GAE: LEARNING A GEOMETRY-NATIVE LATENT SPACE FOR 3D-CONSISTENT WORLD GENERATION

Anonymous authors
Paper under double-blind review

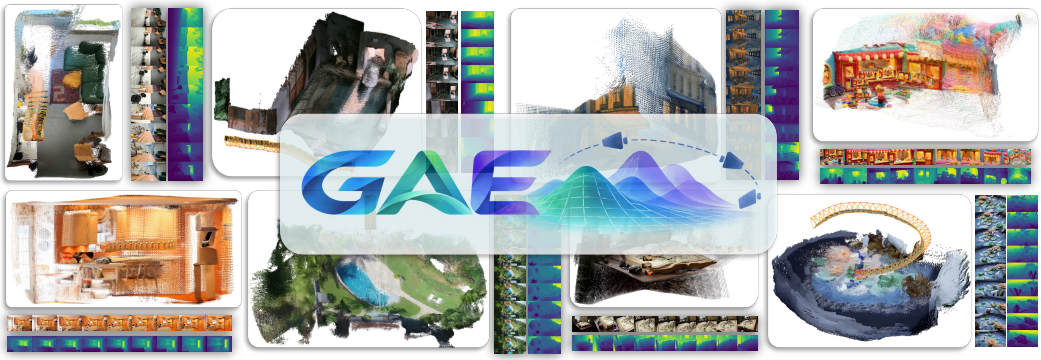

Figure 1: **Geometry-native world generation.** Given a reference image, a camera trajectory, and optional text, GAE jointly generates RGB views and 3D geometry from *a shared, compact latent space*. Each example pairs generated RGB and depth sequences with the resulting 3D point cloud.

ABSTRACT

We present a compact geometry-native latent space as a shared foundation for perception and generation. Visual generators can produce photorealistic frames without preserving a consistent 3D scene. We argue that this is not only a modeling problem but also a representation problem: generators typically evolve appearance-centric latents, while perception models recover geometry in a semantically rich space that encodes cross-view structure. Rather than adding geometry as another output, we reparameterize a geometry foundation model's features into a compact latent space for generation. We realize this shift with the *geometry-native autoencoder* (GAE), whose latent is jointly decodable to appearance, depth, cameras, and point maps. With this state, a standard conditional flow supports diverse generation tasks. In controlled comparisons that hold the generator and training protocol fixed, replacing the latent with GAE improves both visual quality and independently measured 3D coherence: FVD falls by 12.7% and 23.1% on RealEstate10K and DL3DV, and camera-trajectory error is halved on RealEstate10K. Together, these results show that the latent space is central to geometry-consistent generation and can serve as a shared interface between perception and generation. More details can be found on our project page.

1 INTRODUCTION

Photorealistic frames do not guarantee a coherent scene. From a few real views of one scene, geometry foundation models can recover depth, cameras, and point maps that agree in a common 3D coordinate frame (Wang et al., 2024a; 2025a). Video generators, however, can produce views whose recovered geometry drifts and camera trajectories stray from the requested path. *What perception can read, generation still struggles to write.* If video generation is to serve as a foundation for world models, its frames must describe a persistent scene. We trace this requirement upstream, to the latent space the generator is trained to evolve.

A latent state does more than compress the output. It shapes which structure is directly available to the generator and which must be inferred from appearance or introduced elsewhere. A pixel

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.

VAE preserves appearance (Kingma & Welling, 2014; Rombach et al., 2022), while a representation autoencoder organizes semantic content (Zheng et al., 2026; Singh et al., 2026), but neither makes depth, camera geometry, and cross-view relations natively readable. The usual remedy is to add geometry alongside an appearance-native latent, through camera controls, aligned representations, or geometry-aware post-training (He et al., 2025; Huang et al., 2026; An et al., 2026). *We argue that perception and generation should instead share a geometry-native latent space.*

Geometry foundation models offer a natural starting point for such a latent (Wang et al., 2024a; 2025a; Lin et al., 2026). Yet their geometry decoders typically rely on a hierarchy of features rather than a single latent representation. Different levels provide complementary cues, including local detail, correspondence, semantic context, and cross-view structure. This division of labor is effective for perception but awkward for generation: modeling the full hierarchy introduces multiple coupled generative states, whereas selecting one level discards information expected by the geometry decoder. GLD (Jang et al., 2026) illustrates the former approach by using a cascade of flow models, but the underlying mismatch applies more broadly to models that decode geometry from multi-level features. We use DA3 (Lin et al., 2026) as a controlled case study to make this broader perceptual–generative mismatch concrete: no single level simultaneously preserves the semantic, spatial, and cross-view structure needed for geometry-aware generation while providing a smooth, well-conditioned space for generative transport. Earlier features retain stronger scene structure but are harder to model continuously, whereas later features become smoother at the cost of substantial geometric and semantic information. Moreover, individual levels remain highly redundant and anisotropic: despite having 3,072 channels, they span only about 11 effective dimensions, with condition numbers ranging from 10^8 to 10^{16}. This leaves most channels nearly inactive and concentrates useful information along a small number of directions at vastly different scales. Therefore, the geometry is already present, but its raw parameterization is poorly matched to generation.

To reparameterize this hierarchy for generation, we introduce the *geometry-native autoencoder* (GAE). Rather than directly modeling the feature hierarchy, as in GLD (Jang et al., 2026), we argue that a generative reparameterization should satisfy three principles: (1) *Preserve* the geometry already encoded by the perception model; (2) *Unify* appearance and geometry within a single compact representation; and (3) *Organize* that representation for smooth generative transport. To preserve geometry, GAE places a learned bottleneck between DA3's frozen encoder and geometry head, compressing all four feature levels into a single latent and reconstructing the full hierarchy in one pass. Crucially, the original geometry head remains frozen when decoding depth, camera rays, and point maps, preventing the decoder from adapting around information lost during compression. To unify geometry and appearance, a learned RGB head decodes appearance from the same bottleneck, forcing both to coexist in a shared latent. These readouts constrain *what* information the latent preserves, but not *how* it is organized for flow modeling. We therefore organize the bottleneck at two complementary levels: individual tokens and relations among them. Token-wise alignment to co-located C-RADIO features (Heinrich et al., 2025) improves transport smoothness and semantic organization, but on its own sharply degrades pairwise spatial structure. Matching pairwise similarities in the posterior to those in DINOv2 features (Oquab et al., 2024) restores this relational geometry while retaining the token-level gains. GAE thus preserves geometry, unifies it with appearance in a compact latent, and organizes that latent natively for generation.

Using this latent, we train a unified generator based on a standard DiT-style conditional flow (Lipman et al., 2023). The same model supports text-to-image generation, camera-controlled video, and reference-conditioned novel-view synthesis. Condition dropout enables this breadth by exposing the model to different combinations of text, metric Plücker rays (Yin et al., 2026), and reference views. Across regimes, the model jointly generates all target-view latents, which decode natively to both RGB and geometry. Perception and generation therefore share the same latent space.

To isolate the effect of the latent space, we compare GAE with pixel, semantic, and raw geometry latents (Rombach et al., 2022; Singh et al., 2026; Lin et al., 2026) using matched flow models and training protocols. GAE reduces FVD by 12.7% on RealEstate10K (Zhou et al., 2018) and 23.1% on DL3DV (Ling et al., 2024) relative to the strongest competing latent. Independent evaluation of the generated views also shows stronger 3D consistency, with camera-trajectory error roughly halved on RealEstate10K. Together, these results establish a compact geometry-native latent space as a shared foundation for perception and generation.

2 Related Work

Generative representation spaces. Variational and vector-quantized tokenizers separate perceptual compression from generative modeling, with latent diffusion establishing compact codes as generative states (Kingma & Welling, 2014; van den Oord et al., 2017; Esser et al., 2021; Rombach et al., 2022). Representation autoencoders (RAEs) use frozen visual features, extending to text-to-image generation and multi-layer representations (Zheng et al., 2026; Tong et al., 2026; Singh et al., 2026). REPA aligns denoiser features without changing the output state, while latent-diffusability studies show that reconstruction alone does not determine generation quality (Yu et al., 2025a; Zhong et al., 2026). These methods center generation on appearance or semantics; GAE instead seeks a compact state natively readable by a geometry foundation model.

