ViVid-1-to-3: Novel View Synthesis with Video Diffusion Models

Supplementary Material

This supplementary document includes additional content not covered in the main paper. We begin with the implementation details of our framework and then discuss the effectiveness of the design choices. Finally, we provide additional visual samples.

6. Implementation

For novel-view diffusion and video diffusion, we utilize the pre-trained Zero-1-to-3 XL [10, 34] and Zeroscope v2 576w [2], respectively. The resolution of the rendered frames presented in the paper is 256^2, but direct upscaling can be performed using Zeroscope v2 XL [1]. For the denoising scheduler, we employ the DPM solver [38] for both diffusion models instead of DDIM [61] and conduct 50 inference steps. For 360° videos, we set the number of frame F to 24, i.e., 24 frames video. The impact of the number of frames is discussed in Sec. 7. Since we nullify the text prompt conditioning for the video diffusion model, as explained in the main paper, the classifier-free guidance is not applied, and the guidance scale for novel-view diffusion is set to 3.0.

7. Discussion

Effect of number of frames. As our framework incorporates video diffusion, we have the flexibility to modify the number of video frame F in Eq. (4). As shown in Fig. 9, we vary the number of frames to check the 360° rendered images. When each frame is generated independently (equivalent to Zero-1-to-3-XL [10, 34]), the synthesized images lack consistency and exhibit significant deviations from the ground truth samples. Leveraging the video diffusion prior considerably enhances the pose and shape consistency of the generated images, with accuracy improving as more frames are utilized. Therefore, users can adjust the number of frames based on the trade-off between their computational resources and the desired level of consistency.

Prompting strategy. Although we set the prompt for our video diffusion process as a null, i.e., $y = \emptyset$ in Eq. (3), it does not mean that prompting is not possible. We found that in some cases, providing prompt can enhance the quality of our method. As shown in Fig. 10, the text conditioning `sunflowers in a vase` yields images of higher quality with better object-level details and hence improves the quality of novel-view synthesis. This example further highlights the potential use cases of our method for high-resolution and editable novel-view rendering.

8. Additional samples

Additional multi-view samples. In Fig. 11 and Fig. 12, we offer additional samples of multi-view synthesis. Here, we leverage the GSO dataset [12] to facilitate comparisons with ground truth samples. In Fig. 11, we compare our model to compare our model with 2D [10, 34, 35] and 3D methods [33, 44, 62] in the same way as Fig. 5 in the main paper. We additionally present multi-view samples of 2D-based methods [10, 34, 35] including ours in Fig. 12, to verify multi-view consistency and visual quality of the competitive models.

