2023. doi: 10.48550/ARXIV.2312.17142. URL https://doi.org/10.48550/arXiv.2312.17142.

Voleti, V., Yao, C., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., Laforte, C., Rombach, R., and Jampani, V. SV3D: novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *CoRR*, abs/2403.12008, 2024. doi: 10.48550/ARXIV.2403.12008. URL https://doi.org/10.48550/arXiv.2403.12008.

Yoon, J. S., Yu, Z., Park, J., and Park, H. S. Humbi: A large multiview dataset of human body expressions and benchmark challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):623–640, 2021.

Zhang, T., Zhang, K., Lin, J., Louie, W.-Y. G., and Huang, H. Sim2real learning of obstacle avoidance for robotic manipulators in uncertain environments. *IEEE Robotics and Automation Letters*, 7(1):65–72, 2022. doi: 10.1109/LRA.2021.3116700.

Zhao, Y., Yan, Z., Xie, E., Hong, L., Li, Z., and Lee, G. H. Animate124: Animating one image to 4d dynamic scene. *arXiv preprint arXiv:2311.14603*, 2023.

## A. Implementation

We open-source our implementation here, which provides training scripts for running Vid3D on an 8xA100 node.

## B. Related Work

Efficiently generating 3D scenes from 2D inputs has been a long-standing goal in computer vision (Chang et al., 2015). To automate and thus scale the building of static 3D scenes, researchers have recently leveraged advances in generative 2D video models to generate 3D scenes by querying a video model for multiple poses of said scene (Chen et al., 2024; Han et al., 2024). However, these existing methods only generate static 3D data, meaning that there is no temporal component.

**Domain-Specific Datasets** Some existing highly domain-specific 3D video datasets (i.e. human bodies, human faces, etc.) have also been manually curated using a large number of sensors (Chatzitofis et al., 2020; Reimat et al., 2021; Yoon et al., 2021; Pagés et al., 2021). However, the use of these dynamic 3D scenes is limited due to their domain-specific nature. Additionally, it is hard to scale their hardware-intensive video capture technology to produce more scenes.

**Sim2Real** Outside of 3D and 4D video datasets, other works focus on developing sim2real worlds, where robots learn real-world features in simulation (Zhang et al., 2022; Du et al., 2021). These works develop worlds where robots can interact in 3D space over time, but these 4D worlds require physics and temporal correlations to be designed by hand, hence not being able to generate dynamic scenes directly from a sample image.

**Dynamic 3D Scenes** Finally, a number of papers have focused on synthesizing dynamic 3D scenes either by using classifier guidance (Bahmani et al., 2023; Zhao et al., 2023) or by training models that learn how to deform 3D objects (Ren et al., 2023). As stated before, these methods either suffer from high computational cost taking hours to render a single scene, or from rigid assumptions about the temporal structure of 3D scenes, limiting these methods' applicability for modeling arbitrary dynamic 3D scenes. Our work instead leverages the strong generalization capabilities of 2D video models to flexibly generate 3D scenes without the need to train complex auxiliary models.

## C. Varying Scene Motion

*Table 3.* CLIP-I score values for Vid3D for different temporal seed motion scores. This result shows that there is a slight loss in quality for scenes with more motion.

| Motion Score | CLIP-I |
|---|---|
| 120 (Baseline) | 0.8946 |
| 160 | 0.8893 |
| 200 | 0.8897 |

We ablate how the degree of motion in the scene's temporal seed (i.e., 2D video outline) affects the quality of the resulting 3D generations. The model we use for temporal seeding, Stable Video Diffusion, generates videos conditioned on a "motion score", which corresponds to the time and spatial average of the optical flow maps of the video. We thus vary the degree of motion in temporal seeds by varying the motion score.

We quantitatively measure the impact of varied motion in terms of CLIP-I score in Table 3. While the initial increase of conditioned motion score from 120 to 160 degrades the CLIP-I score of the renderings by 0.0053, there is almost no difference between the CLIP-I metrics for motion scores of 160 and 200. This result suggests that while there is some initial loss in quality due to increased motion, past a given motion score the quality of Vid3D is relatively robust to added motion.

We provide a qualitative example of varying the motion score in Figure 4. For each motion score, we render a 2D video from the 3D video using a fixed camera view. We observe greater variability in poses between frames for the higher motion score. Furthermore, the added motion is achieved without a noticeable drop in rendering quality.

