I2V3D: Controllable image-to-video generation with 3D guidance

Supplementary Material

A. User Study

In this section, we further conduct a user study to evaluate the effectiveness of our approach. We compare our method with the baselines mentioned for both human-like characters videos (AnimateAnyone [20], MagicPose [7], ISculpting [69]) and nonhuman-like object videos (DragAnything [61], MotionBooth [60], ISculpting [69]).

The user study includes 30 animations. For each animation, participants were shown the original input image and rendered videos, followed by animations generated by the baseline methods and our approach, arranged in random order. Each animation was repeated three times to ensure participants had ample time to assess and compare the results. Participants were then asked to answer three questions by selecting the best option:

- **Alignment:** Select the video that best preserves the motion of the objects and camera in the rendered video and the appearance of the input image.
- **Consistency:** Select the video with the least flicker.
- **Overall quality:** Select the video with the fewest artifacts.

We collected a total of 24 questionnaires from participants aged 20 to 55, including 7 with CG and CV backgrounds and 17 from other fields. The results of the user study, as shown in Fig. 10, indicate that our method outperforms the other methods in all three factors. It achieved preferred rates of 0.811 and 0.814 for alignment, 0.856 and 0.814 for consistency, and 0.844 and 0.828 for video quality.

B. Ablation Study

Single Stage vs. Two Stages. We conducted an ablation study using a single-stage approach, where the input image served as the first frame, and geometric guidance was applied as described in Sec.3.3. As shown in third row of Fig.8, the sea turtle loses texture because the image-to-video model struggles when the first and subsequent frames differ substantially. Moreover, the water color shifts over time due to accumulated errors, whereas our two-stage approach preserves the input image's visual details throughout the entire video sequence.

Reconstructed 3D Background Meshes. As discussed in Sec.3.1, we reconstruct the mesh for the background as well, which offers two key advantages. First, reconstructed meshes enable meaningful interactions with the background. For instance, as shown in the first panel of Fig.9, our method allows the goose to walk behind the fire hydrant, even when it is partially occluded. In contrast, using the image plane as the background restricts the goose to the frontal plane, causing it to block the fire hydrant rather than moving behind it. Second, the reconstructed background meshes provide 3D guidance after camera movement. By comparison, relying on the image plane results in an invisible black region when the camera moves, as shown in the second panel of Fig. 9, where the camera moves backward. While the diffusion model can fill in the black region using its generalization capability, the generated content is often unrealistic, such as additional goal frames or soccer balls. In contrast, our reconstructed meshes guide the expanded scene, ensuring that the result is more realistic. Third, background reconstruction facilitates large camera movement in video generation, as shown in Fig. 12.

3D-Guided Video Generation vs. Video Refinement We also conduct an ablation study using video refinement to evaluate the effectiveness of our 3D-guided video generation method, as shown in Fig. 11. Following the Renoise strategy [12, 37], we add random noise to the video up to a certain noise level and then denoise it to obtain refined images. Specifically, we use SVD [2], adopting the input image as the first frame and setting t to 0.4. However, this approach fails to correct the background region, and due to the discontinuity between the first frame and subsequent frames, it results in noticeable blurriness. Additionally, we apply TokenFlow [15], a video-to-video translation method based on a text-to-image diffusion model. For a fair comparison, we use our fine-tuned SDXL LoRA. The results indicate that despite these refinements, noticeable flickering issues persist. In contrast, our method outperforms these approaches by effectively reducing artifacts while maintaining the smoothness of the video.

C. Limitations

Despite the effectiveness of our method, it has three main limitations. First, it cannot fully resolve artifacts caused by inaccurate 3D reconstruction, especially when geometric errors are significant. For example, the human hands in Fig.13, first row, exhibit distortions that stem from inaccurately reconstructed geometry. Second, the detailed appearance of objects in the generated video does not always perfectly align with the input image. The discrepancy, such as the variation from paper to leaf in Fig. 13, second row, arises from the limited capability of customizing a diffusion model with LoRA to capture fine-grained details. Lastly, the generated video is not entirely smooth, as some small

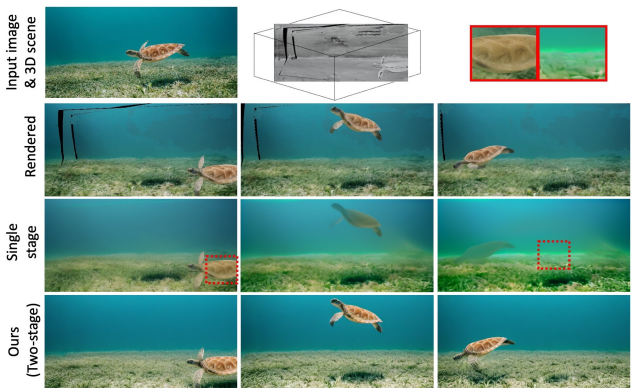

Figure 8. Ablation on single-stage and two-stage video generation. The red boxes highlight texture loss and error accumulation.

