VID2ORBIT: ADAPTING VIDEO DIFFUSION AS MULTI-VIEW DIFFUSION FOR SINGLE-IMAGE SCENE RECONSTRUCTION

Anonymous authors
Paper under double-blind review

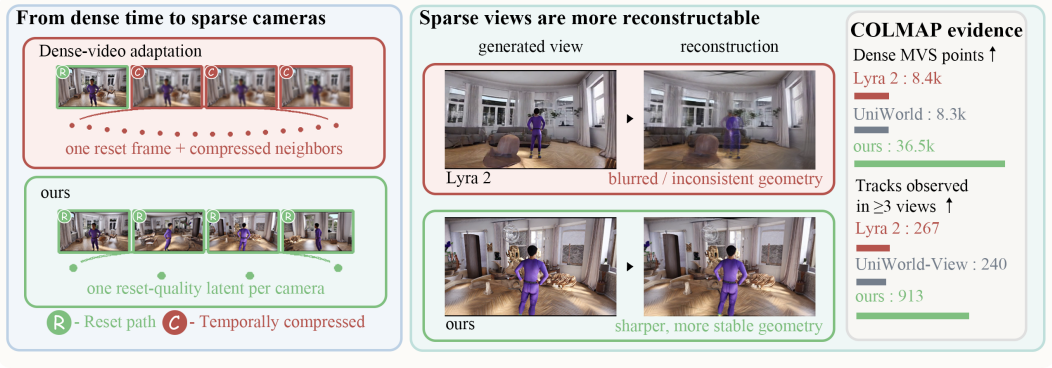

Figure 1: **VID2ORBIT modifies what video diffusion produces for reconstruction.** Left: dense-video adaptation retains the native temporal codec, in which a reset frame is followed by lower-fidelity temporally compressed neighbors. VID2ORBIT instead maps every target camera to the native single-frame reset path, producing sparse camera-indexed views. Right: this representation preserves a difficult rear view and yields a sharper ReSplat reconstruction; the same 20 views also produce more dense MVS points and long COLMAP tracks than Lyra 2 and UniWorld-View. The left panel is schematic; Figures 4 and 5 provide the systematic qualitative comparisons.

ABSTRACT

Video diffusion models provide strong priors for single-image 3D reconstruction, but their temporal representation is designed for dense video rather than sparse reconstruction views. In the native video VAE, the reset frame follows a higher-fidelity single-frame path, whereas subsequent frames are temporally compressed; under viewpoint motion, these neighboring frames provide limited new angular coverage while carrying larger codec errors. We verify this effect without diffusion sampling: resetting the frozen VAE for every ground-truth view improves round-trip PSNR by 6.25 dB at 18° view spacing. Motivated by this finding, we introduce VID2ORBIT, which repurposes a pretrained video diffusion model as a sparse multi-view generator. Each requested camera uses the VAE's native single-frame path, while the diffusion transformer jointly generates the complete view set. We further introduce camera-aligned feature exchange: intermediate multi-view features are aggregated in a shared 3D field and read back along camera rays. The resulting camera-indexed views provide a practical interface to COLMAP and Gaussian reconstruction.

1 INTRODUCTION

Reconstructing a 3D scene from a single image requires inferring surfaces and regions that are not visible in the input. A practical approach is to generate additional camera views and use them as observations for multi-view reconstruction Shen et al. (2026); Li et al. (2026). Video diffusion models

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

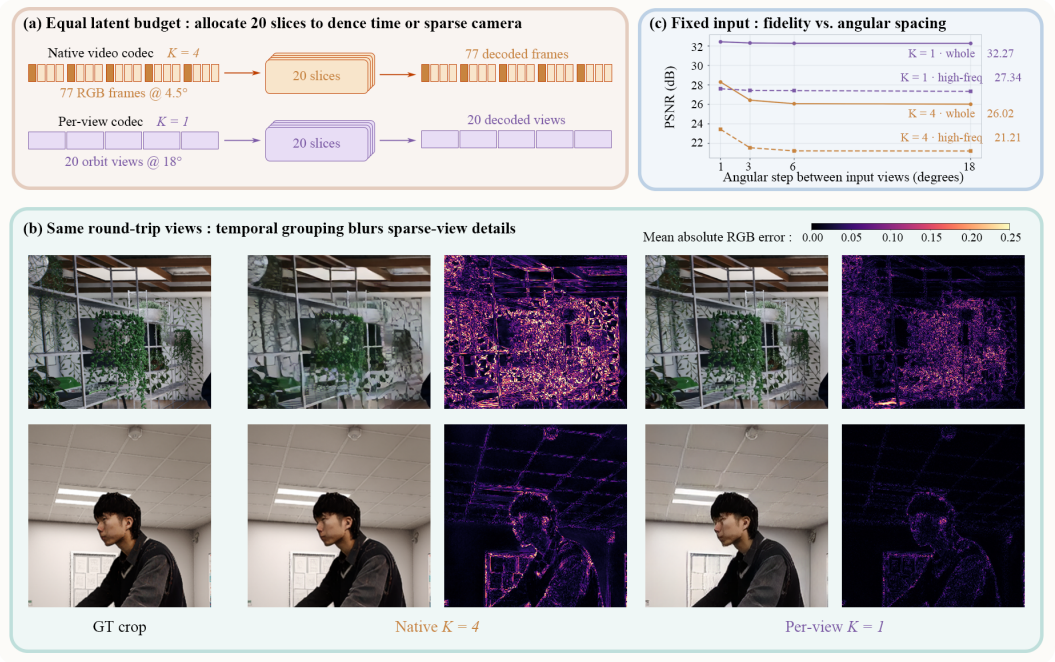

Figure 2: **Per-view VAE processing preserves sparse-orbit details.** (a) Allocation schematic: native $K = 1$ uses 20 latent slices to represent 77 dense frames, whereas per-view $K = 1$ assigns the same number of slices to 20 orbital views. (b) Round-trip crops and their mean absolute RGB error show that temporal grouping removes fine background and facial details. (c) Frozen-VAE round trips over the 20 scored views following a common reset image show the same trend in whole-image PSNR and in PSNR over the top 25% of pixels by ground-truth Laplacian magnitude: $K = 4$ remains stable as angular spacing increases, while $K = 4$ degrades and preserves less high-frequency detail.

are attractive for this purpose because large-scale pretraining provides rich priors for appearance, unseen-content completion, and viewpoint variation. However, visually plausible frames alone are insufficient: reconstruction requires fine detail, camera adherence, and consistency across views. Especially on person-centered scenes, which require completing unseen human and background surfaces while preserving fine-grained appearance and foreground–background consistency.