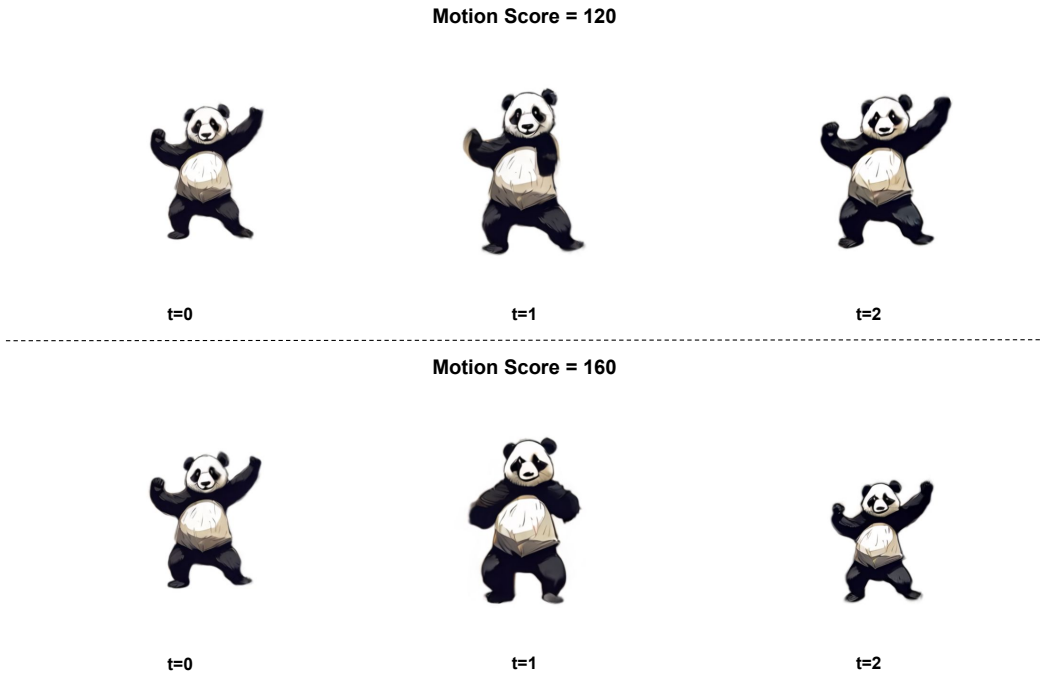


*Figure 4.* Rendering of singular frame from 3D videos generated by the same reference image but with different amount of motion synthesized. As desired, the higher motion score video has higher variability, along with similar rendering quality to the lower motion score, demonstrating robustness to motion.

## D. Qualitative Analysis

We compare the performance qualitatively of Vid3D to DreamGaussian4D and Animate124 on four examples (Ren et al., 2023; Zhao et al., 2023) in Figure 5, Figure 6, Figure 7, and Figure 8. We note that only four examples are chosen because Animate124 requires 10 hours of textual inversion and 7 hours of classifier guidance in order to produce a singular video, in comparison to the ten minutes from our method.

We observe that Vid3D has high quality generated 3D videos on multiple axes. First, the video hallucinates less than Animate124, which frequently changes the color and appearance of objects, such as in the space shuttle and the astronaut, along with higher resolution. Second, DreamGaussian4D has high quality reference-view renders but struggles on alternate views, such as those for the space shuttle. On the other hand, Vid3D achieves high performance over all axes.

In Figure 9, we present further examples of generations from Vid3D, which demonstrate strong coherence both temporally and over multiple views. This demonstrates that Vid3D can learn strong 3D temporal priors even though it solely relies on 2D temporal priors and 2D video diffusion models.

### D.1. Failure Modes

We address various failure modes of Vid3D in Figure 10. The first is a lack of quality in the generation of multi-views, which leads to issues like the face of the skeleton not being properly rendered. We plan to hopefully address this by potentially training our own multi-view generation model. The next issue is hallucination, where new objects, like the orange ball are added to existing scenes. This issue arises because of the flexibility of our algorithm–we do not simply deform the scene. However, this means that occassionally, small aberrations like the orange ball may appear. Finally, we encounter the Janus problem common in 3D generation (Armandpour et al., 2023). For example, in the third frame of the bike, the bike appears twice, which is an issue that can only be addressed through better training of multi-view models.