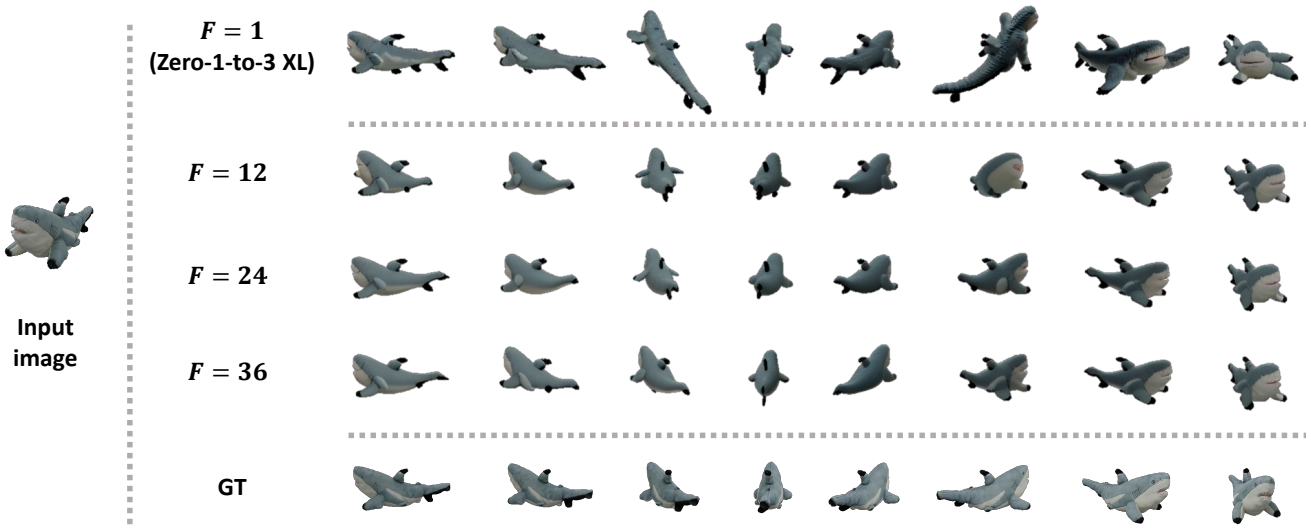

Figure 9. **Effect of number of frames –** We present an example of our framework by varying the number of frames.

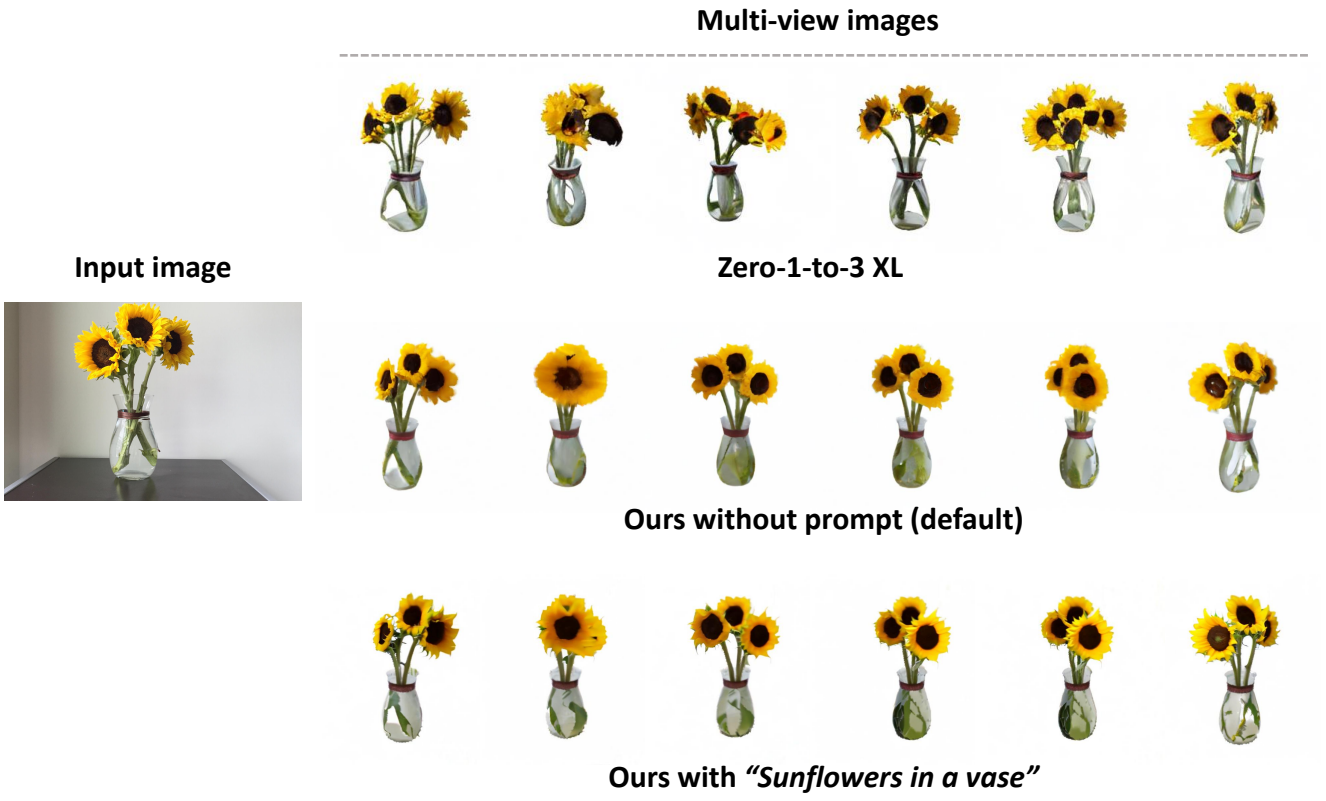

Figure 10. **Effect of prompting –** We show the effect of prompting on novel-view synthesis with our approach.