To address this gap, existing methods adapt video diffusion to novel-view generation through camera conditioning and geometric guidance (He et al., 2025; Yu et al., 2024; Ren et al., 2025; Shen et al., 2026). These systems inherit the native video representation, raising a more basic question: does the video codec already discard evidence needed for reconstruction? To isolate this factor, we encode and decode ground-truth orbital views with the frozen video VAE, without diffusion sampling. The round trip alone produces visible losses in fine background structure and human appearance, as shown in Figure 2.

The native VAE processes the reset frame through a single-frame path but temporally groups each subsequent four-frame block into one latent slice. This creates uneven fidelity: the reset frame remains relatively sharp, whereas the compressed views lose more high-frequency detail as the viewpoint changes. Because densely neighboring views add little angular coverage yet can introduce these larger codec errors, we instead reset the VAE temporal state for every requested camera. Across the 20 scored orbit views following a common reset image, this change improves PSNR from 26.02 to 32.27 dB at 18° spacing.

Inspired by this finding, we introduce VID2ORBIT, which repurposes a pretrained video diffusion model as a reconstruction-ready multi-view generator. Every target camera uses the frozen VAE's native single-frame encoding and decoding paths, establishing a one-to-one correspondence between cameras, latent maps, and output images without modifying the VAE. Using this path for a densely

sampled video would increase the latent sequence length by roughly four times. Our reconstruction setting instead requests 20 cameras distributed over a complete orbit: 20 latent slices that natively represent 77 dense video frames are allocated to 20 distinct views. All slices remain in one transformer sequence and are jointly generated through cross-view attention.

The explicit camera-to-latent correspondence also enables geometric exchange during denoising. We introduce a camera-aligned readout that aggregates intermediate features from all views into a shared 3D feature field. Each view reads a residual from this field along its camera rays, allowing its tokens to incorporate spatial evidence from the complete view set before image decoding. The field is predicted from the current multi-view features and requires no external depth or ground-truth geometry at inference.

We evaluate VID2ORBIT through codec round trips, novel-view comparisons with recent baselines, and downstream reconstruction with COLMAP and ReSplat. On the synthetic benchmark, its generated views yield 45.8% more dense COLMAP points than the strongest compared baseline under the fixed-pose reconstruction protocol. A compact paired study further finds that GT-guided training of the shared readout improves dense reconstruction yield by 12.39% on a fixed real-world set.

2 RELATED WORK

Multi-view diffusion for 3D reconstruction. Pose-conditioned image diffusion expands one image into camera-specific views for reconstruction (Liu et al., 2023; 2024b; Shi et al., 2024; Long et al., 2023). Era3D and PSHuman specialize this interface for objects and people, while CAT3D targets scene-level NeRF and 3DGS reconstruction (Li et al., 2024a;b; Gao et al., 2024). These methods establish sparse generated views as reconstruction observations, but often rely on image-diffusion backbones with task-specific posed data, cross-view modules, or domain priors such as SMPL-X. VID2ORBIT instead starts from a large video model, retaining its pretrained completion and appearance priors while converting dense frames into camera-indexed views.

Camera-controlled and geometry-aware video diffusion. CameraCtrl and ViewCrafter steer video generation with poses or camera rays, while GEN3C and ReconX render estimated geometry into target cameras (He et al., 2025; Yu et al., 2024; Sun et al., 2024; Ren et al., 2025; Liu et al., 2024a). The latter improve spatial control but inherit errors from monocular depth, pose, or point-cloud estimation. Longer-range systems use recurrent caches or extended trajectories (Shen et al., 2026; Zhou et al., 2026; Chen et al., 2024b; Suzuki et al., 2026). These methods retain the native video codec, where reset and compressed frames have different fidelity and several RGB frames can share one latent index. VID2ORBIT first establishes one camera per slice, then uses this mapping for camera-aligned exchange through a predicted 3D field without external geometry at inference. Correspondence, representation alignment, and attention supervision provide related internal guidance (Nam et al., 2025; Yu et al., 2025; Kwon et al., 2025; Wang et al., 2025); our readout instead aggregates all view-indexed features in one field and reads them back along camera rays.

Per-view video representations. FrameCrafter is the closest representation-level precedent (Wu et al., 2026). It independently encodes posed observations with the single-frame Wan VAE path, removes temporal positional embeddings, and predicts query views with query-centered ray conditioning. Its main setting uses several observed views and evaluates a synthesized query as novel-view prediction. VID2ORBIT instead completes an otherwise unseen full orbit from one image and evaluates the complete set through SfM, MVS, and Gaussian reconstruction. AnyRecon also uses a frame-wise VAE, but couples it to an external geometry initializer, persistent 3D memory, view retrieval, block-sparse attention, diffusion distillation, and iterative reconstruction (Chen et al., 2026). VID2ORBIT retains the pretrained VAE and joint transformer, changes their frame-to-latent contract, and keeps reconstruction modular. Its contribution is the resulting joint sparse-orbit generator and camera-aligned grid, rather than per-view encoding alone.