Geometry-aware visual generation. Novel-view synthesis reconstructs radiance fields or Gaussian primitives, or uses pose-conditioned diffusion to generate unseen content from text or sparse views (Mildenhall et al., 2020; Kerbl et al., 2023; Liu et al., 2023; Shi et al., 2024; Liu et al., 2024; Gao et al., 2024; Yu et al., 2024; 2025b). Camera-controlled video encodes trajectories through conditioning or positional encoding (Wang et al., 2024b; He et al., 2025; Yin et al., 2026), while joint-output methods generate RGB with depth or normals (Stan et al., 2023; Krishnan et al., 2025; Kwon et al., 2025; Guizilini et al., 2025). Other methods incorporate geometry through feature alignment, conditioning, coupled latents, scene-level tokens, or geometric rewards (Wu et al., 2026; Wan et al., 2026; Huang et al., 2026; Mi et al., 2026; Xiang et al., 2026; Wei et al., 2026; An et al., 2026). Across these designs, geometry augments an appearance-led state as a control, modality, auxiliary representation, or objective; GAE instead derives the sole evolving latent from a geometry foundation model.

Geometry foundation models as generative states. Geometry foundation models have progressed from paired point-map regression and matching to many-view, persistent, reference-free, and scaled static or dynamic reconstruction (Wang et al., 2024a; Leroy et al., 2024; Lu et al., 2025; Wang et al., 2025b;a; Lin et al., 2026; Wang et al., 2026c;b). Pretrained representations also support planning, geometry forecasting, and stochastic world modeling (Zhou et al., 2025a; Sun et al., 2026; Porcher et al., 2026). The closest visual-generation methods model geometry-foundation features: GLD cascades selected DA3 or VGGT levels and propagates the remaining hierarchy (Jang et al., 2026); latent Riemannian flow matching jointly evolves four normalized VGGT levels on their product manifold (Weijler et al., 2026); and OneWorld (Gao et al., 2026) augments Pi3 features with appearance and semantics to form a unified 3D representation. These methods retain high-dimensional generative states. GAE instead learns a compact Euclidean reparameterization of the full hierarchy, evolves it with a standard flow model, and reconstructs the hierarchy for frozen geometry readout and learned RGB decoding.

3 Method

GAE places the 3D inductive bias in the generated state. As shown in Fig. 2, it first learns a codec that maps frozen DA3 multi-level features to a compact latent with RGB and geometry readouts, then freezes the codec and learns a conditional flow in its standardized latent space:

$$\underbrace{\mathbf{I}_{1:V} \xrightarrow{\mathcal{E},\mathcal{P}} \mathbf{X}_{1:V} \xrightarrow{\mathrm{Enc}_\phi} \mathbf{z}_{1:V} \xrightarrow{\mathrm{Dec}_\phi} \hat{\mathbf{X}}_{1:V}}_{\text{Stage 1: train the codec } \mathcal{A}_\phi}, \qquad \underbrace{\bar{\mathbf{z}}^{(1)}_{1:V} \xrightarrow{\mathrm{Flow}_\psi(\cdot|\mathcal{C})} \hat{\bar{\mathbf{z}}}_{1:V} \xrightarrow{\mathrm{Std}^{-1}} \hat{\mathbf{z}}_{1:V}}_{\text{Stage 2: train the flow model } \mathbf{f}_\psi}, \tag{1}$$

here $\mathcal{E}$ is the frozen DA3 encoder and $\mathcal{P}$ normalizes and fuses its four feature levels. Stage 1 trains the codec to reconstruct $\mathbf{X}_{1:V}$; after freezing, its standardized posterior means provide the flow targets, and sampling starts from $\bar{\mathbf{z}}^{(1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The condition set $\mathcal{C}$ may contain text, camera rays, and clean reference latents; only target-view latents evolve, while these signals remain fixed controls. Both outputs are decoded from the sampled latent:

$$\hat{\mathbf{I}}_{1:V} = \mathcal{D}^{\mathrm{rgb}}_\theta(\hat{\mathbf{z}}_{1:V}), \qquad \hat{\mathbf{G}}_{1:V} = \mathcal{H}_{\mathrm{DPT}}(\mathrm{Dec}_\phi(\hat{\mathbf{z}}_{1:V}))\,. \tag{2}$$

The frozen DPT head reads geometry from the reconstructed hierarchy, while a learned head reads RGB directly. Unlike a pixel VAE, the state remains geometry-decodable; unlike raw-feature diffusion, it is compact enough for one flow network. We write $\mathbf{z}$ for the codec latent, $\bar{\mathbf{z}}$ for its standardized form, and $\bar{\mathbf{z}}^{\mathrm{ref}}$ for clean reference evidence.

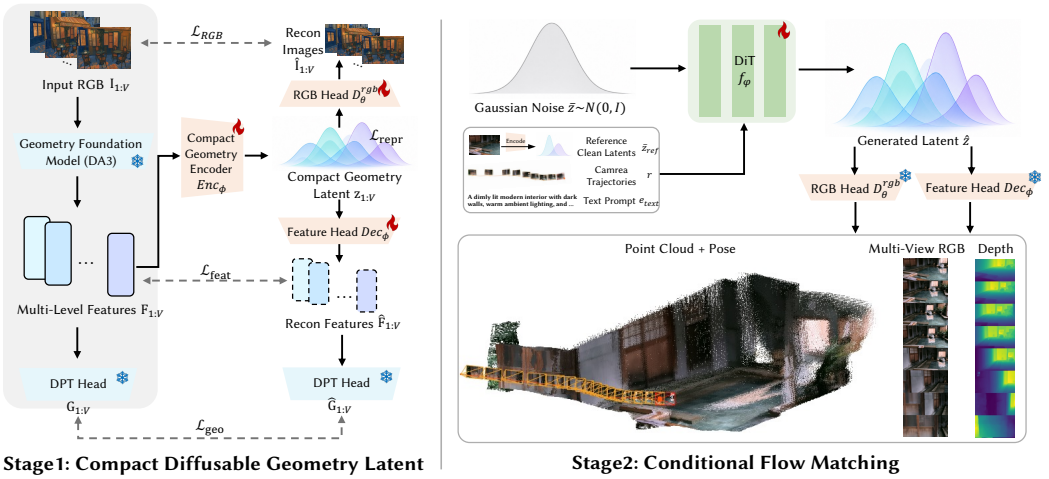

Figure 2: **Overview of Geometry-Native Autoencoder (GAE).** Stage 1 compresses frozen multi-level geometry features into a compact latent and reconstructs the hierarchy for frozen geometry readout and RGB decoding. Stage 2 applies conditional flow matching to generated-view latents, with clean reference latents, camera rays, and text as controls.

3.1 GEOMETRY-NATIVE CODEC

The DA3 hierarchy presents two unsatisfactory choices for generation. Modeling all four raw levels preserves the information needed for geometry decoding but requires a level-wise generative cascade; selecting one level permits a single flow model but discards complementary information and retains the poor conditioning identified in Sec. 4.1. GAE instead learns the state between a frozen DA3 encoder and its frozen geometry head: it fuses the hierarchy, compresses it into one spatial latent, and reconstructs all levels for joint RGB and geometry decoding.

From a feature hierarchy to one state. Given V input views, the frozen DA3 encoder $\mathcal{E}$ produces the four-level hierarchy $\mathbf{F}_v^{0:3} := \{\mathbf{f}_{v,\ell}\}_{\ell=0}^{3}$, where $\mathbf{f}_{v,\ell} \in \mathbb{R}^{N\times C_\ell}$ contains the patch tokens for view v at level ℓ and $N = h_p w_p$. The shallow L0 features preserve stronger cross-view correspondence (xLNC*), semantic organization (LNC@5), and spatial structure (LDS and SRSS), but have less smooth transport (higher ρ), while the deeper L3 features provide the complementary trade-off of lower ρ and weaker structure. This distinction motivates fusing the levels as complementary inputs to the frozen geometry decoder rather than selecting a single level (Tab. 2).

Before fusion, we normalize each level using per-channel training-set statistics $\mathbf{m}_\ell$ and $\mathbf{s}_\ell$. We then reshape the patch tokens to their spatial grids and concatenate them along the channel dimension:

$$\mathbf{X}_v = \mathcal{P}(\mathbf{F}_v^{0:3}) := \Big\|_{\ell=0}^{3} \mathrm{reshape}\left(\frac{\mathbf{f}_{v,\ell} - \mathbf{m}_\ell}{\mathbf{s}_\ell + \epsilon}\right) \in \mathbb{R}^{C_x\times h_p\times w_p}, \qquad C_x = \sum_{\ell=0}^{3} C_\ell, \tag{3}$$

where $\|$ denotes channel-wise concatenation. Thus, $\mathcal{P}$ is a fixed operator that converts the four DA3 token matrices into the single fused tensor $\mathbf{X}_v$. Level-wise normalization prevents high-variance channels from dominating this representation.