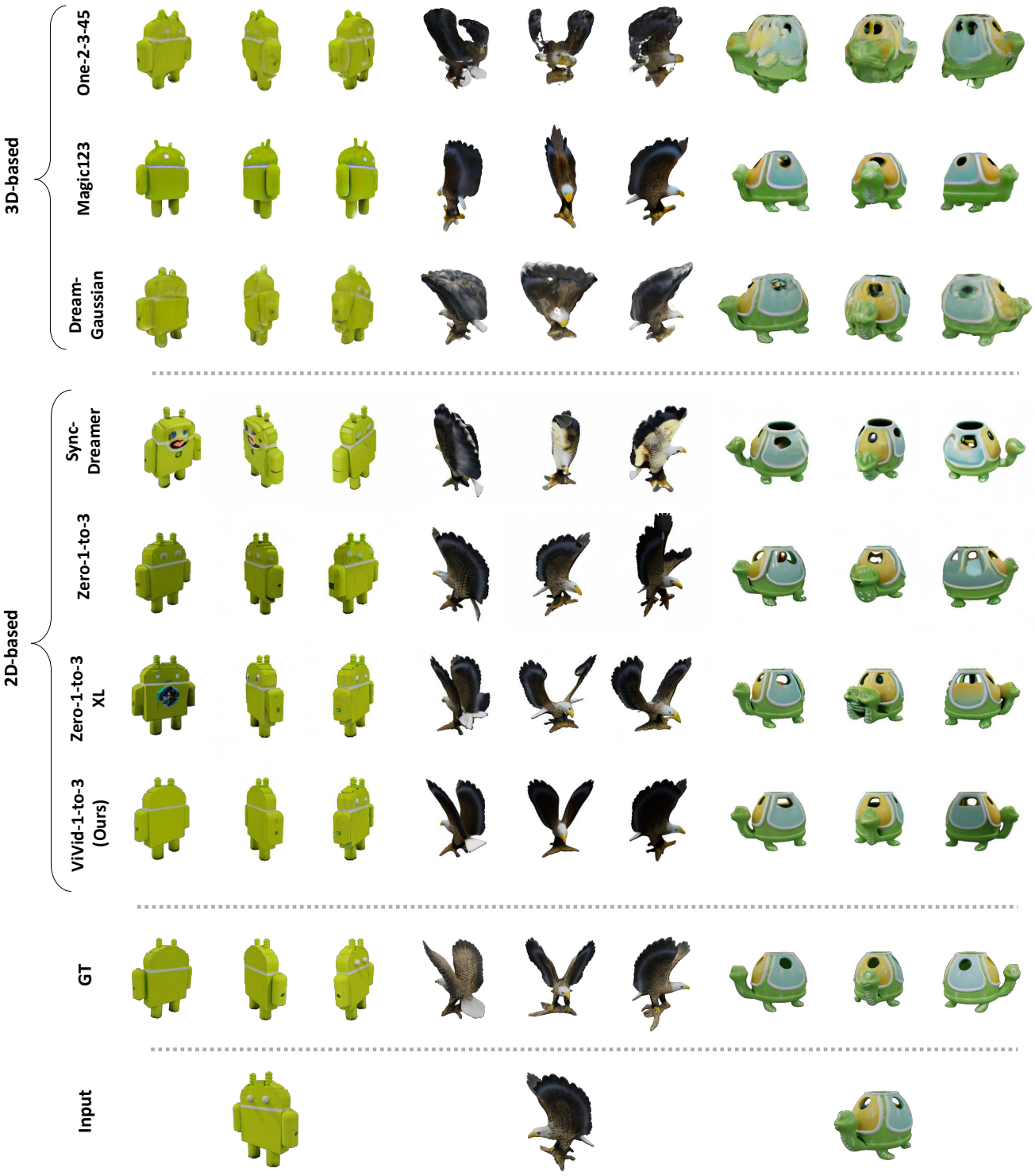

Figure 11. **Additional qualitative results –** Additional novel-view generation samples from GSO [12], and comparison with 2D and 3D methods.