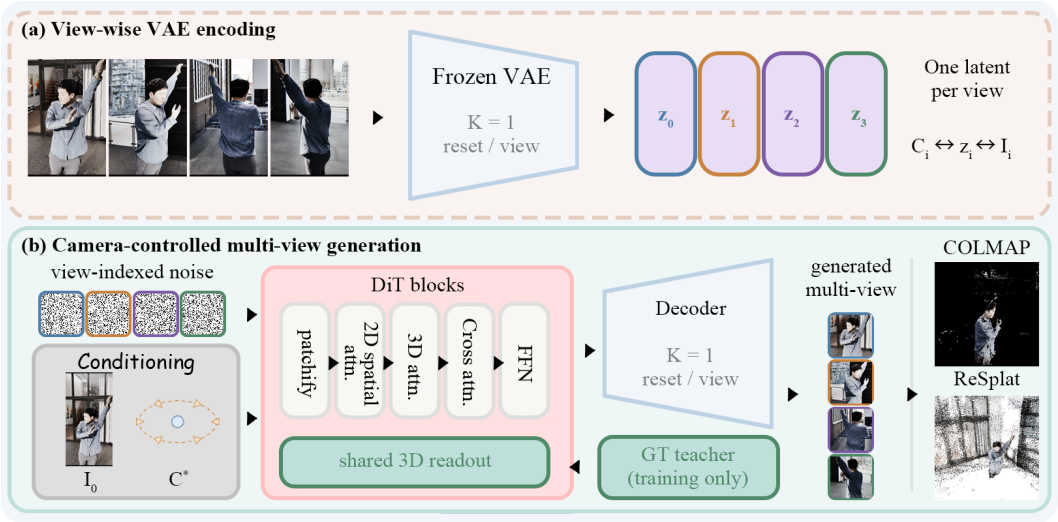

Figure 3: **VID2ORBIT pipeline.** (a) The frozen VAE is reset for every requested camera, producing one latent map per view and preserving the correspondence $C_i \leftrightarrow z_i \leftrightarrow I_i$. (b) All view-indexed latents are denoised in one camera-conditioned DiT sequence. The shared readout aggregates intermediate features in 3D and returns camera-aligned residuals, while the GT teacher provides training-only guidance. Per-view decoding then supplies the same RGB–camera set to external COLMAP or ReSplat reconstruction.

3 METHOD

3.1 TASK AND OVERVIEW

Given one RGB image I_0 of a person and the surrounding scene, the system instantiates a fixed orbit

$$\mathcal{C}^\star = \{C_i^\star\}_{i=0}^{N-1}, \qquad N = 20, \tag{1}$$

with cameras separated by 18° from 0° to 342°. VID2ORBIT generates the corresponding view set

$$\hat{\mathcal{I}} = G_\theta(I_0; \mathcal{C}^\star) = \{\hat{I}_i\}_{i=0}^{N-1}. \tag{2}$$

The semicolon distinguishes the user image from the system-defined orbit. Outputs are serialized by azimuth for the diffusion transformer, but each image retains its intended camera identity. The views and cameras are then passed to an external reconstruction method.

VID2ORBIT makes two changes to a pretrained camera-controlled video model. First, every requested camera uses the frozen VAE's native single-frame path, while all resulting latent slices are denoised jointly by the video transformer. Second, an intermediate transformer layer exchanges information through a shared 3D feature field aligned with the target cameras. Figure 3 shows the complete pipeline.

Figure 3 also makes the dependency between the two components explicit. The reset codec first assigns every latent grid to exactly one requested camera. This correspondence makes it possible to project intermediate tokens into a shared field and read them back along the same rays; the exchange would be ambiguous under four-frame temporal groups. The field only mediates denoising, and the final outputs remain ordinary RGB images with known cameras, so downstream reconstructors require no modification.

3.2 RESETTING THE VIDEO CODEC FOR SPARSE VIEWS

The Wan2.1 VAE used by our parent model (Wan Team et al., 2025) processes the first frame through a single-frame path and temporally compresses every later group of four frames. With M latent slices, it therefore decodes

$$T_{\text{native}}(M) = 1 + 4(M-1) \tag{3}$$

RGB frames. This is efficient for dense video, but it creates uneven fidelity between the reset frame and later compressed frames. It also prevents one latent index from denoting one camera observation.

Let E and D denote the frozen VAE and s its causal temporal state. We reuse its first-frame path as

$$E_1(I) = E(I; s \leftarrow \varnothing), \qquad D_1(z) = D(z; s \leftarrow \varnothing). \tag{4}$$

During training, each target view is encoded independently and the slices are concatenated along the view axis:

$$z_i^0 = E_1(I_i), \qquad Z^0 = \mathrm{Cat}_{\mathrm{view}}(z_0^0, \ldots, z_{N-1}^0). \tag{5}$$

At inference, the model starts from N noise slices and decodes each predicted slice with D_1. We call this reset policy $K = 1$, in contrast to the native four-frame grouping $K = 4$. No VAE layer or parameter is changed.

The latent slices still form one transformer sequence, so self-attention connects all views throughout denoising. VID2ORBIT changes the frame-to-latent contract rather than generating images independently. Under a 20-slice budget, native $K = 4$ represents 77 densely sampled frames, whereas $K = 1$ represents 20 distinct cameras. We use the latter allocation because the compressed intermediate frames add little angular coverage and exhibit larger codec errors, while 20 evenly spaced views cover the full orbit.

3.3 JOINT CAMERA-CONDITIONED DENOISING

We initialize the generator from a Wan2.1-Fun 14B image-to-video model with an existing camera adapter. The adapter keeps its inherited Plücker-ray packing and interpolation interface. The VAE, base DiT, and camera adapter remain frozen, while rank-64 LoRA modules (Hu et al., 2022) update the Q/K/V/O projections of self- and cross-attention. Thus, $K = 1$ exposes a camera identity for every slice but does not alter the camera parameterization.

For clean multi-view latents Z^0, noise ϵ, and time t, we use

$$Z^t = (1 - t)Z^0 + t\epsilon, \qquad v^* = \epsilon - Z^0, \tag{6}$$

and train the conditional velocity predictor with

$$\mathcal{L}_{\mathrm{rgb}} = \mathbb{E}\Big[\big\|v_\theta(Z^t, t, c(I_0, \mathcal{C}^\star)) - v^*\big\|_2^2\Big]. \tag{7}$$

All 20 slices are noised and predicted together. At inference, the only conditions are I_0 and the predefined orbit. The source-facing image is generated rather than copied from the input.