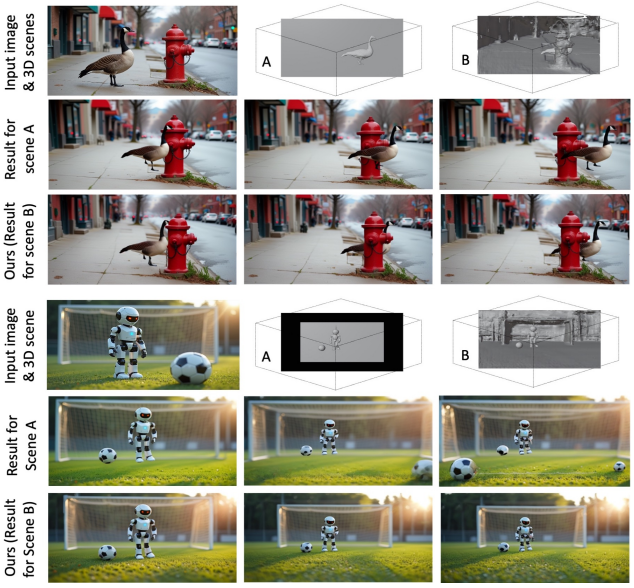

Figure 9. Ablation on 3D reconstruction for background. Scene A and B are w/o and w/ background reconstruction respectively.

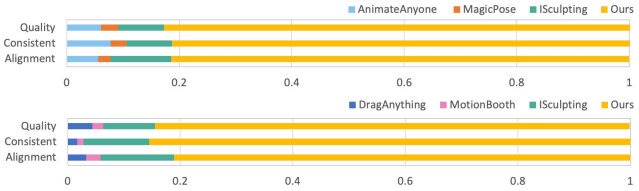

Figure 10. Results of user study. Our method has the best preferred rate for all video alignment, consistency and quality on both human-like characters and nonhuman-like objects..

flickering persists. This reflects the limitations of current open-source video diffusion models. Fortunately, since our I2V3D framework is general and not tied to specific reconstruction and generation models, we believe these problems can be addressed with advancements in 3D modeling techniques and improved image [11, 27] and video generation models [68].

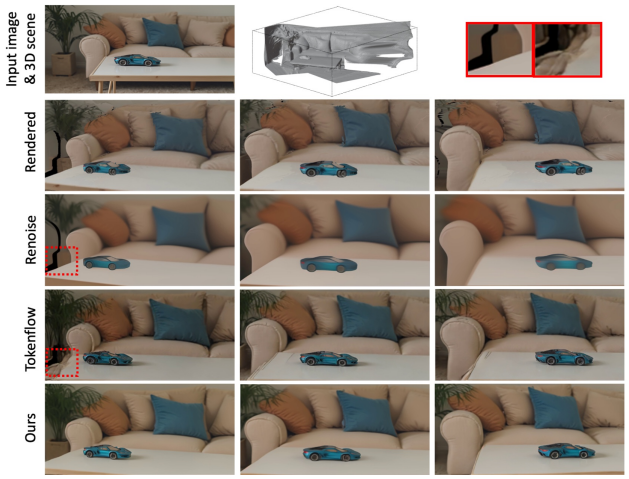

Figure 11. Ablation on 3D-guided video generation vs. video refinement. The red boxes highlight failures in refinement.

Figure 12. Generate video with large camera movement.

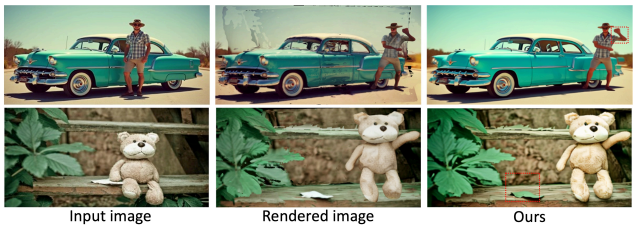

Figure 13. Our limitations. 1st row: Artifacts caused by the coarseness of the reconstructed mesh. 2nd row: insufficient customization provided by LoRA.

D. Comparison with DiT-based Methods

In addition to the comparisons presented in Section 4.2, we further evaluate our method against two state-of-the-

art DiT-based approaches for controllable video generation: Diffusion-as-Shader [16] and Go-with-the-Flow [4]. Both methods show limitations in fine-grained control, especially for object rotation. As illustrated in Figure 14, Go-with-the-Flow [4] produces distorted outputs with noticeable artifacts, while Diffusion-as-Shader [16] fails to generate rotations that are consistent with the rendered geometry. This may be due to misalignment between the first frame and subsequent frames, or a lack of sufficient rotation examples in their training data.

In contrast, our method leverages inversion and feature injection to achieve more precise and geometry-consistent control. Moreover, our keyframe generation allows flexible starting points without the constraints of I2V pipelines.