These issues are all common issues faced by 3D generation models, and do not take away from the value of this experiment. We thus still maintain that 3D temporal knowledge may not be necessary for 3D video generation.

*Figure 5.* A qualitative comparison of Animate124, DreamGaussian4D, and Vid3D for a seed image of an astronaut riding a horse. Here, Vid3D both creates accurate representations from multiple angles, but also does not recolor the horse like Animate124 or have worse renders from a non-reference view like DreamGaussian4D.

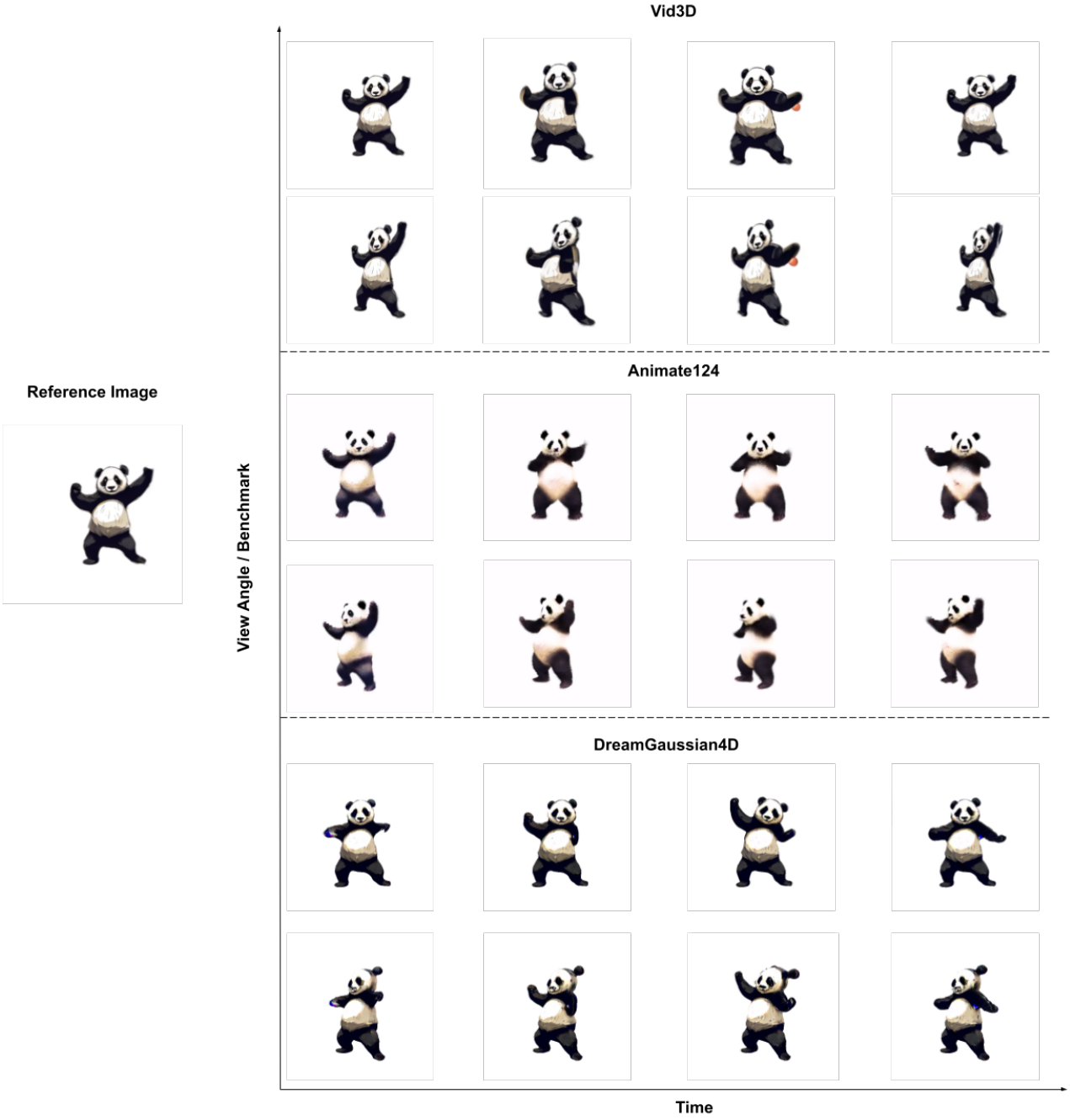


*Figure 6.* A qualitative comparison of Animate124, DreamGaussian4D, and Vid3D for a seed image of a dancing panda. Here, although Vid3D includes the hallucination of a orange ball (discussed in Figure 10), in comparison to DreamGaussian4D, Vid3D does not hallucinate only one ear to the panda.