3.4 CAMERA-ALIGNED FEATURE EXCHANGE

The one-to-one camera–latent mapping gives every intermediate token grid an explicit projection model. We reshape a DiT hidden state into view-indexed grids and project normalized features from all views into a common volume. For each voxel, a small 3D network maps the mean, variance, and coverage of valid projected features to a feature field $\mathcal{V}_f$ and density field $\mathcal{V}_\sigma$. The volume covers the person-centered region defined by the prescribed orbit. It is rebuilt from the current hidden features on every forward pass and serves as an internal exchange field, not as the final reconstruction.

For a token ray, let σ_k be the density and δ_k the length of interval k. The first-hit weights are

$$\alpha_k = 1 - \exp(-\sigma_k \delta_k), \qquad w_k = \alpha_k \prod_{j<k} (1 - \alpha_j). \tag{8}$$

The ray feature $\bar{f} = \sum_k w_k \mathcal{V}_f(x_k)$ is projected back to the DiT width and added to the corresponding token. Every camera therefore reads from the same predicted field while the original transformer stream remains intact.

Unlike a post-hoc 3D decoder, this exchange occurs before image decoding and returns only a residual to the original DiT stream. It preserves the pretrained synthesis pathway while using camera geometry to route evidence across views. Because the field is rebuilt from current tokens, inference requires neither an external depth estimator nor persistent scene state.

We train the camera-aligned readout jointly with $\mathcal{L}_{\mathrm{rgb}}$ and a geometric loss on visible first hits and free space. A GT-guided variant adds a frozen copy of the shared branch as a teacher. Using training-time depth and person masks, the teacher supplies a target readout residual and first-hit distribution. The student continues to construct and read its own predicted field. The additional losses are a normalized residual MSE and $D_{\mathrm{KL}}(q_{\mathrm{GT}} \| p_{\mathrm{pred}})$. The teacher and all GT geometry are removed at inference. Implementation and sampling details are provided in Appendix B.

4 EXPERIMENTS

We evaluate the two parts of VID2ORBIT at different scales. Frozen-VAE round trips isolate the effect of per-view reset. Full generation and reconstruction comparisons test the sparse-orbit interface against recent baselines. A smaller paired study evaluates GT-guided training of the shared readout.

4.1 EXPERIMENTAL SETUP

Model and data. We adapt a Wan2.1-Fun 14B camera-controlled parent model. The VAE, base DiT, and inherited camera adapter remain frozen, while rank-64 LoRA modules update the Q/K/V/O projections of self- and cross-attention. Training uses 3,362 synthetic 20-view orbits at 832×480, spanning 10 scene assets and 706 identities. Appendix A describes the rendering pipeline, and Appendix B gives optimization details.

Baselines. We compare GEN3C (Ren et al., 2025), FlashWorld (Li et al., 2026), Lyra 2 (Shen et al., 2026), and UniWorld-View (Zhou et al., 2026). We normalize orbit direction and select the nearest target azimuths, while retaining each method's native cameras, field of view, and scale. Thus, the comparison evaluates delivered orbit interfaces rather than pixel-aligned cameras.

Inference. Unless stated otherwise, VID2ORBIT samples 20 views with 30 diffusion steps, shift 5, and image guidance scale 2. The input image and predefined orbit are the only conditions.

Reconstruction protocols. MVSplat360, FlashWorld, Lyra, Gen3R, and OneWorld couple generative priors to learned 3D representations through dedicated decoders, alignment objectives, or iterative reconstruction (Chen et al., 2024a; Li et al., 2026; Bahmani et al., 2026; Huang et al., 2026; Gao et al., 2026). We instead test whether the generated, posed RGB views provide a reusable interface to independent reconstruction systems. Fixed-pose COLMAP (Schönberger & Frahm, 2016; Schönberger et al., 2016) consumes each view with its intended camera and reports sparse points, fused dense points, tracks spanning at least three views, and retained reprojection error. These statistics measure the evidence preserved through feature matching and fusion rather than complete 3D accuracy. ReSplat (Xu et al., 2025) independently consumes the same RGB views and cameras to produce a renderable Gaussian representation. Together, the two protocols measure matchable multi-view evidence and stability under learned reconstruction, complementing evaluations based on epipolar matches, predicted geometry, or COLMAP yield (Yu et al., 2023; Asim et al., 2025; Shin et al., 2026).

4.2 SPARSE-VIEW CODEC FIDELITY

We first remove diffusion sampling. Both variants process a common reset image followed by the same 20 ground-truth orbit views; the reset image is processed identically and excluded from scoring. Per-view $K = 1$ clears temporal state before every evaluated view and uses one slice per view, whereas native $K = 4$ processes the reset once and represents the 20 evaluated views with five compressed slices. This experiment compares fidelity at different latent rates, not matched-compute performance.

Per-view PSNR remains nearly constant with spacing, whereas native $K = 4$ drops from 28.30 to 26.02 dB; the gap grows from 4.13 dB at 1° to 6.25 dB at 18°. Figure 2 localizes the loss to background texture, facial boundaries, and hair, directly exposing errors in compressed frames.

Table 1: Frozen-VAE round-trip PSNR over 20 orbit views following an unscored common reset image. Higher is better.

Angular step	1°	3°	6°	18°
Native $K = 4$	28.30	26.44	26.07	26.02
Per-view $K = 1$	**32.43**	**32.30**	**32.27**	**32.27**



Figure 4: **Single-image orbital-view generation.** FlashWorld, GEN3C, Lyra 2, UniWorld-View, and VID2ORBIT are shown on two real-world inputs at the nominal azimuths marked on the right. We normalize orbit direction and select the nearest target view while retaining each method's native cameras and fields of view. The comparison exposes errors that arise during generation, before any reconstruction: human or scene geometry can deform, the realized viewpoint can disagree with the requested rotation, and identity, layout, or illumination can drift across views. In these examples, VID2ORBIT more consistently retains the input subject and scene structure over the displayed orbit.

4.3 ORBITAL-VIEW GENERATION

Figure 4 compares selected wide-baseline directions for two real-world inputs, testing person identity, unseen-side completion, camera adherence, and scene continuity. FlashWorld renders its predicted Gaussians; the other methods generate RGB directly.