Compact geometry codec. The feature codec $\mathcal{A}_\phi := (\mathrm{Enc}_\phi, \mathrm{Dec}_\phi)$ compresses the fused tensor into a grid-shaped latent $\mathbf{z}_v \in \mathbb{R}^{C_z\times h_p\times w_p}$ and reconstructs the four-level hierarchy:

$$\hat{\mathbf{F}}_v^{0:3} = \mathrm{Dec}_\phi(\mathbf{z}_v), \qquad \mathbf{z}_v = \boldsymbol{\mu}_\phi(\mathbf{X}_v) + \boldsymbol{\sigma}_\phi(\mathbf{X}_v) \odot \boldsymbol{\epsilon}_v, \quad \boldsymbol{\epsilon}_v \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{4}$$

where $\boldsymbol{\mu}_\phi(\mathbf{X}_v)$ and $\boldsymbol{\sigma}_\phi(\mathbf{X}_v)$ are predicted by the encoder. GAE preserves the DA3 patch grid and compresses only along channels, yielding $C_z \in \{64, 128\}$ rather than the 3,072 channels of a single raw DA3 level. The encoder and decoder use lightweight convolutional pyramids with spatial self-attention, retaining the DA3 patch layout while allowing long-range mixing. We sample $\mathbf{z}_v$ during codec training, but use the posterior mean $\boldsymbol{\mu}_v$ deterministically for flow training and inference. A small KL weight regularizes the bottleneck without forcing it toward a heavily regularized generative-VAE regime.

Joint RGB and geometry decoding. In the notation above, Dec_ϕ includes reconstruction of the normalized fused tensor followed by the fixed inverse split and denormalization, and therefore re-

Table 2: **Latent-space diagnostics.** Single-/multi-view diffusability and structure metrics.

Latent	$C\times H\times W$	Diffusability (single-view) $\rho\downarrow$	$\kappa\downarrow$	effrank↑	Diffusability (multi-view) $\rho\downarrow$	$\kappa\downarrow$	effrank↑	Structure (single-view) LNC@5↑	LDS↑	SRSS↑	Structure (cross-view) xLNC* ↑
Pixel VAE, single-image (SD-VAE)	4×32×32	0.888	**3.7**	3.0	0.765	**4.1**	3.1	0.051	0.262	0.324	0.124
Pixel VAE, video (WAN2.1 VAE)	16×32×32	0.847	127	5.3	0.609	160	4.3	0.084	0.234	0.313	0.402
Semantic RAE (RAEV2)	1024×16×16	0.711	1.8×10^{4}	**289.3**	0.640	6.0×10^{4}	**93.3**	**0.932**	0.171	0.220	0.587
Raw geometry (DA3-GIANT L0)	3072×18×18	0.876	2.8×10^{8}	11.3	0.793	3.9×10^{8}	9.3	0.353	0.102	0.123	**0.642**
Raw geometry (DA3-GIANT L3)	3072×18×18	0.769	6.6×10^{16}	11.5	0.539	2.0×10^{11}	9.0	0.054	0.054	0.070	0.380
GLD (L0/L1 stage avg.)	–	0.854	9.5×10^{7}	34.8	0.742	1.7×10^{8}	22.7	0.274	0.103	0.131	0.594
Ours (GAE-128)	128×18×18	0.674	227	51.3	0.564	356	45.2	0.498	0.444	0.553	0.591
Ours (GAE-64)	64×18×18	**0.634**	53.1	37.0	**0.519**	117	33.2	0.488	**0.631**	**0.767**	0.569

turns $\hat{\mathbf{F}}_v^{0:3} = \{\hat{\mathbf{f}}_{v,\ell}\}_{\ell=0}^{3}$. A separate learned RGB head renders appearance directly from the compact latent, whereas the original frozen DA3 dense-prediction transformer (DPT) head reads geometry from the reconstructed hierarchy:

$$\hat{\mathbf{I}}_v = \mathcal{D}_\theta^{\mathrm{rgb}}(\mathbf{z}_v), \qquad \hat{\mathbf{G}}_v = \mathcal{H}_{\mathrm{DPT}}(\hat{\mathbf{F}}_v^{0:3}) = \mathcal{H}_{\mathrm{DPT}}(\mathrm{Dec}_\phi(\mathbf{z}_v)), \tag{5}$$

where $\hat{\mathbf{G}}_v$ contains depth, rays, and derived point maps. Consequently, RGB and geometry are two readouts of the same compact state. Keeping $\mathcal{H}_{\mathrm{DPT}}$ frozen prevents the geometry head from adapting to information lost by the codec: successful geometry reconstruction must remain readable by the original DA3 head. Unlike our raw single-layer baselines, GAE reconstructs the entire hierarchy directly and requires no backbone propagation between the generated state and geometry decoding.

Codec objective. We first require the bottleneck to reconstruct both the hierarchy and its downstream readouts. As shown in Eq. 6: $\mathcal{L}_{\mathrm{feat}}$ is a weighted level-wise feature reconstruction loss, $\mathcal{L}_{\mathrm{rgb}}$ combines pixel and perceptual reconstruction terms, and $\mathcal{L}_{\mathrm{geo}}$ supervises depth and ray outputs. The geometry targets are pseudo-targets obtained by applying the frozen DA3 head to the original features, rather than direct ground truth. $\mathcal{L}_{\mathrm{repr}}$ shapes how information is organized inside the bottleneck rather than what it reconstructs; the next section motivates this term.

$$\mathcal{L}_{\mathrm{codec}} = \mathcal{L}_{\mathrm{feat}} + \lambda_{\mathrm{kl}}\mathcal{L}_{\mathrm{kl}} + \mathcal{L}_{\mathrm{rgb}} + \mathcal{L}_{\mathrm{geo}} + \mathcal{L}_{\mathrm{repr}}. \tag{6}$$

3.2 Organizing the latent for generation

Reconstruction determines what the bottleneck preserves, but not how it is organized: a reconstruction-only codec may recover RGB and geometry while remaining difficult for a flow model. Tab. 1 summarizes the three-step progression; metric usage is summarized in Sec. 4.1, with full definitions in Appendix B and the complete metric breakdown in Appendix Tab. 8. The reconstruction baseline is compact and well-conditioned, but has weak transport smoothness (ρ) and semantic neighborhoods (LNC@5).

Table 1: **Latent shaping ablation.** Diffusability metrics are averaged over single- and multi-view protocols.

Representation obj.	Diffusability (view-averaged) $\bar{\rho}\downarrow$	$\bar{\kappa}\downarrow$	Structure (single-view) LNC@5↑	LDS↑	SRSS↑	Structure (cross-view) xLNC* ↑
Raw geometry (L0)	0.835	3.35×10^{8}	0.353	0.102	0.123	0.642
Raw geometry (L3)	0.654	3.3×10^{16}	0.054	0.054	0.070	0.380
None	0.725	355	0.310	0.240	0.300	0.590
$\mathcal{L}_{\mathrm{tok}}$	0.635	310	0.490	0.020	0.030	0.580
$\mathcal{L}_{\mathrm{tok}} + \mathcal{L}_{\mathrm{struct}}$	**0.619**	**292**	**0.498**	**0.444**	**0.553**	**0.591**

We first align each posterior-mean token $\boldsymbol{\mu}_{v,i}$ (location i, view v) with its co-located frozen C-RADIO feature (Heinrich et al., 2025). This improves transport smoothness (ρ) and semantic neighborhoods (LNC@5), but independent token alignment can destroy spatial similarities (LDS and SRSS), sharply reducing both and slightly weakening cross-view correspondence (xLNC*). A frozen DINOv2 teacher (Oquab et al., 2024) therefore supplies pairwise spatial similarities, which we match in the raw posterior space to restore neighborhood geometry.

With teacher tokens $\mathbf{c}_{v,i}$ and $\mathbf{d}_{v,i}$ bilinearly aligned to the latent grid, the objectives are:

$$\mathcal{L}_{\mathrm{tok}} = \mathbb{E}_{v,i}[1 - \langle \widehat{g_\eta(\boldsymbol{\mu}_{v,i})}, \hat{\mathbf{c}}_{v,i}\rangle], \qquad \mathcal{L}_{\mathrm{struct}} = \mathbb{E}_{v,i\neq j}[(\langle \hat{\boldsymbol{\mu}}_{v,i}, \hat{\boldsymbol{\mu}}_{v,j}\rangle - \langle \hat{\mathbf{d}}_{v,i}, \hat{\mathbf{d}}_{v,j}\rangle)^2]. \tag{7}$$

We combine them as $\mathcal{L}_{\mathrm{repr}} = \lambda_{\mathrm{repa}}(\mathcal{L}_{\mathrm{tok}} + \lambda_\mu \mathcal{L}_{\mathrm{struct}})$, where λ_{repa} and λ_μ weight the overall representation loss and the structural term.

