Supplementary Material of VFusion3D: Learning Scalable 3D Generative Models from Video Diffusion Models

Junlin Han[1,2]*, Filippos Kokkinos[1]*, and Philip Torr[2]

[1] GenAI, Meta
[2] Torr Vision Group, University of Oxford
* Equal contribution
junlinhan@meta.com, fkokkinos@meta.com, philip.torr@eng.ox.ac.uk

1 Training Details

Emu Video Fine-tuning. Following the EMU Video [3], we freeze the spatial convolutional and attention layers of Emu Video, while only fine-tuning the temporal layers. We use the standard diffusion loss for this fine-tuning process. The Emu Video model is fine-tuned over a period of 5 days using 80 A100 GPUs, with a total batch size of 240 and a learning rate of 1×10^{-5}. Although the 3D consistency continues to improve with extended fine-tuning, we do not observe any decline in visual quality. One possible explanation is the static nature of the spatial layers and the image-conditioned network, which ensures that the generated 360° videos maintain high fidelity with the high-frequency texture components of the input.

VFusion3D. The architecture of VFusion3D is identical to that of LRM [4], except for two minor modifications. The first is the use of DINOv2 [8] instead of DINO [1] to enhance image feature extraction capabilities. The second modification is that we have reduced the number of NeRF MLP layers to 4 instead of 10, as empirically suggested by the LRM ablation study.

In addition to the training details provided in the main paper, we use 0.95 as the second beta parameter of the AdamW optimizer [7]. We apply a gradient clipping of 1.0 and a weight decay of 0.05. This weight decay is only applied to weights that are neither bias nor part of normalization layers. We use Bfloat16 precision for training and inference.

Fine-tuning with 3D Data. We use 32 GPUs to fine-tune the pre-trained VFusion3D model with 3D data. At this stage, we also employ the L2 loss function for novel view supervision. The model undergoes fine-tuning with a dataset of 100K rendered multi-view images over 10 epochs, adhering to a cosine learning rate schedule. We set the initial learning rate as 1×10^{-4}. All other parameters remain consistent with the VFusion3D pre-training phase.

2 Visualizations and Test-time Processing

Video Comparison Results. We provide video comparison results in the project page that cover all the qualitative results presented in the main paper. Our project page also includes additional comparisons. These additional results are based on the Single Image 3D Reconstruction and Text-to-3D Generation experiments discussed in the main paper. All input images used were never seen by the model during training.

Test-time Processing. Following standard procedures [4, 10], we utilize a heuristic function to process all input images during testing. The initial steps involve eliminating the image background using rembg with isnet-general-use [9], then extracting the salient object. Following this, we adjust the size of the salient object to an appropriate scale and position it in the center of the input image.

3 Benchmarking on Public Datasets

We present a comparison using the GSO [2] dataset, from which we render 996 objects (excluding some redundant or very similar objects) with each object having 16 rendered views. These views are rendered with a fixed elevation angle of 20°. The azimuth angle is sampled uniformly between 0° and 360°. The input image is the first rendered view, namely, the view with an azimuth of 0°.

Results are shown in Table 1, which demonstrates that VFusion3D significantly outperforms the baselines by large margins. This further demonstrates that utilizing synthetic data to scale 3D generative models holds great promise.

Method	PSNR↑	SSIM↑	LPIPS↓
OpenLRM	14.21	0.820	0.226
LGM	13.53	0.811	0.254
VFusion3D	**20.89**	**0.848**	**0.127**

Table 1: Performance Comparison on the GSO Dataset: VFusion3D significantly outperforms the baselines by substantial margins.

4 Discussions on Limitations and Potential Solutions

The limitations section of the main paper presents that the fine-tuned video generator does not always yield flawless results. This is particularly noticeable in scenarios involving vehicles and texts, where the model sometimes generates multi-view results that lack 3D consistency. Additional examples of this are presented in Figure 1.

One way to assess the failure rate and check the pixel-level multi-view correspondence is to run Structure from Motion (SfM) algorithms, which require pixel-level correspondences, on generated multi-view frames. We run VGGSfM [13] on

Fig. 1: Samples of failure cases generated by the fine-tuned video diffusion model. In these instances, our fine-tuned video diffusion model struggles to generate high-quality multi-view sequences of text-related content and vehicles, resulting in distortions and 3D inconsistencies. Most of these failure cases are subsequently filtered out by the designed filter.