We read each row as a camera-indexed view set rather than as independent, visually plausible frames. Several baseline outputs remain locally realistic but fail at the set level: the person or scene deforms, the viewpoint under- or overshoots the requested rotation, or identity, layout, and illumination change between azimuths. These discrepancies are already present before reconstruction and therefore cannot be attributed to the downstream 3D model. Across the displayed views, VID2ORBIT better preserves the input subject and surrounding structure while following the prescribed orbit. The real-image comparison remains qualitative because the hidden views are unavailable and the native camera protocols differ.

4.4 DOWNSTREAM RECONSTRUCTION SUPPORT

We reconstruct from 20 selected views per method. Figure 5 pairs fixed-pose COLMAP fused points with ReSplat renders for one synthetic and one real input. The former shows where multi-view evidence survives matching and fusion; the latter shows how the same observations behave in a learned renderable representation.

VID2ORBIT recovers 45.8% more dense points than UniWorld-View, more than doubles long tracks, and obtains the lowest retained reprojection error. Figure 5 complements these synthetic-set averages with synthetic and real reconstructions. The displayed results retain sharper people and more stable scene structure, though duplicated limbs, ghosting, and missing surfaces remain.



Figure 5: **Reconstruction from generated orbital views.** Lyra 2, UniWorld-View, and VID2ORBIT are compared on synthetic (top) and real (bottom) inputs. Each row shows the input, fixed-pose COLMAP MVS points, and ReSplat renders at 90°, 180°, and 270°, using the same downstream pipelines. The MVS cloud shows where the generated views retain mutually matchable evidence; sparse or diffuse regions indicate failed correspondence. ReSplat renders expose the downstream effects of inconsistent observations, including softened structure, duplicated geometry, and view-dependent ghosting. In the displayed cases, VID2ORBIT supports a denser person-centered reconstruction and more stable novel-view renders.

Table 2: Fixed-pose COLMAP reconstruction over 20 held-out synthetic inputs. Methods contribute 20 views with their prescribed cameras. Higher is better except for reprojection error.

Method	Sparse pts. ↑	Dense pts. ↑	Tracks ≥ 3 ↑	Reproj. px ↓
Lyra 2	284.45	4,104.95	166.20	1.427
UniWorld-View	460.95	9,738.50	242.35	1.414
VID2ORBIT	**833.90**	**14,197.60**	**635.70**	**1.233**

4.5 SHARED-READOUT DISTILLATION

We evaluate GT-guided readout training with a paired 500-step study. Both arms start from the same checkpoint, train on the same three examples, and contain the same DiT LoRA, shared 3D field, RGB loss, and geometric loss. Only the distilled arm adds the teacher residual and first-hit targets from Section 3.4. We evaluate both checkpoints with the same seed on Synthetic 20 and Real 20.

On Real 20, distillation increases dense points by 12.39% and long tracks by 8.53%, while reducing reprojection error by 4.75%. Figure 6 shows local improvements in foreground detail and nearby structure. Thus, GT-guided training can strengthen reconstruction evidence from camera-aligned exchange; Appendix B gives full statistics.

5 DISCUSSION AND LIMITATIONS

Sparse views reallocate video capacity. VID2ORBIT replaces a dense sequence of uneven codec fidelity with fewer camera-indexed observations. This choice fits a fixed full-orbit reconstruction

Table 3: Dense COLMAP points with and without GT-guided distillation; both variants contain the shared readout.

Set	w/o distill.	+ GT-guided	Change
Synthetic 20	**12,407.10**	12,355.10	−0.42%
Real 20	5,854.80	**6,580.30**	+12.39%

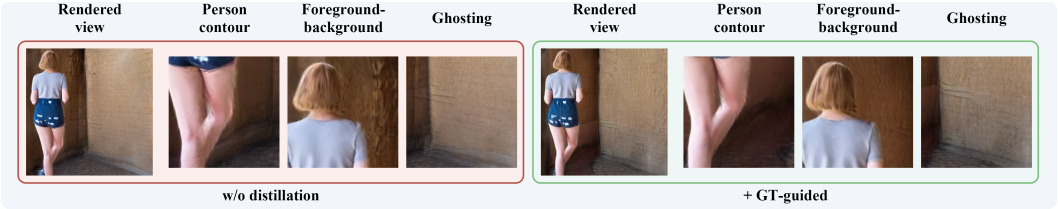

Figure 6: **Local effects of GT-guided readout distillation.** The columns isolate the full rendered view, person contour, foreground–background boundary, and a region affected by ghosting. With the same real input, view, seed, and reconstruction protocol, GT guidance yields a cleaner contour, clearer foreground–background separation, and less ghosting.

task, but it does not make additional views free: latent length, memory, and transformer cost still grow with the number of requested cameras. Twenty views are the operating point used here rather than a universal optimum.

Reconstructability does not imply physical truth. A single image cannot determine the hidden appearance or geometry of the person or scene. The model can generate a mutually compatible completion without recovering the real unobserved surface. An intended camera also does not guarantee exact pose adherence: changes in shape, orientation, or scale can appear as camera error during reconstruction. Synthetic targets permit direct image comparison, whereas real photographs support plausibility and internal-consistency analysis only.

Reconstruction metrics are complementary. COLMAP rewards stable local matches and discards inconsistent candidates, while ReSplat can trade geometric accuracy against its learned rendering prior. Point yield, reprojection error, and render quality must therefore be interpreted together. The current model is trained mainly on centered standing people in indoor and semi-indoor scenes. Close portraits, unusual poses, multiple people, fine hair, transparent objects, and broader outdoor conditions remain challenging.

Shared-readout gains remain domain dependent. GT-guided distillation improves fixed-pose reconstruction support on Real 20 but does not reproduce the same aggregate gain on Synthetic 20. The paired study isolates the distillation targets because both arms already contain the shared field and geometric loss. Measuring the complete branch against matched RGB-only adaptation remains a separate experiment.