Here g_η maps posterior tokens to the C-RADIO dimension and hats denote ℓ_2 normalization. C-RADIO organizes token semantics, while DINOv2 restores relational structure directly in the flow latent. Unlike REPA (Yu et al., 2025a), which supervises an intermediate denoiser representation, both terms shape the codec latent before generative training; the teachers and projector are then

Table 3: **Geometry reconstruction from encoded latents.** RealEstate10K and DL3DV results against Pi3 references.

	RealEstate10K (Pi3 ref)						DL3DV (Pi3 ref)					
Method	AbsRel↓	δ_1 ↑	Cham.↓	PMap↓	ATE↓	RPEr° ↓	AbsRel↓	δ_1 ↑	Cham.↓	PMap↓	ATE↓	RPEr° ↓
Pi3 on real (ref)	—	—	—	—	0.009	0.17	—	—	—	—	0.011	0.26
Gen3R	0.085	0.906	0.430	0.829	0.038	0.63	0.114	0.862	0.237	0.542	0.022	0.72
GLD	0.271	0.694	0.857	1.580	0.039	1.23	0.303	0.608	0.462	1.083	0.026	1.14
Raw geometry (DA3-GIANT L0)	0.178	0.759	0.635	1.160	0.068	1.33	0.234	0.639	0.575	1.130	0.057	1.09
Ours (GAE-128)	0.090	0.905	0.400	0.700	**0.006**	**0.19**	0.125	**0.862**	**0.145**	**0.433**	**0.008**	**0.25**
Ours (GAE-64)	**0.090**	**0.905**	**0.386**	**0.691**	0.007	0.20	**0.125**	0.861	0.145	0.433	0.008	0.27

discarded. After codec training, we freeze the codec, take the posterior mean $\mathbf{z}_v := \boldsymbol{\mu}_v$ as the latent, and standardize each channel using training-set statistics: $\bar{\mathbf{z}}_v = (\mathbf{z}_v - \mathbf{m}_z)/(\mathbf{s}_z + \epsilon)$.

3.3 Conditional flow matching and generation

Flow matching. After freezing the codec, we model only its standardized posterior means. Given target latents $\bar{\mathbf{z}}^{\mathrm{gt}}$ and noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, flow matching interpolates between them and trains a conditional transformer to predict the velocity; the path and objective are given in Appendix A.1, Eq. 11. Following RAEv2 (Singh et al., 2026), the network uses clean-latent prediction, converted to velocity. A single transformer jointly models all target views, allowing cross-view interaction directly in the compact geometry-native state. Stage 1 trains the codec and RGB decoder with the geometry backbone and DPT head frozen; Stage 2 freezes the codec and trains this flow model for text-to-image and view-conditioned generation. At inference, Gaussian target latents are integrated from $t = 1$ to 0, denormalized, and decoded into RGB and geometry through Eq. 2. Architecture, optimization, and solver details are provided in the supplementary material.

Reference, camera, and text conditioning. GAE encodes the K observed reference views using only those views as context and prepends the resulting clean latents $\bar{\mathbf{z}}^{\mathrm{ref}}$ to the noisy all-view tokens. All V output slots remain flow variables; reference tokens are assigned $t = 0$, participate in self-attention, and are excluded from the prediction head. Camera control uses metric Plücker-ray embeddings in query/key self-attention (Yin et al., 2026), while frozen language-model features enter through cross-attention. Independent condition dropout supports text-only, camera-controlled, reference-conditioned, and jointly conditioned generation with the same flow model. Detailed token construction and the conditioning ablations are provided in the supplementary material.

4 Experiments

Experimental setup and ranking convention. The controlled comparison includes two pixel codecs (the single-image SD-VAE (Rombach et al., 2022) and the video WAN2.1 VAE), the official semantic representation autoencoder RAEV2 (Singh et al., 2026) with its frozen DINOv3-L encoder (Siméoni et al., 2025), the raw DA3-GIANT layers L0 and L3 (Lin et al., 2026), and our codecs GAE-64 and GAE-128, which compress the full DA3 hierarchy and retain its native geometry readout. We set: $\lambda_{\mathrm{repa}} = 0.25$ and $\lambda_\mu = 8.0$. All variants use the same flow-model family, camera conditioning, training budget, and sampling protocol on RealEstate10K (Zhou et al., 2018) and DL3DV (Ling et al., 2024), and are evaluated on 64 held-out scenes with nine 252^2 views, one reference, 50 Euler steps, and CFG $= 2$. The final-model showcase in Sec. 4.3.3 uses a separate qualitative protocol: 81 views at 672×378, one reference view, 50 Euler steps, and CFG $= 2$. These long-rollout and text-conditioned examples are qualitative and are not part of the comparison. For ranking, blue cells mark the best and second-best controlled variants; displayed ties use unrounded values. Gray rows are external complete pipelines excluded from the latent ranking: GLD (Jang et al., 2026) generates a DA3 feature cascade, whereas Gen3R (Huang et al., 2026) is a pretrained geometry-generation system.

4.1 Latent-space study

We measure transport complexity with ρ (Zhong et al., 2026), spectral conditioning with κ and effective rank, semantic neighborhoods with LNC@5 (Zhong et al., 2026), spatial similarities with LDS (Singh et al., 2026; 2025) and SRSS (Zhong et al., 2026; Singh et al., 2025), and cross-view retrieval with xLNC*. Full definitions are in the Appendix. Tab. 2 reports the raw diagnostics; an

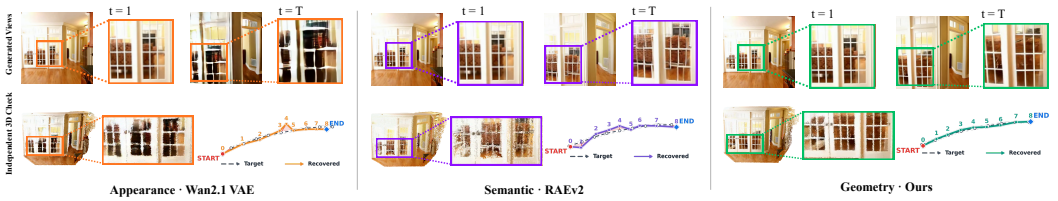

Figure 3: **Independent 3D check across latent families.** Generated views, detail crops, reconstructed point clouds, and recovered camera trajectories under matched conditioning.

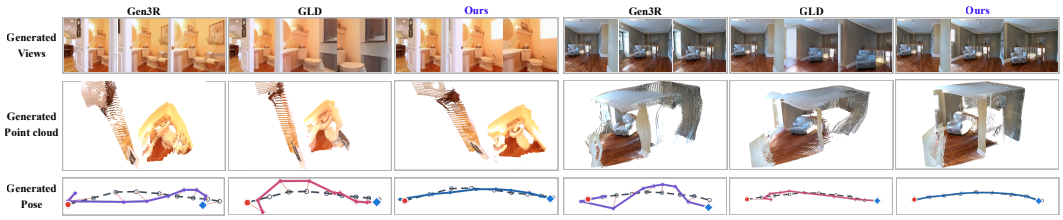

Figure 4: **Multi-frame RGB, geometry, and camera-pose comparison.** Generated views, point clouds, and recovered poses under matched conditioning.

overview of intrinsic representation quality and matched DiT generation is provided in Appendix Fig. 6. No baseline is strong on every axis. Pixel VAEs are well-conditioned but weakly structured, semantic RAEv2 has strong semantic neighborhoods but no native geometry readout, and raw DA3 features are geometry-native but high-dimensional and poorly conditioned. Within DA3, L0 retains the strongest cross-view correspondence, whereas L3 lowers ρ at the cost of spatial and cross-view structure. Both use 3,072 channels but have an effective rank of only about 10. By fusing the hierarchy, GAE reduces the state to 64 or 128 channels and provides the best overall balance of transport, conditioning, and structure among the geometry-native representations. GLD is reported only as an L0/L1 stage-average proxy because it models the two levels with a cascade.

4.2 Codec Reconstruction Quality

4.2.1 RGB Reconstruction

We report PSNR and LPIPS for paired fidelity and rFID/rFVD for the distributional gap between reconstructed and real images/videos. Higher PSNR and lower LPIPS, rFID, and rFVD are better. A compact latent risks discarding the fine-grained spatial detail retained by raw geometry features. We therefore evaluate codec reconstruction independently of downstream generation. Tab. 4 measures RGB fidelity. A separately trained RGB head on raw DA3 L0 reconstructs appearance well, but the same 3072-channel feature is ill-conditioned and nearly rank-degenerate (Tab. 2). Raw L3 loses substantially more reconstructable appearance information. GAE-128 matches or slightly exceeds raw L0 across PSNR, LPIPS, rFID, and rFVD while using 24 times fewer channels.