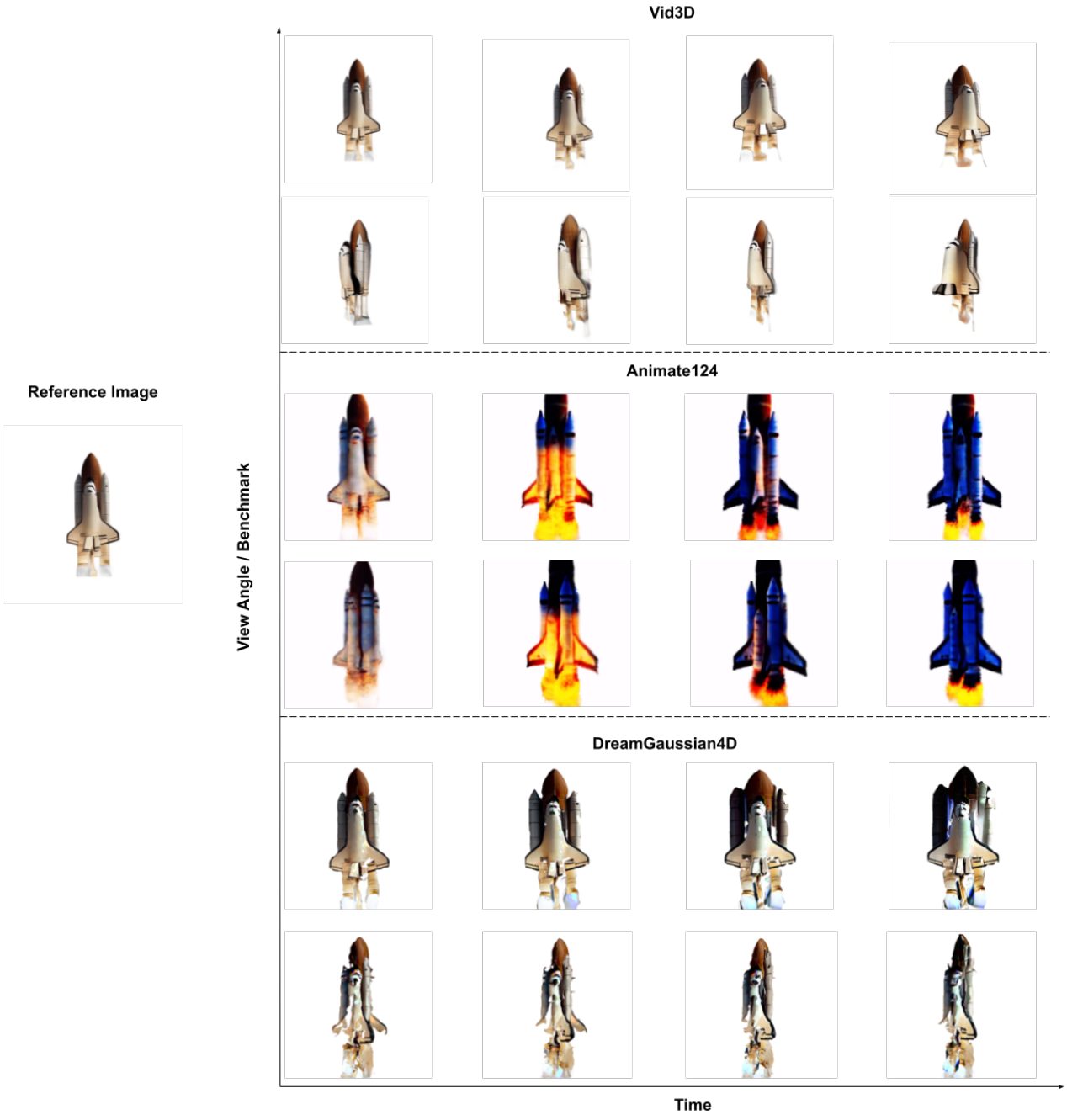


*Figure 7.* A qualitative comparison of Animate124, DreamGaussian4D, and Vid3D for a seed image of a space shuttle. Here, the quality of the exhaust is higher in Vid3D, with the exhaust recoloring the rocket in Animate124 and the space shuttle being deformed in the side angle in DreamGaussian4D.