6 Conclusion

We presented VID2ORBIT, which repurposes a pretrained video diffusion model as a sparse multi-view generator for single-image 3D reconstruction. Resetting the frozen VAE for every target camera preserves the high-fidelity single-frame path, while the video DiT continues to generate the complete orbit jointly. A camera-aligned shared field further exchanges intermediate features across views before decoding. Per-view reset improves frozen-VAE round-trip PSNR by 6.25 dB at 18° spacing, and the generated views support COLMAP and ReSplat reconstruction across synthetic and real inputs. The results show that a video model's temporal interface can be redesigned around distinct reconstruction views rather than dense playback frames.

AI USE STATEMENT

Generative AI tools were used to assist manuscript outlining, drafting, language editing, literature-search planning, and code editing. The authors reviewed the resulting prose, citations, claims, experimental records, and code, and take responsibility for the final content. **This statement must be reviewed and updated by all authors before submission so that it fully reflects their use of generative AI tools.**

ETHICS STATEMENT

The method synthesizes novel views of people and scenes and may reproduce biases present in its training data or be misused to create misleading visual content. The work uses licensed or research-authorized assets and data according to their respective terms. A release should document data provenance, model limitations, and appropriate use. No identity or demographic claims are made from generated images.

REPRODUCIBILITY STATEMENT

Section 3 specifies the view representation and geometric supervision, and Section 4 reports the training and evaluation protocols. Appendix B records implementation details, data-split qualifications, and metric definitions. We will provide configuration files, evaluation scripts, and an anonymous code package with the submission.

REFERENCES

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. MEt3R: Measuring multi-view consistency in generated images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Sherwin Bahmani, Tianchang Shen, Jiawei Ren, Jiahui Huang, Yifeng Jiang, Haithem Turki, Andrea Tagliasacchi, David B. Lindell, Zan Gojcic, Sanja Fidler, Huan Ling, Jun Gao, and Xuanchi Ren. Lyra: Generative 3D scene reconstruction via video diffusion model self-distillation. In *International Conference on Learning Representations*, 2026.

Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. *arXiv preprint arXiv:2306.16940*, 2023.

Yuedong Chen, Chuanxia Zheng, Haofei Xu, Bohan Zhuang, Andrea Vedaldi, Tat-Jen Cham, and Jianfei Cai. MVSplat360: Feed-forward 360° scene synthesis from sparse views. *arXiv preprint arXiv:2411.04924*, 2024a.

Yutian Chen, Shi Guo, Renbiao Jin, Tianshuo Yang, Xin Cai, Yawen Luo, Mingxin Yang, Mulin Yu, Linning Xu, and Tianfan Xue. AnyRecon: Arbitrary-view 3D reconstruction with video diffusion model. *arXiv preprint arXiv:2604.19747*, 2026.

Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3D: Video diffusion models are effective 3D generators. *arXiv preprint arXiv:2403.06738*, 2024b.

Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole. CAT3D: Create anything in 3D with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024.

Sensen Gao, Zhaoqing Wang, Qihang Cao, Dongdong Yu, Changhu Wang, Tongliang Liu, Mingming Gong, and Jiawang Bian. OneWorld: Taming scene generation with 3D unified representation autoencoder. *arXiv preprint arXiv:2603.16099*, 2026.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for text-to-video generation. In *International Conference on Learning Representations*, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Jiaxin Huang, Yuanbo Yang, Bangbang Yang, Lin Ma, Yuewen Ma, and Yiyi Liao. Gen3R: 3D scene generation meets feed-forward reconstruction. *arXiv preprint arXiv:2601.04090*, 2026.

Minkyung Kwon, Jinhyeok Choi, Jiho Park, Seonghu Jeon, Jinhyuk Jang, Junyoung Seo, Minseop Kwak, Jin-Hwa Kim, and Seungryong Kim. CAMEO: Correspondence-attention alignment for multi-view diffusion models. *arXiv preprint arXiv:2512.03045*, 2025.

Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, Wenping Wang, Qifeng Liu, and Yike Guo. Era3D: High-resolution multiview diffusion using efficient row-wise attention. *arXiv preprint arXiv:2405.11616*, 2024a.

Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Mengfei Li, Xiaowei Chi, Siyu Xia, Wei Xue, Wenhan Luo, Qifeng Liu, and Yike Guo. PSHuman: Photorealistic single-view human reconstruction using cross-scale diffusion. *arXiv preprint arXiv:2409.10141*, 2024b.

Xinyang Li, Tengfei Wang, Zixiao Gu, Shengchuan Zhang, Chunchao Guo, and Liujuan Cao. FlashWorld: High-quality 3D scene generation within seconds. In *International Conference on Learning Representations*, 2026.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. DL3DV-10K: A large-scale scene dataset for deep learning-based 3D vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. ReconX: Reconstruct any scene from sparse views with video diffusion model. *arXiv preprint arXiv:2408.16767*, 2024a.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3D object. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. SyncDreamer: Generating multiview-consistent images from a single-view image. In *International Conference on Learning Representations*, 2024b.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3D: Single image to 3D using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*, 2023.

Jisu Nam, Soowon Son, Dahyun Chung, Jiyoung Kim, Siyoon Jin, Junhwa Hur, and Seungryong Kim. Emergent temporal correspondences from video diffusion transformers. *arXiv preprint arXiv:2506.17220*, 2025.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-informed world-consistent video generation with precise camera control. *arXiv preprint arXiv:2503.03751*, 2025.

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, 2016.

Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, Sanja Fidler, Jiahui Huang, Huan Ling, Jun Gao, and Xuanchi Ren. Lyra 2.0: Explorable generative 3D worlds. *arXiv preprint arXiv:2604.13036*, 2026.

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. MVDream: Multi-view diffusion for 3D generation. In *International Conference on Learning Representations*, 2024.

Minjung Shin, Hyunin Cho, Sooyeon Go, Jin-Hwa Kim, and Youngjung Uh. MVCustom: Multi-view customized diffusion via geometric latent rendering and completion. In *International Conference on Learning Representations*, 2026.

Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, Yikai Wang, et al. DimensionX: Create any 3D and 4D scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 2024.