Table 4: **RGB reconstruction.** ImageNet for Single-view; multi-view with five video datasets.

Latent	Single-view			Multi-view		
	PSNR↑	LPIPS↓	rFID↓	PSNR↑	LPIPS↓	rFVD↓
Pixel VAE, single-image (SD-VAE)	24.44	0.141	26.9	28.97	0.112	40.5
Pixel VAE, video (WAN2.1 VAE)	27.49	0.125	21.8	28.12	0.156	50.6
Semantic RAE (RAEV2)	22.08	0.134	8.8	24.73	0.094	32.1
Raw geometry (DA3-GIANT L0)	28.64	0.038	5.5	34.70	**0.014**	4.4
Raw geometry (DA3-GIANT L3)	19.90	0.219	25.3	23.98	0.122	76.2
Ours (GAE-128)	**28.76**	**0.036**	**5.4**	**34.78**	**0.014**	**4.3**
Ours (GAE-64)	27.30	0.061	7.1	32.37	0.040	9.1

4.2.2 Geometry Reconstruction

Tab. 3 evaluates geometry after encoding and decoding real frames. AbsRel and δ_1 measure relative depth error and threshold accuracy; Chamfer and PMap measure point-set and point-map discrepancy; ATE and RPEr measure trajectory and relative-rotation error. The raw-L0 baseline (DA3-GIANT L0) propagates L0 through the frozen DA3 backbone to recover the remaining hierarchy before the DPT head. Both GAE variants improve on this baseline across depth, point-map, and pose metrics on both datasets. Thus, the compact codec preserves geometry-relevant information in addition to the RGB fidelity measured above. Compared with the external systems, GAE gives lower camera errors than Gen3R (Huang et al., 2026) and GLD (Jang et al., 2026) and lower point-

Table 5: **Camera-conditioned appearance and 3D consistency.** Selected FVD, PSNR, ATE, Reproj., and MEt3R metrics; complete metrics are in the supplement.

Latent	RealEstate10K					DL3DV				
	FVD↓	PSNR↑	ATE↓	Reproj.↓	MEt3R↓	FVD↓	PSNR↑	ATE↓	Reproj.↓	MEt3R↓
Pixel VAE, single-image (SD-VAE)	258.6	19.32	0.0072	0.0093	0.1385	373.2	17.05	0.0073	0.0108	0.1542
Pixel VAE, video (WAN2.1 VAE)	362.9	16.52	0.0225	0.0105	0.1441	596.7	14.31	0.0243	0.0258	0.1777
Semantic RAE (RAEV2)	379.4	17.60	0.0085	0.0104	0.1263	453.4	16.01	0.0104	0.0130	0.1549
Raw geometry (DA3-GIANT L0)	298.6	18.31	0.0085	0.0096	0.1420	376.5	16.68	0.0110	0.0129	0.1721
Raw geometry (DA3-GIANT L3)	488.9	16.82	0.0112	0.0124	0.1575	584.8	16.22	0.0097	0.0151	0.1754
Ours (GAE-128)	233.4	19.54	0.0041	0.0075	0.1213	345.2	17.39	0.0056	0.0080	**0.1347**
Ours (GAE-64)	**225.7**	**20.02**	**0.0034**	**0.0068**	**0.1208**	**287.0**	**18.00**	**0.0056**	**0.0077**	0.1370
GLD	445.1	14.91	0.0124	0.0133	0.1660	587.4	14.51	0.0136	0.0241	0.2183
Gen3R	269.7	17.79	0.0090	0.0072	0.1157	580.5	15.62	0.0096	0.0143	0.1662

Table 6: **Geometry generation from sampled latents.** RealEstate10K and DL3DV results against Pi3 references.

Method	RealEstate10K (Pi3 ref)						DL3DV (Pi3 ref)					
	AbsRel↓	δ_1 ↑	Cham.↓	PMap↓	ATE↓	RPEr$^\circ$ ↓	AbsRel↓	δ_1 ↑	Cham.↓	PMap↓	ATE↓	RPEr$^\circ$ ↓
Pi3 on real (ref)	—	—	—	—	0.009	0.17	—	—	—	—	0.011	0.26
Gen3R	0.164	0.807	0.727	1.428	0.046	0.69	0.232	0.688	0.587	1.180	0.031	0.91
GLD	0.271	0.694	0.858	1.580	0.039	1.23	0.323	0.594	0.462	1.097	0.028	1.19
Raw geometry (DA3-GIANT L0)	0.157	0.822	0.540	1.058	0.019	0.35	0.245	0.715	0.330	0.815	0.019	0.48
Ours (GAE-128)	0.139	0.849	**0.432**	**0.813**	0.013	0.37	0.225	0.750	0.325	0.832	0.016	0.40
Ours (GAE-64)	**0.134**	**0.857**	0.501	0.924	**0.010**	**0.33**	**0.204**	**0.765**	**0.262**	**0.698**	**0.014**	**0.36**

cloud errors than GLD; Gen3R is slightly better on some depth metrics but uses a separate geometry latent. Pixel codecs and RAEV2 are omitted because they have no native geometry readout.

4.3 Flow-matching generation quality

We compare all latent families under the same flow-matching backbone, training budget, and camera-conditioning mechanism. This controlled setup isolates the impact of the generated latent representation on both appearance quality and 3D consistency.

4.3.1 RGB generation and 3D consistency

Tab. 5 reports the selected FVD and PSNR results; Complete FID, LPIPS, and SSIM results are given in Appendix Tab. 9. Among the controlled variants, GAE-64 achieves the lowest FVD and highest PSNR on both datasets, reducing FVD by 12.7% on RealEstate10K and 23.1% on DL3DV relative to the best non-GAE latent. Additional visualization examples are provided in Appendix Fig. 7. Under identical reference and camera trajectories, GAE-64 better preserves object identity and fine structure as viewpoints move away from the reference, consistent with Tab. 5.

For 3D consistency, Tab. 5 reports camera and multi-view consistency using independent VGGT (Wang et al., 2025a) and MEt3R (Asim et al., 2025). ATE measures global trajectory error after Sim(3) alignment, Reproj. measures normalized round-trip error for high-confidence points, and MEt3R measures pairwise feature disagreement across views. The two GAE variants occupy the top two controlled positions on the VGGT metrics. GAE-64 reduces ATE by 52.8% on RealEstate10K and 23.3% on DL3DV; it also gives the best controlled MEt3R on RealEstate10K (0.1208), while GAE-128 is best on DL3DV (0.1347). The DA3-GIANT backbone-overlap cross-check is reported in Appendix Tab. 10. Fig. 3 further shows that GAE yields a more coherent 3D scene and more faithful camera motion under the same reference view and prescribed trajectory.

4.3.2 Geometry generation

We separately evaluate the geometry decoded from sampled latents against Pi3 reconstructed from the real target frames at the same viewpoints. This protocol measures native depth, point-map, and camera quality without scoring each method against its own generated frames. AbsRel and δ_1 measure depth agreement, Chamfer and PMap measure point-map agreement, and ATE and RPEr

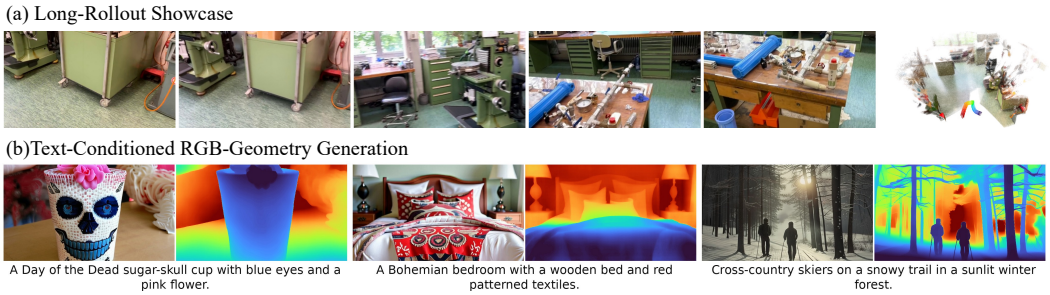

Figure 5: **Additional qualitative capabilities.**

measure camera-trajectory error. Tab. 6 shows that the geometry-native latents consistently improve over raw DA3-GIANT features. On RealEstate10K, GAE-64 gives the best depth and pose results, while GAE-128 leads on Chamfer and PMap; on DL3DV, GAE-64 is best on all six metrics, which indicates that the latent shaping preserves the geometric structure needed for both depth and camera generation. Jointly decoded RGB, geometry, and camera trajectories are shown in Fig. 4.