1000 fine-tuned EMU Video-generated multi-view sequences. Among these sequences, VGGSfM is able to obtain reasonable pose estimations (at least accurate poses for 8 views) over 70% of the time. Examples are presented in Figure 2. This demonstrates that the generated sequences generally have reasonable pixel-level 3D correspondence, but there is significant room for further improvements. Enforcing stronger multi-view consistency constraints [5,11,12,14] during the video fine-tuning stage is a plausible solution. This could involve the use of epipolar constraints and 3D-aware Plücker ray embeddings.

5 Conversation to Meshes

We use the marching cubes algorithm [6] to extract meshes from the generated NeRF results. Visualizations of sample converted meshes are shown in Figure 3.

References

1. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the International Conference on Computer Vision (ICCV) (2021) 1
2. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2553–2560. IEEE (2022) 2

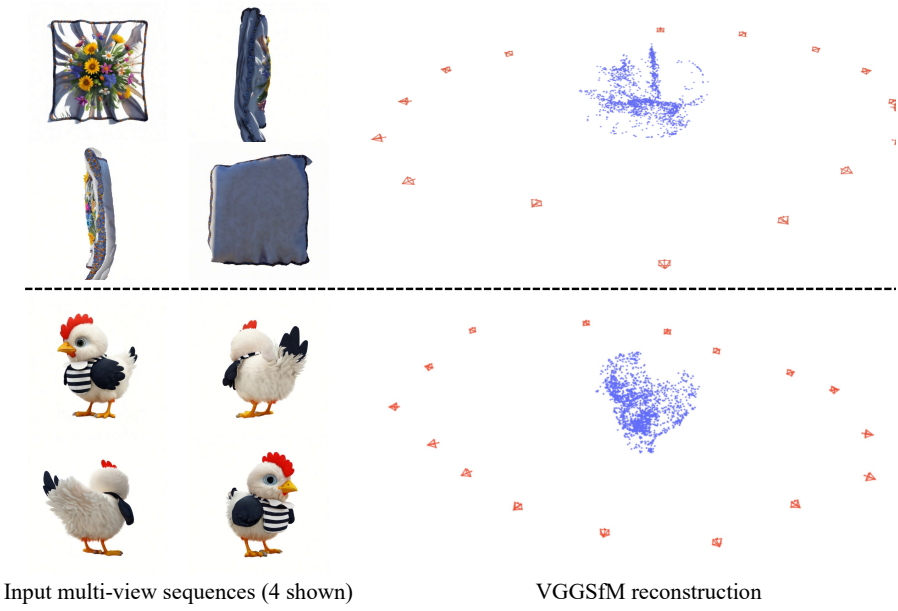

Fig. 2: VGGSfM reconstruction on all 16 generated frames. Our fine-tuned EMU Video typically produces multi-view sequences with reasonable pixel-level correspondences.

3. Girdhar, R., Singh, M., Brown, A., Duval, Q., Azadi, S., Rambhatla, S.S., Shah, A., Yin, X., Parikh, D., Misra, I.: Emu video: Factorizing text-to-video generation by explicit image conditioning. arXiv preprint arXiv:2311.10709 (2023) 1
4. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d. ICLR (2024) 1, 2
5. Kant, Y., Siarohin, A., Wu, Z., Vasilkovsky, M., Qian, G., Ren, J., Guler, R.A., Ghanem, B., Tulyakov, S., Gilitschenski, I.: Spad: Spatially aware multi-view diffusers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10026–10038 (2024) 3
6. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. In: Seminal graphics: pioneering efforts that shaped the field, pp. 347–353 (1998) 3
7. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 1
8. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 1
9. Qin, X., Dai, H., Hu, X., Fan, D.P., Shao, L., Gool, L.V.: Highly accurate dichotomous image segmentation. In: ECCV (2022) 2
10. Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z.: Lgm: Large multi-view gaussian model for high-resolution 3d content creation. arXiv preprint arXiv:2402.05054 (2024) 2

Fig. 3: Samples of converted meshes. We can create detailed and accurate meshes from the generated NeRF results in seconds.

11. Tang, S., Zhang, F., Chen, J., Wang, P., Yasutaka, F.: Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. arXiv preprint 2307.01097 (2023) 3
12. Voleti, V., Yao, C.H., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., Laforte, C., Rombach, R., Jampani, V.: Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. arXiv preprint arXiv:2403.12008 (2024) 3
13. Wang, J., Karaev, N., Rupprecht, C., Novotny, D.: Vggsfm: Visual geometry grounded deep structure from motion (2023) 2
14. Yang, J., Cheng, Z., Duan, Y., Ji, P., Li, H.: Consistnet: Enforcing 3d consistency for multi-view images diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7079–7088 (2024) 3