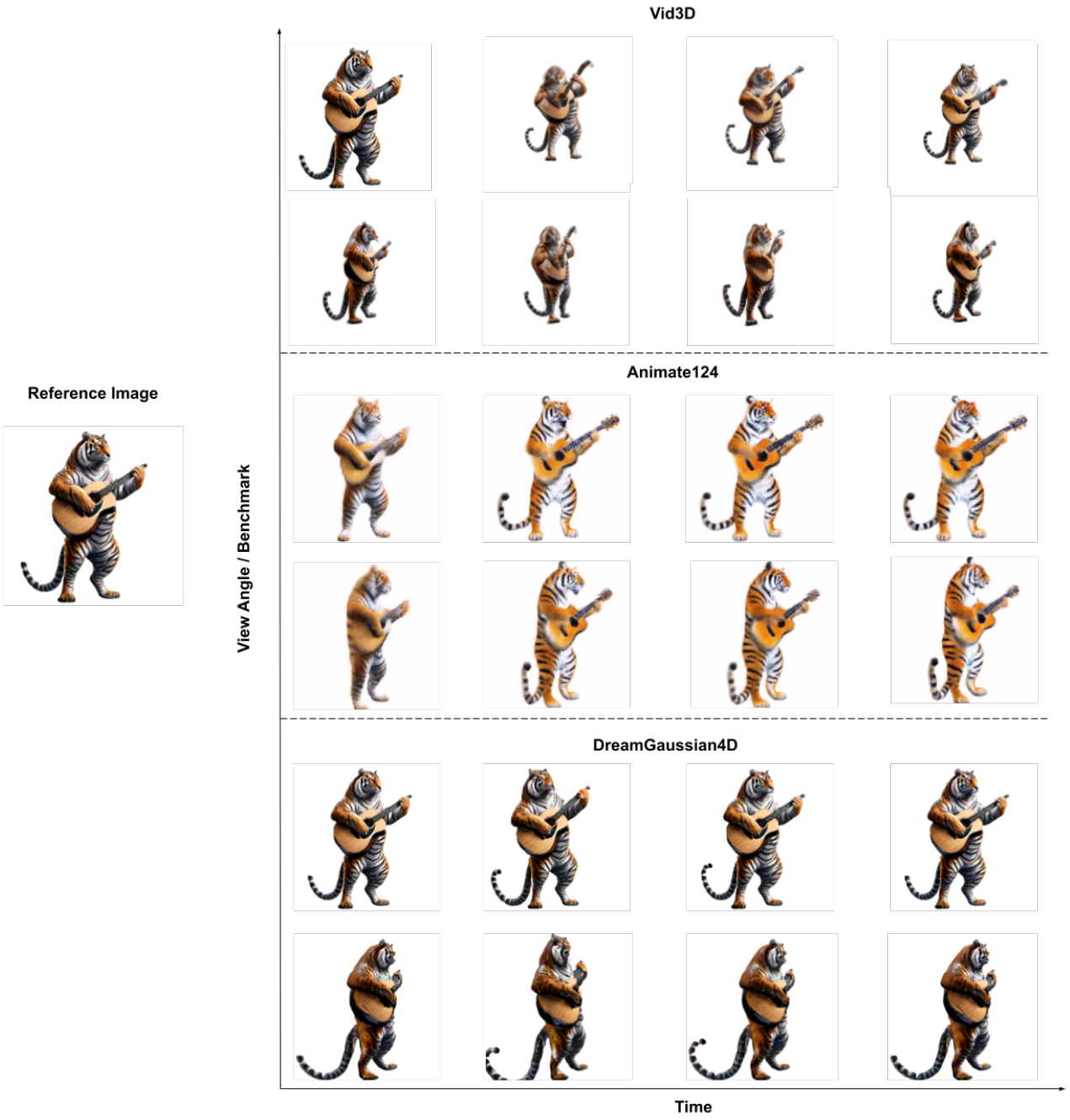


*Figure 8.* A qualitative comparison of Animate124, DreamGaussian4D, and Vid3D a seed image of a tiger playing the guitar. Here, we notice that our method has higher coherence to the source image than Animate124, keeping the same style and not distorting the shape of the tiger like DreamGaussian4D.