4.3.3 Additional qualitative capabilities

The same flow model also supports long rollouts and text-conditioned RGB–geometry generation. Fig. 5 shows one representative row from each capability: an 81-view rollout at 672×378 with its decoded point cloud and a set of text-conditioned RGB/depth pairs. The full rollouts, additional scenes, and prompts are provided in the supplementary material.

4.4 Ablation studies

The latent diagnostics in Tab. 1 show that the complete representation objective improves transportability and semantic organization while preserving spatial and cross-view structure. Tab. 7 examines whether these gains transfer to downstream generation and separately evaluates the effect of preserving metric camera scale. The first block ablates the codec backbone and representation objective. On DA3-GIANT, $\mathcal{L}_{\text{tok}}$ substantially improves FVD and PSNR over the unregularized codec. Adding $\mathcal{L}_{\text{struct}}$ to form the complete $\mathcal{L}_{\text{repr}}$ further improves FVD, PSNR, and VGGT ATE while maintaining MEt3R. The complete objective outperforms the unregularized DA3-GIANT codec across all four metrics. With the same latent width and objective, DA3-GIANT improves appearance quality and camera accuracy over DA3-LARGE, whereas DA3-LARGE retains a small MEt3R advantage, which indicates that representation shaping and backbone capacity provide complementary gains. The second block compares per-scene-normalized poses with metric-pose conditioning following SCoPE (Yin et al., 2026). Retaining metric translation scale reduces the scale-sensitive SE(3) ATE from 0.0346 to 0.0093 and improves FVD, PSNR, and MEt3R. Thus, metrically grounded camera conditioning improves both trajectory fidelity and overall generation quality.

Table 7: **Ablation study of codec and pose scale.** Diag. denotes the block-specific metric per header.

Setting	FVD ↓	PSNR ↑	MEt3R ↓	Diag. ↓
Codec design (Diag. = VGGT ATE)				
LARGE-128, with $\mathcal{L}_{\text{repr}}$	270.5	19.12	**0.1192**	0.0048
GIANT-128, no $\mathcal{L}_{\text{repr}}$	266.4	19.19	0.1226	0.0044
GIANT-128, with $\mathcal{L}_{\text{tok}}$	241.9	19.40	0.1212	0.0042
GIANT-128, with $\mathcal{L}_{\text{repr}}$	**233.4**	**19.54**	0.1213	**0.0041**
Pose scale (Diag. = SE(3) ATE, no Sim(3))				
Per-scene normalized pose	350.7	16.28	0.1290	0.0346
Metric pose	**225.7**	**20.02**	**0.1208**	**0.0093**

5 Conclusion

We presented GAE, a compact reparameterization of multi-level geometry features that is jointly decodable to RGB and geometry and organized for flow modeling. Under a controlled generator and training protocol, this latent improves reconstruction, visual generation, 3D consistency, and camera control over pixel, semantic, and raw geometry representations. Additional larger-scale experiments further confirm that these gains persist with more data, higher resolution, and longer generation sequences. Together, these results show that geometry can define the state in which generation occurs, providing a shared interface between perception and generation.

AI USE STATEMENT

In preparing this submission package, we used a generative AI assistant to organize the LaTeX source, remove identifying metadata, correct a typographical error, and verify compilation. We did not use generative AI for the research claims, mathematical derivations, data generation, method design, experiment design, implementation, or interpretation reported in this paper. The authors reviewed all AI-assisted changes and take responsibility for the final content of this work.

ETHICS STATEMENT

This work does not collect new human-subject data or conduct user studies. We use publicly released benchmark data and pretrained models under their respective licenses and terms. The method generates images, video, and scene geometry, which can be repurposed for synthetic media; we present it for research evaluation, do not target real individuals, and do not report personal identifiers or sensitive attributes. We follow the ICLR Code of Ethics in the design, evaluation, and reporting of this work.

REPRODUCIBILITY STATEMENT

The main paper specifies the geometry-native codec, flow model, conditioning, and evaluation design in Sections 3–4. Appendix A gives the flow objective, conditioning implementation, training and sampling protocols; Appendix B defines the latent diagnostics and metrics; and Appendix C provides complete quantitative tables and additional qualitative results. The experiments use publicly released benchmark data and pretrained geometry models, with the matched settings and evaluation procedures described in the paper and appendix.

REFERENCES

Zhaochong An, Orest Kupyn, Théo Uscidda, Andrea Colaco, Karan Ahuja, Serge Belongie, Mar Gonzalez-Franco, and Marta Tintore Gazulla. VGGRPO: Towards world-consistent video generation with 4d latent reward. In *ECCV*, 2026.

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. MEt3R: Measuring multi-view consistency in generated images. In *CVPR*, 2025.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021.

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. In *NeurIPS*, 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.

Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):7457–7476, 2022.

Vitor Guizilini, Muhammad Zubair Irshad, Dian Chen, Greg Shakhnarovich, and Rares Ambrus. Zero-shot novel view and depth synthesis with multi-view geometric diffusion. In *CVPR*, 2025.

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 *ICLR*, 2025.

Greg Heinrich, Mike Ranzinger, Hongxu Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. RADIOv2.5: Improved baselines for agglomerative vision foundation models. In *CVPR*, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jiaxin Huang, Yuanbo Yang, Bangbang Yang, Lin Ma, Yuewen Ma, and Yiyi Liao. Gen3R: 3d scene generation meets feed-forward reconstruction. In *CVPR*, 2026.

Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2821–2830. IEEE, 2018.

Wooseok Jang, Seonghu Jeon, Jisang Han, Jinhyeok Choi, Minkyung Kwon, Seungryong Kim, Saining Xie, and Sainan Liu. Repurposing geometric foundation models for multi-view diffusion. In *ECCV*, 2026.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 2023.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

Akshay Krishnan, Xinchen Yan, Vincent Casser, and Abhijit Kundu. Orchid: Image latent diffusion for joint appearance and geometry generation. In *ICCV*, 2025.

Byung-Ki Kwon, Qi Dai, Hyoseok Lee, Chong Luo, and Tae-Hyun Oh. JointDiT: Enhancing rgb-depth joint modeling with diffusion transformers. In *ICCV*, 2025.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with MASt3R. In *ECCV*, 2024.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *ICLR*, 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 *CVPR*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023.

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 *ICCV*, 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 *ICLR*, 2024.

Jiahao Lu, Tianyu Huang, Peng Li, Zhiyang Dou, Cheng Lin, Zhiming Cui, Zhen Dong, Sai-Kit Yeung, Wenping Wang, and Yuan Liu. Align3r: Aligned monocular depth estimation for dynamic videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22820–22830. IEEE, 2025.

Zhenxing Mi, Yuxin Wang, and Dan Xu. One4d: unified 4d generation and reconstruction via decoupled lora control. In *European Conference on Computer Vision*, pp. 38–56. Springer, 2026.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

François Porcher, Nicolas Carion, Karteek Alahari, and Shizhe Chen. Flow matching in feature space for stochastic world modeling. *arXiv preprint arXiv:2606.29059*, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. MVDream: Multi-view diffusion for 3d generation. In *ICLR*, 2024.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure? In *The Fourteenth International Conference on Learning Representations*, 2025.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. *arXiv preprint arXiv:2605.18324*, 2026.

Gabriela Ben Melech Stan, Diana Wofk, Scottie Fox, Alex Redden, Will Saxton, Jean Yu, Estelle Aflalo, Shao-Yen Tseng, Fabio Nonato, Matthias Müller, and Vasudev Lal. LDM3D: Latent diffusion model for 3d. *arXiv preprint arXiv:2305.10853*, 2023.

Xiangyu Sun, Shijie Wang, Fengyi Zhang, Lin Liu, Caiyan Jia, Ziying Song, Zi Huang, and Yadan Luo. VGGT-World: Transforming vggt into an autoregressive geometry world model. In *ECCV*, 2026.

Shengbang Tong, Boyang Zheng, Ziteng Wang, Bingda Tang, Nanye Ma, Ellis Brown, Jihan Yang, Rob Fergus, Yann LeCun, and Saining Xie. Scaling text-to-image diffusion transformers with representation autoencoders. In *CVPR Findings*, 2026.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Yuhao Wan, Lijuan Liu, Jingzhi Zhou, Zihan Zhou, Xuying Zhang, Dongbo Zhang, Shaohui Jiao, Qibin Hou, and Ming-Ming Cheng. GeoWorld: Providing full-frame geometry features to facilitate 3d scene generation. In *ECCV*, 2026.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Chang Zeng, Yanxi Zhou, Xiao-Xiao Long, et al. Spatialvid: A large-scale video dataset with spatial annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42592–42603, 2026a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *CVPR*, 2026b.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3d vision made easy. In *CVPR*, 2024a.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *ICLR*, 2026c.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A unified and flexible motion controller for video generation. In *SIGGRAPH*, 2024b.