Keito Suzuki, Kunyao Chen, Lei Wang, Bang Du, Runfa Blark Li, Peng Liu, Ning Bi, and Truong Nguyen. HumanOrbit: 3D human reconstruction as 360° orbit generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Joachim Tesch, Giorgio Becherini, Prerana Achar, Anastasios Yiannakidis, Muhammed Kocabas, Priyanka Patel, and Michael J. Black. BEDLAM2.0: Synthetic humans and cameras in motion. *arXiv preprint arXiv:2511.14394*, 2025.

Wan Team, Ang Wang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Qi Wu, Khiem Vuong, Minsik Jeon, Srinivasa Narasimhan, and Deva Ramanan. FrameCrafter: Novel view synthesis as video completion. *arXiv preprint arXiv:2604.08500*, 2026.

Haofei Xu, Daniel Barath, Andreas Geiger, and Marc Pollefeys. ReSplat: Learning recurrent gaussian splats. *arXiv preprint arXiv:2510.08575*, 2025.

Jason J. Yu, Fereshteh Forghani, Konstantinos G. Derpanis, and Marcus A. Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. THuman2.0: A 3D real human dataset. In *IEEE/CVF International Conference on Computer Vision*, 2021.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, and Ying Shan. ViewCrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.

Haiyang Zhou, Wangbo Yu, Chaoran Feng, Xunyu Zhou, Yonghong Tian, and Li Yuan. UniWorld-View: Large-baseline view synthesis via video diffusion models. *arXiv preprint arXiv:2608.04701*, 2026.

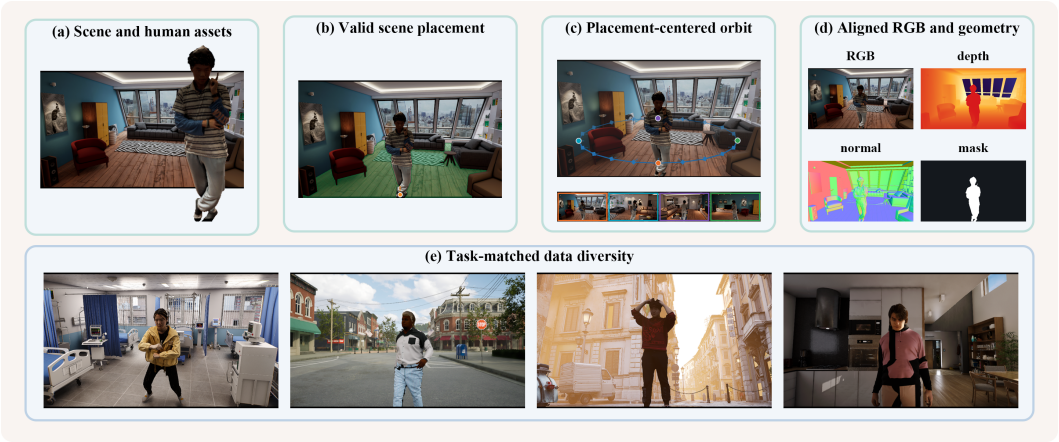

Figure 7: **Task-matched synthetic data.** The upper row shows scene and human selection, valid placement, construction of a 20-view person-centered orbit, and aligned RGB, depth, normals, and masks. The lower row illustrates the appearance and environment diversity used by the renderer.

A Task-Matched Training Data

Adapting a video model to the fixed orbit requires aligned RGB views, cameras, and training-time geometry. We use an Unreal Engine 5 pipeline derived from the asset and annotation structure of BEDLAM and BEDLAM2.0 (Black et al., 2023; Tesch et al., 2025). It imports human assets, finds valid placements in indoor or semi-indoor environments, fits a person-centered camera orbit, and exports RGB, metric depth, surface normals, person masks, intrinsics, and camera-to-world transforms. Only RGB and cameras are needed at inference.

The current instantiation keeps the person and scene static and renders 20 horizontal views at 832×480, separated by 18°. The filtered adaptation set contains 3,362 orbits from 10 scene assets and 706 human identities. It uses THuman-derived people (Yu et al., 2021) and is much smaller than general captured collections such as DL3DV-10K (Ling et al., 2024). Its role is to provide the exact orbit and geometry supervision required by the method.

A six-case development suite spans four scene assets and two identities and has already informed model development. A separate 52-example candidate pool is identifier-disjoint from the filtered adaptation manifest, although its assets may overlap with earlier parent-model training. Real photographs are evaluated without hidden-view ground truth. The current distribution emphasizes centered standing people with visible indoor or semi-indoor context. Broader poses, multiple people, and outdoor conditions remain outside this training instance.

B Additional Implementation and Evaluation Details

B.1 Native temporal mapping and per-view reset

For $T = 1 + 4m$ RGB frames, the native Wan causal VAE emits $1 + m$ temporal latent slices: one for the reset frame and one for every subsequent group of four. A budget of 20 native slices therefore decodes 77 video frames, whereas applying the single-frame path independently to 20 slices decodes 20 views. Our fixed-input round-trip comparison scores 20 orbit views following a common reset image; the reset image is processed identically by both variants and omitted from the metric. Our implementation invokes the native first-frame encode and decode paths separately for every requested camera and clears the temporal cache before each invocation. All codec and diffusion experiments use the same frozen VAE weights.

For a fixed set of 21 input images, native temporal grouping produces 6 slices while per-view processing produces 21. Table 1 is therefore a fidelity-versus-representation comparison at different latent rates. The same-latent-budget illustration in Figure 2(a) instead fixes 20 slices and changes the number of represented RGB frames.

B.2 TRAINING AND INFERENCE DETAILS

The adaptation set contains 3,362 20-view examples at 832×480. We use rank-64 LoRA on the Q/K/V/O projections of self- and cross-attention, corresponding to approximately 209.7M trainable parameters. The global batch size is 8 with gradient accumulation 2 across four GPUs. AdamW uses learning rate 5×10^{-6}, $(\beta_1, \beta_2) = (0.9, 0.95)$, zero weight decay, gradient clipping at 1, and 50 warmup steps.