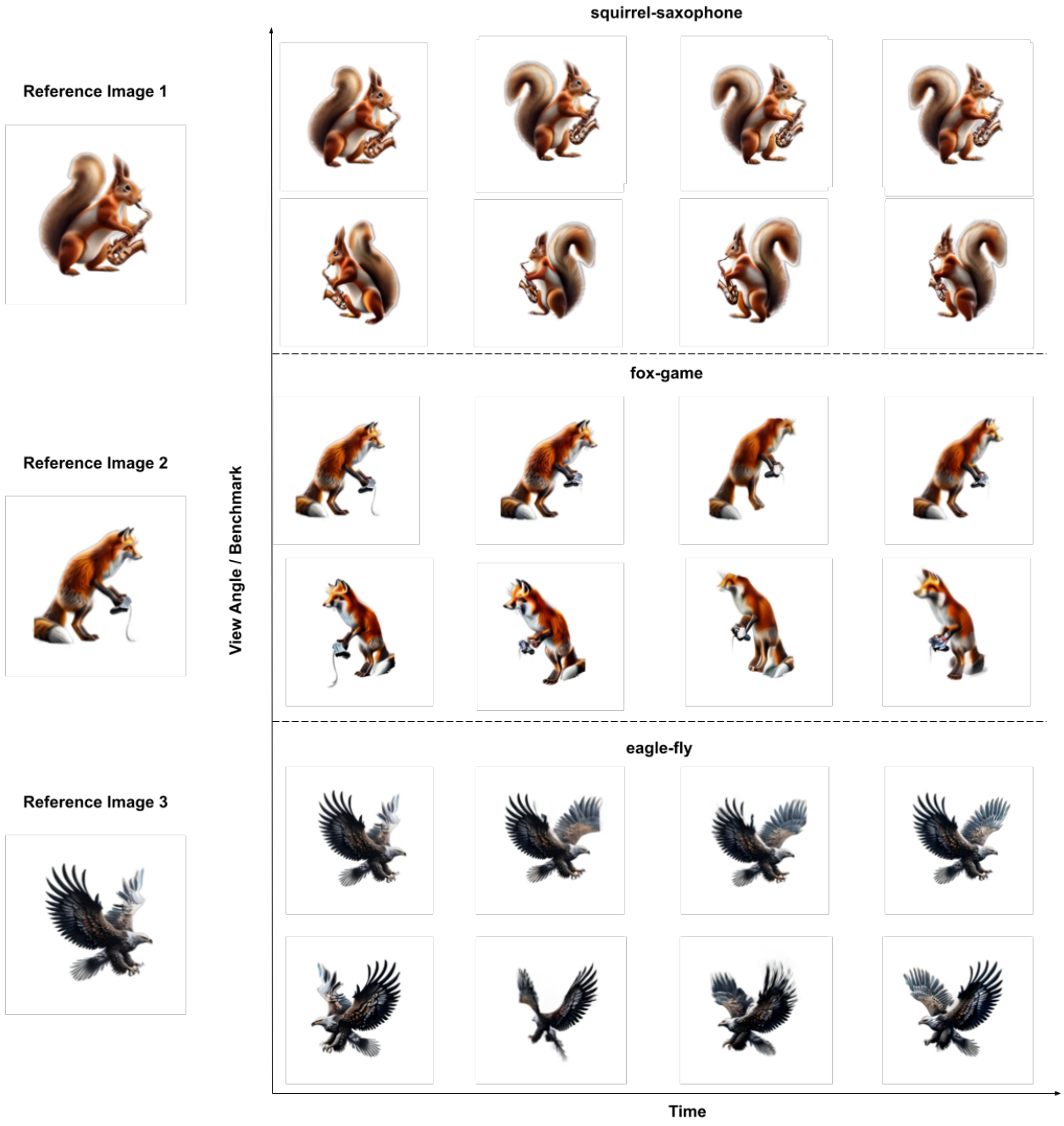


*Figure 9.* Further examples of high quality generations from Vid3D, demonstrating generalization over multiple timesteps and views.