Dongxu Wei, Qi Xu, Zhiqi Li, Hangning Zhou, Cong Qiu, Hailong Qin, Mu Yang, Zhaopeng Cui, and Peidong Liu. Any 3d scene is worth 1k tokens: 3d-grounded representation for scene generation at scale. *arXiv preprint arXiv:2604.11331*, 2026.

Lisa Weijler, Irene Ballester, Guofeng Mei, Tolga Birdal, and Pedro Hermosilla. Latent riemannian flow matching for geometry-grounded 3d foundation models. *arXiv preprint arXiv:2607.19120*, 2026.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. In *ICLR*, 2026.

Xunzhi Xiang, Zixuan Duan, Yabo Chen, Zhengxuan Wei, Guiyu Zhang, Zixiao Gu, Zhe Gao, Haibin Huang, Chi Zhang, Qi Fan, and Xuelong Li. VideoWeave: Unlocking geometric consistency in video generation via joint geometry-video modeling. *arXiv preprint arXiv:2606.14162*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12–22. IEEE, 2023.

Minghao Yin, Jiahao Lu, Wenbo Hu, Wang Zhao, Ying Shan, and Kai Han. SCoPE: Sightline-coordinate positional encoding for video diffusion transformers. *arXiv preprint arXiv:2606.27345*, 2026.

Jason J. Yu, Tristan Aumentado-Armstrong, Fereshteh Forghani, Konstantinos G. Derpanis, and Marcus A. Brubaker. PolyOculus: Simultaneous multi-view image-based novel view synthesis. In *ECCV*, 2024.

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 *ICLR*, 2025a.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. ViewCrafter: Taming video diffusion models for high-fidelity novel view synthesis. *IEEE TPAMI*, 2025b.

Yuchen Zhang, Nikhil Keetha, Chenwei Lyu, Bhuvan Jhamb, Yutian Chen, Yuheng Qiu, Jay Karhade, Shreyas Jha, Yaoyu Hu, Deva Ramanan, et al. Ufm: A simple path towards unified dense correspondence with flow. *arXiv preprint arXiv:2506.09278*, 2025.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *ICLR*, 2026.

Tianxiong Zhong, Xingye Tian, Xuebo Wang, Xin Tao, and Pengfei Wan. Diffusing in the right space: A systematic study of latent diffusability. *arXiv preprint arXiv:2606.03578*, 2026.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *ICML*, 2025a.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM TOG*, 2018.

Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, et al. Omniworld: A multi-domain and multi-modal dataset for 4d world modeling. *arXiv preprint arXiv:2509.12201*, 2025b.

A Implementation and evaluation details

A.1 Flow-matching parameterization

After freezing the codec, we train only on standardized posterior means. Let $\bar{\mathbf{z}}^{\mathrm{gt}}_{1:V}$ denote the target latent and $\bar{\mathbf{z}}^{(1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ Gaussian noise. We interpolate between them using

$$\bar{\mathbf{z}}_t = (1-t)\bar{\mathbf{z}}^{\mathrm{gt}} + t\bar{\mathbf{z}}^{(1)}, \qquad t \in [0,1], \tag{8}$$

whose target velocity is $\mathbf{u}_t = \bar{\mathbf{z}}^{(1)} - \bar{\mathbf{z}}^{\mathrm{gt}}$. Following RAEv2 (Singh et al., 2026), the network predicts the clean latent,

$$\hat{\bar{\mathbf{z}}}^{\mathrm{gt}} = \mathbf{f}_\psi(\bar{\mathbf{z}}_t, t, \mathcal{C}), \tag{9}$$

which is converted to velocity by

$$\hat{\mathbf{u}}_t = \frac{\bar{\mathbf{z}}_t - \hat{\bar{\mathbf{z}}}^{\mathrm{gt}}}{\max(t, t_\epsilon)}, \qquad t_\epsilon = 0.05. \tag{10}$$

We optimize the flow objective

$$\mathcal{L}_{\mathrm{flow}} = \mathbb{E}\Big[\rho(t)\,\big\|\hat{\mathbf{u}}_\psi(\bar{\mathbf{z}}_t, t, \mathcal{C}) - (\boldsymbol{\epsilon} - \bar{\mathbf{z}}^{\mathrm{gt}})\big\|_2^2\Big], \tag{11}$$

with an optional time weight $\rho(t)$. The timestep is shared across all views in a clip. We use a DDT-style transformer (Singh et al., 2026) with 1×1 latent tokens and cross-view self-attention; temporal models also use frame-axis rotary embeddings. The controlled models use the same encoder–decoder transformer recipe: 28 encoder blocks of width 768 and six decoder blocks of width 2048. Only the latent input/output projections change when the codec width changes from 128 to 64 channels.

A.2 Inference-aligned multimodal conditioning

The conditioning design follows a single principle: every signal is constructed from information that is genuinely available at inference time and is kept separate from the variables evolved by the flow. In particular, only the V requested output latents form the noisy ODE state. Observed images are converted once into clean geometry-native evidence tokens, camera parameters specify a spatial field over both observed and requested views, and text supplies global semantic context. We denote these signals by $\mathcal{C} = \{\bar{\mathbf{z}}^{\mathrm{ref}}, \mathbf{r}, \mathbf{e}_{\mathrm{text}}\}$. They are injected through complementary attention pathways and remain fixed throughout sampling, preventing the conditioning evidence from being corrupted or implicitly reconstructed by the ODE.

Reference tokens. Geometry encoders are set-based, so encoding a reference jointly with unavailable target views would leak target-set context during training and create a train–test mismatch. We instead encode the K observed references jointly, but using only those references as the encoder context available at inference:

$$\bar{\mathbf{z}}^{\mathrm{ref}} = \mathrm{Std}\Big[\mathrm{Enc}^{\mu}_{\phi}\big[\mathcal{P}\big(\mathcal{E}(\mathbf{I}^{\mathrm{ref}}_{1:K})\big)\big]\Big]. \tag{12}$$

The resulting tokens preserve the same geometry-native representation used for the generated views, allowing evidence and state tokens to interact directly in self-attention. They are assigned $t = 0$, prepended to the noisy all-view tokens, and excluded from the prediction head. Consequently, their role is purely contextual: they can be attended to at every transformer layer but are never updated by the numerical solver. Every output slot—including a requested output at a reference camera—remains an ordinary noisy flow variable. This separation avoids repeatedly clamping a partially denoised reference slot and keeps the same ODE parameterization for conditional and unconditional sampling.

Camera rays. For each latent cell we form a world-space Plücker ray $(\mathbf{d}, \mathbf{m})$, where $\mathbf{m} = \mathbf{o} \times \mathbf{d}$, and decompose it as

$$\mathbf{d}, \qquad \hat{\mathbf{m}} = \frac{\mathbf{m}}{\|\mathbf{m}\| + \epsilon}, \qquad s = \log(\|\mathbf{m}\| + \epsilon). \tag{13}$$

Here $\mathbf{d}$ describes viewing direction, $\hat{\mathbf{m}}$ captures the orientation of the ray moment, and the log-scale channel s retains metric translation magnitude when metric camera sidecars are available. Following SCoPE (Yin et al., 2026), the resulting embedding modulates query/key self-attention, making pairwise token interaction explicitly dependent on the corresponding camera rays rather than treating pose as a single global vector. Rays are evaluated at the DiT (Peebles & Xie, 2023) token grid, preserving pixel–ray alignment when the latent patch size changes. Reference tokens receive the rays of their observed cameras, while noisy target tokens receive rays from the requested output cameras; the same spatial representation therefore relates evidence to the views being generated.

Text. Prompts are encoded by a frozen language model and projected to the transformer cross-attention width. Unlike reference and camera signals, which are aligned with latent locations and enter self-attention, text acts as global semantic context through cross-attention. Independent dropout of text, reference, and camera conditions trains a single set of weights on the subsets of $\mathcal{C}$ encountered at inference. The same model can consequently operate in text-only, camera-controlled, reference-conditioned, or jointly conditioned image and video generation regimes, and the dropped-condition branches also provide the unconditional predictions used for classifier-free guidance (Ho & Salimans, 2022).

A.3 Two-stage training, sampling, and evaluation protocols

We distinguish two training configurations: a controlled setting for comparing latent representations and a separate setting for the final model used in qualitative demonstrations. Evaluation covers three complementary aspects: codec reconstruction, conditional generation, and native geometry readout.