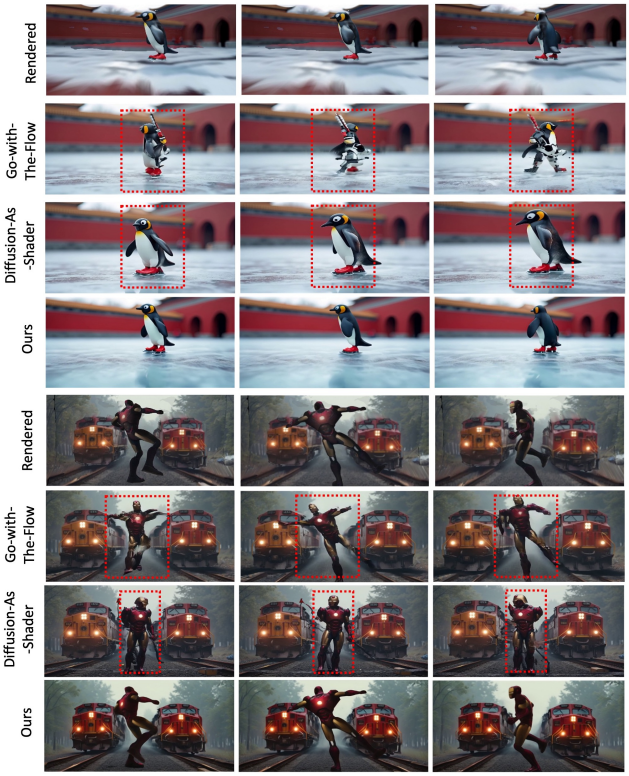

Figure 14. Comparison with DiT-based methods, Diffusion-as-Shader [16] and Go-with-the-Flow [4]. Both methods struggle with fine-grained control such as object rotation.

E. Quality Upgrades with DiT-based Models

Our framework is model-agnostic and can benefit from better model. For example, the keyframe generation stage can replace SDXL [40] with Flux [28], and the interpolation stage can swap SVD [2] for CogVideoX [68] or WAN 2.1 [53] for higher fidelity. As shown in Figure 15, benefiting from the improved detail preservation capability of Flux [28], the white paper on the desk remains consistent with the input image instead of turning green, and the girl's face and skirt exhibit finer details. Moreover, thanks to the strong motion modeling ability of WAN [53], the clouds in the sky appear more dynamic.

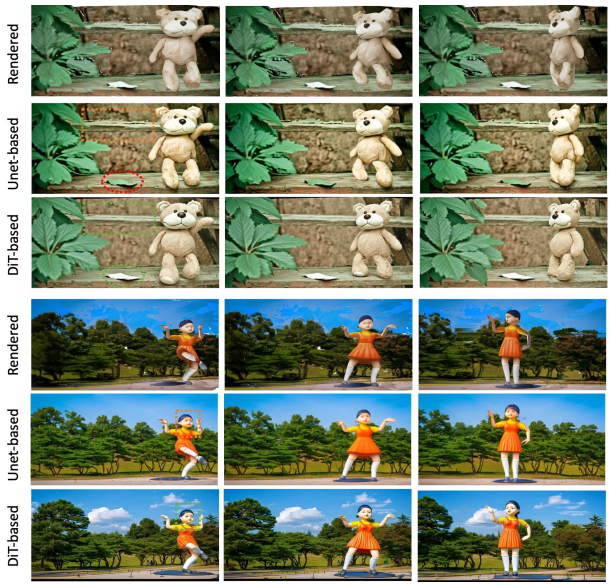

Figure 15. Enhancing quality with DiT-based models. Our framework is training-free and adaptable to the latest models. Results demonstrate improved details and more dynamic motion.

F. Reconstruction Error Handling

While the 3D reconstruction methods we adopt are generally robust, occasional artifacts such as missing topology may still occur. Our system is designed to tolerate such imperfections. When the geometry is unreliable, users can adaptively reduce the strength of geometric control by lowering the ControlNet scale and decreasing the number of feature injection steps, allowing the diffusion prior to compensate for the degraded guidance. As shown in Figure 16, even in the presence of noticeable topology distortion on the astronaut's face and missing mesh in the background, our method produces plausible results by reducing the 3D guidance strength by half.

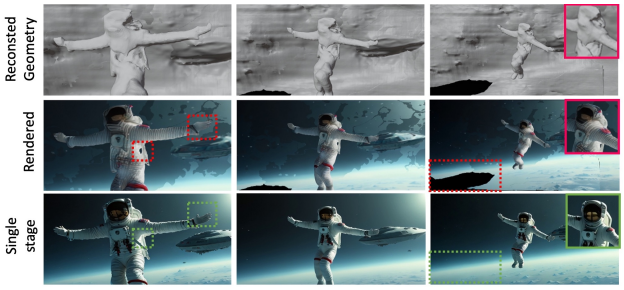

Figure 16. Experiments on error handling by reduced guidance. 1st row: Erroneously reconstructed mesh. 2nd row: Rendered result with significant artifacts. 3rd row: Our generated result. (errors corrected)