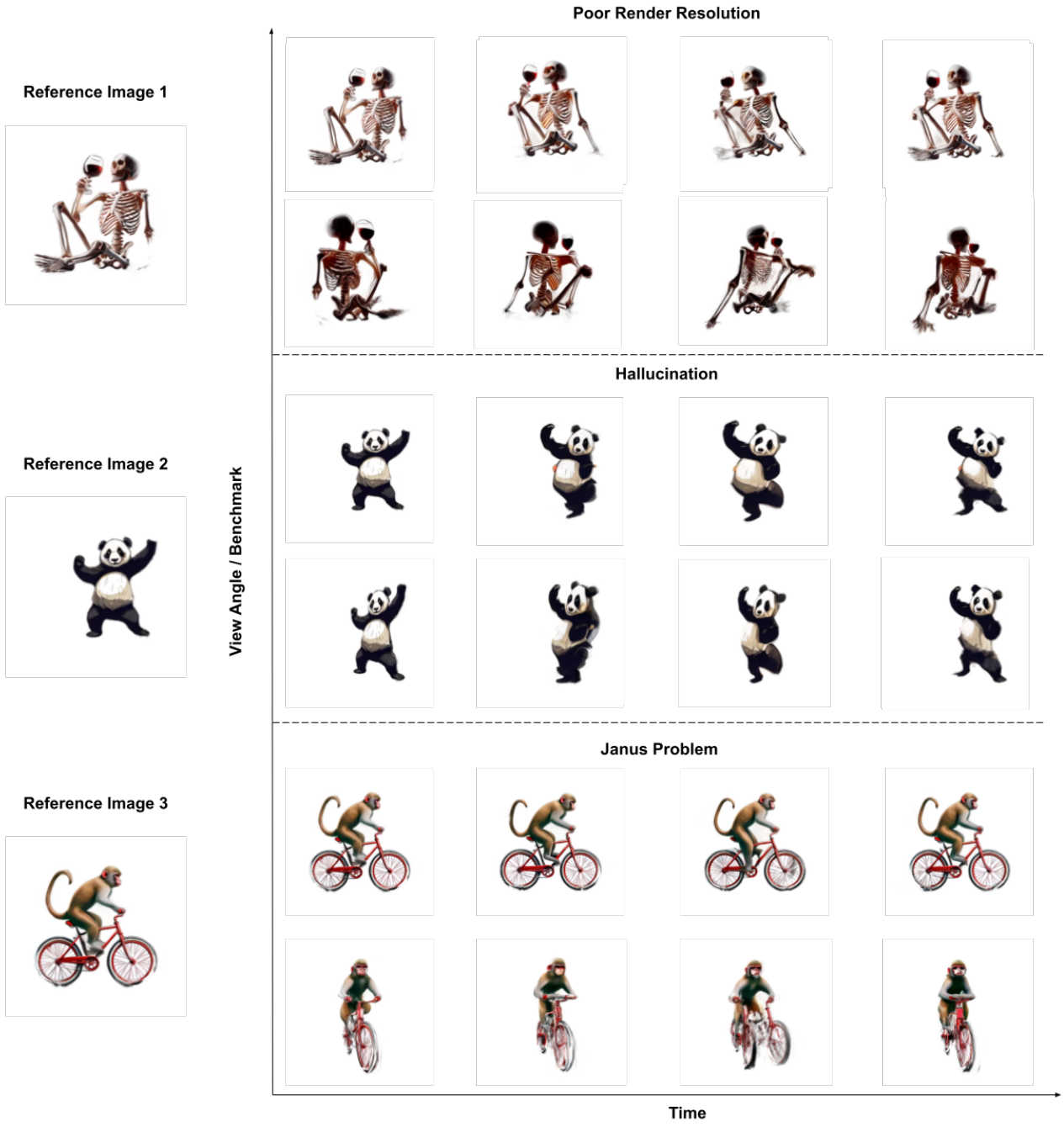


*Figure 10.* A depiction of various failure modes for Vid3D. The first is poor render quality, where the skeleton has its facial features blurred due to the failure of the 3D modeling to appropriately represent it. The second is hallucination, where a orange ball appears near the panda. The final is the Janus problem, where the same object appears twice, which is a known problem in 3D generation and once again appears (Armandpour et al., 2023).