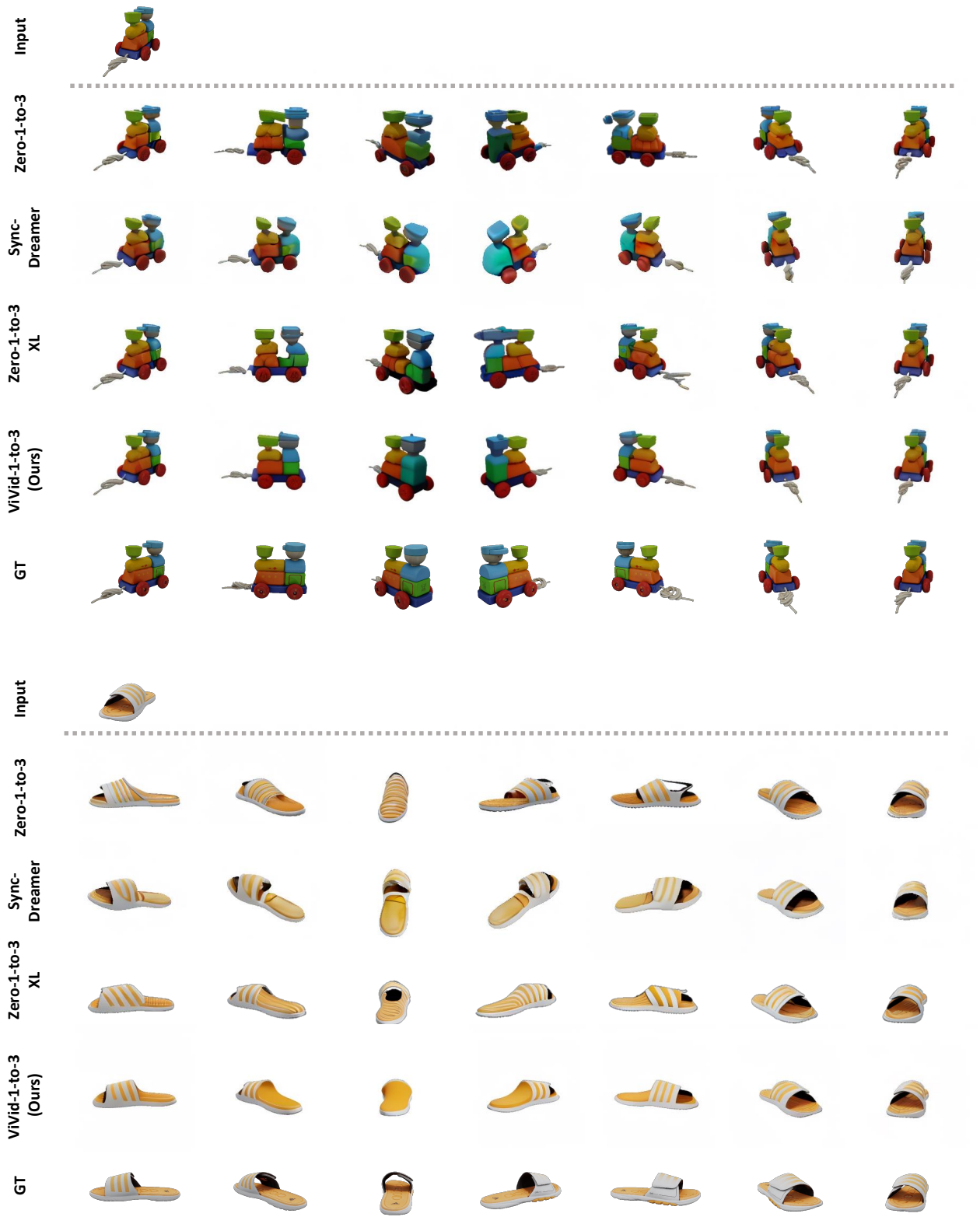

Figure 12. **Comparison to 2D methods –** Additional multi-view synthesis samples of 2D-based methods on GSO [12] dataset.