Persistent trainable parameters, optimizer moments, and distributed reductions use FP32; forward and backward operations use BF16. The first-view latent and its image embedding are jointly dropped with probability 0.1 while the camera condition is retained. Text conditioning is empty. At inference we use 30 sampling steps, shift 5, and image guidance scale 2 unless a table states otherwise.

The inherited camera adapter retains the parent's Plücker-ray packing and interpolation convention. The view-discrete VAE changes the association between output images and latent slices but does not change this adapter. The source-facing output is sampled by the model and is not replaced with the input image after generation.

B.3 SHARED-READOUT IMPLEMENTATION AND PAIRED TRAINING

The shared branch reads the output of `core.blocks[24]` using zero-based indexing. A hidden tensor of shape $[1, 31200, 5120]$ is rearranged into $[20, 30, 52, 5120]$, normalized, and projected to 24 channels before camera-based aggregation. The shared field has resolution $[D, H, W] = [32, 48, 32]$ with 24 feature channels and one density channel. Each token ray is integrated over 48 finite intervals inside the orbit-derived bounding volume. Coordinates follow an OpenCV convention relative to the first camera and normalized by the orbit radius.

The ordinary geometric term supervises first-hit probability on visible person surfaces and free space before an observed surface. For distillation, each low-resolution token retains four high-resolution depth and mask samples. A 3×3 dilation and erosion define the boundary band; person interior, boundary, and valid background contribute at most 128 rays each and are averaged first within, then across, regions. Background rays whose observed surface lies inside or in front of the shared volume are treated as unknown. An empty-space target is used only when the observed depth clears the volume exit by a safety margin.

Both paired arms start from `adapt_gt/step-100.safetensors`, whose history includes the RGB-adapted ancestor followed by shared-branch adaptation. They train on three examples for 500 optimizer steps with global batch size 1. LoRA and shared-branch learning rates are 10^{-6} and 10^{-4}; AdamW uses $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, zero weight decay, and gradient clipping at 1. Geometry-branch drop is 0.15 and image-condition drop is 0.1. Both arms rebuild FP32 master parameters and optimizer states and use the same seed 73, noise, sigma, condition-drop, and supervision-sampling sequence.

The teacher is a frozen copy of the geometric branch rather than a second full DiT. It consumes the current student block features through a stop-gradient path, so its targets are recomputed rather than cached. The two runs both average about 21.8 seconds per step over steps 26–500 and peak near 54.67 GiB allocated memory. The ordinary implementation retains an unused teacher copy, so these measurements compare the executed paired code paths rather than the theoretically minimal memory cost of ordinary training.

B.4 DATA-SPLIT QUALIFICATIONS

Exact identifiers in the six-case development suite are absent from the filtered adaptation manifest. This does not establish that every human or scene asset was absent from all historical parent-model training. The suite has also informed method and checkpoint development, so aggregate results from it are diagnostic rather than an untouched generalization estimate.

The 52-example candidate pool is identifier-disjoint from the filtered adaptation manifest. Its final role, scene composition, and all-history overlap will be audited after the model configuration is

Table 4: Full paired evaluation of GT-guided readout distillation. Both variants contain the same shared readout and geometric loss. Tracks span at least three views; reprojection error is in pixels.

Set	Variant	Sparse ↑	Dense ↑	Tracks ↑	Reproj. ↓
Synthetic 20	Without distillation	**679.80**	**12,407.10**	**523.65**	1.206
	With distillation	639.85	12,355.10	483.80	**1.189**
Real 20	Without distillation	236.75	5,854.80	150.00	1.425
	With distillation	**243.05**	**6,580.30**	**162.80**	**1.357**

frozen. Real-image examples are reported separately because they do not provide ground-truth hidden views or geometry.

The reported readout-distillation evaluation contains the fixed synthetic and real sets of 20 inputs each. Every case uses generation seed 0.

B.5 Full shared-readout results

The RGB-adapted ancestor obtains 833.90/14,197.60 mean sparse/dense points on Synthetic 20 and 263.20/6,462.55 on Real 20. The corresponding values are 679.80/12,407.10 and 236.75/5,854.80 without distillation, and 639.85/12,355.10 and 243.05/6,580.30 with distillation. Because the paired arms share a later initialization, these values do not isolate the complete shared branch against RGB-only adaptation.

B.6 Fixed-pose COLMAP protocol

We supply the intended intrinsics and camera poses to prevent pose estimation from absorbing generation errors. Sparse feature extraction and matching are followed by triangulation, official GPU PatchMatch stereo at a maximum image size of 832, and stereo fusion. The current configuration uses 10 source images and five photometric-plus-geometric PatchMatch iterations. The fused point cloud provides the dense-point statistic.

Twenty supplied cameras in the reconstruction workspace do not mean that COLMAP successfully estimated or registered 20 cameras; registration is bypassed by the fixed-pose protocol. Sparse points, track length, reprojection error, and dense fused points describe the evidence retained after matching and reconstruction.

Point yield depends on stable texture, local contrast, high-frequency detail, view overlap, occlusion, source-image selection, and reconstruction thresholds. It is evidence that a view set supports this common reconstruction protocol, not a new or pure measure of geometric accuracy.

All reported means are arithmetic averages over cases rather than point-weighted aggregates. The 95% intervals for dense-point mean changes are obtained by resampling cases and equal $[+3.42\%, +22.15\%]$ on real 20 and $[-4.49\%, +3.52\%]$ on synthetic 20. They describe one fixed generation seed, not variation across training or generation runs. Distillation also reduces the mean number of detected SIFT features by 2.56% on real 20 and 3.83% on synthetic 20, so higher retained-observation fractions can partly reflect a smaller denominator.

B.7 ReSplat and baseline conventions

ReSplat receives the generated RGB images together with their intrinsics and camera-to-world transforms. It does not use the COLMAP sparse or dense point cloud. This separation allows COLMAP and ReSplat to reveal different consequences of view inconsistency.

Camera-controlled video baselines may generate more than 20 frames. For each such method, the final comparison will report the native output length, the target trajectory, and the frame-selection rule used to form the 20-view reconstruction input. Differences in field of view, scale gauge, orbit shape, crop, and lossy video encoding will be disclosed rather than described as a perfectly identical image pipeline.