Two-stage training and shared sampling procedure. Stage 1 jointly trains $\mathcal{A}_\phi$ and $\mathcal{D}_\theta^{\mathrm{rgb}}$ on mixed single- and multi-view data, with $\mathcal{E}$ and $\mathcal{H}_{\mathrm{DPT}}$ frozen. Stage 2 freezes all Stage 1 modules and trains only the conditional flow model. Both configurations use the same sampling procedure: reference images are encoded once as clean conditions, while all requested output slots are initialized with Gaussian noise. We jointly integrate their velocity field from $t = 1$ to $t = 0$, then denormalize and decode the resulting latents through Eq. 2. Quantitative comparisons and qualitative demonstrations both use 50 Euler steps and a CFG scale of 2. For each controlled backend, the latent normalization statistics used during training are also applied at inference.

Controlled comparison: architecture and training data. To isolate the effect of latent representation, we place all compared backends within the same DDT-style flow architecture and match their training and sampling protocols. The shared transformer comprises 28 encoder blocks of width 768 and six decoder blocks of width 2048. It predicts clean latents, uses an eight-block auxiliary base head, and incorporates frozen Qwen3-0.6B (Yang et al., 2025) text features through cross-attention. The shared flow backbone contains approximately 0.93B parameters; total parameter counts differ slightly because each latent shape requires its own input and output projections.

Training clips contain nine 252×252 views drawn from an equal mixture of RealEstate10K (Zhou et al., 2018) and DL3DV (Ling et al., 2024), with one to four prefix views serving as references. Each model is trained on eight GPUs. The controlled comparison therefore holds the flow backbone, data mixture, optimization schedule, and sampling protocol fixed while varying the latent encoder, decoder, and the projections required to interface with them. Architecture- or conditioning-specific modifications are introduced only in the corresponding ablations.

Controlled comparison: optimization. We train in BF16 using AdamW with $\beta = (0.9, 0.95)$, zero weight decay, gradient clipping at 1.0, and an EMA decay of 0.9995. The learning rate increases to 10^{-4} over 2k warmup steps and then follows a cosine schedule toward 10^{-6} over a 100k-step horizon. Text, camera, and reference conditions are independently dropped with probabilities 0.10, 0.05, and 0.05, respectively. A single-view text-to-image batch is interleaved every three multi-view updates.

Codec evaluation. We first evaluate the latent representations through direct codec readout, independently of flow-model generation. The latent-space study in Section 4.1 uses ImageNet-S (Gao et al., 2022) for single-view diagnostics. Its multi-view diagnostics use video clips from five sources: RealEstate10K (Zhou et al., 2018), DL3DV (Ling et al., 2024), MVS-Synth (Huang et al., 2018), ScanNet++ (Yeshwanth et al., 2023), and Open-Sora Plan (OSP) (Lin et al., 2024). The RGB reconstruction study in Section 4.2 uses ImageNet for single-view reconstruction and reports the average multi-view reconstruction performance over the same five video sources. These codec evaluations are separate from the held-out scene pools used for the generation comparison below.

Generation evaluation and ranking. The main generation comparison uses shared held-out pools of 64 scenes from each of RealEstate10K and DL3DV, totaling 128 scenes. Each example contains nine views at 252×252, with one clean reference view and a prescribed camera trajectory. The frame interval is 10 for RealEstate10K and 1 for DL3DV. All backends use these same examples and the shared sampling procedure described above.

SD-VAE (Rombach et al., 2022), WAN2.1 VAE (Wan et al., 2025), and RAEV2 (Singh et al., 2026) produce 256×256 images in this setting. We resize their generated outputs to 252×252 before evaluation to match the other backends and ground-truth clips. FID and FVD measure frame- and clip-level distributional quality, respectively, whereas LPIPS, PSNR, and SSIM compare generated frames with their paired targets along the prescribed trajectory. Blue cells rank only the seven latent variants trained with the shared flow recipe. GLD and Gen3R are reported as external system-level references and are excluded from this controlled ranking.

Camera accuracy and cross-view consistency. To evaluate camera accuracy, we apply VGGT or DA3-GIANT to each complete generated nine-view clip and compare the recovered camera trajectory with the ground-truth trajectory provided by the original dataset. We align the recovered trajectory to the ground truth using Umeyama Sim(3), then compute absolute trajectory error (ATE), relative translation error (RPEt), and relative rotation error (RPEr). VGGT is independent of all evaluated latent backbones and serves as the independent evaluator; DA3-GIANT provides an additional cross-check with backbone overlap.

For reprojection evaluation (Jang et al., 2026), recovered world points are aligned to the dataset camera frame using Sim(3). High-confidence points are transferred between views and projected back to the source view. The resulting round-trip 2D pixel error is normalized by the image diagonal. We report Reproj@50, retaining the top half of points by estimator confidence.

We compute MEt3R using its official implementation (Asim et al., 2025). For each pair of generated views, it recovers pairwise geometry, upsamples visual features with FeatUp, and renders the features into the opposite view using visibility and z-buffer occlusion masks. The score measures symmetric cosine-feature disagreement and is averaged over all $\binom{V}{2}$ view pairs. Both Reproj and MEt3R are lower-is-better metrics.

Native geometry evaluation. Complementing the RGB-based evaluation above, we directly assess the geometry decoded from the latent representation. Decoded depth and point maps are compared with reference predictions obtained by applying Pi3 (Wang et al., 2026c) to the real target frames. Depth is aligned using a per-view median scale, while point clouds use the Sim(3) transformation estimated from the camera trajectories. Native camera ATE and RPEr are evaluated against the ground-truth dataset trajectory. This protocol evaluates the model's own geometry readout, separately from the post-hoc reconstruction of generated RGB frames used for camera and consistency evaluation.

Final model for qualitative demonstrations. The final model is trained separately for longer, higher-resolution generation and is excluded from the controlled ranking. It uses a frozen space–time codec built on DA3-GIANT features at 672×378 and trains the flow model on 81-view clips with one to four prefix reference views. Its multi-domain training mixture includes RealEstate10K (Zhou et al., 2018), DL3DV (Ling et al., 2024), ScanNet++ (Yeshwanth et al., 2023), MVS-Synth (Huang et al., 2018), SpatialVID (Wang et al., 2026a), OmniWorld (Zhou et al., 2025b), training data curated and released by Open-Sora Plan (Lin et al., 2024), and several internally curated datasets.

We train the 0.93B-parameter flow transformer on 80 GPUs in BF16 using AdamW, with 500 warmup steps to a base learning rate of 5×10^{-5}, followed by cosine decay to 10^{-6} by 60k steps. Gradient clipping is set to 1.0 and EMA decay to 0.9995. The parameter count refers only to the trainable Stage 2 generator and excludes the frozen DA3 encoder, space–time codec, and text encoder. Text, camera, and reference conditions are independently dropped with probabilities 0.10, 0.20, and 0.20, respectively. Single-view text-to-image batches at the same aspect ratio are interleaved every four multi-view updates. We use shifted logit-normal time sampling, with a shift of 6.364 for 81-view training and 4.5 for the single-view stream. For the showcased results, all 81 requested views are sampled jointly from one reference using the shared sampling procedure above. Camera conditioning is retained in the unconditional CFG branch.

B LATENT-SPACE DIAGNOSTIC DEFINITIONS

Table 2 characterizes seven complementary properties of a latent representation: transport ambiguity (ρ), covariance conditioning (κ and effective rank), semantic neighborhoods (LNC@5), within-image spatial structure (LDS and SRSS), and cross-view correspondence (xLNC*). The transport and semantic-neighborhood perspectives follow the latent-diffusability study of Zhong et al. (2026), while LDS and SRSS follow the spatial diagnostics introduced and released by Singh et al. (2025). The covariance quantities are standard spectral statistics. We introduce xLNC* as a multi-view extension of latent-neighbor consistency, using dense correspondences to define cross-view positives.

Evaluation data and preprocessing. For the single-view columns, we use ImageNet-S (Gao et al., 2022) at resolution 256, which provides ImageNet class labels and pixel-level foreground masks. For the multi-view columns Zhou et al. (2018); Ling et al. (2024); Huang et al. (2018); Yeshwanth et al. (2023); Lin et al. (2024), each short scene clip defines one group, so views of the same scene determine the within-group variation. We map spatial neighborhoods to image coordinates before sampling token pairs, making the distance thresholds comparable across latent grids. Tokens are ℓ_2-normalized before cosine-based measurements. The table reports the raw representation consumed by the flow model, without diagnostic-only PCA removal or feature centering.

Within-group velocity fraction ρ (lower is better). Let $\mathbf{\Sigma}_{\text{within}}$ denote the covariance of latent variation within a semantic class for single-view evaluation, or within a scene clip for multi-view evaluation, and let $\mathbf{\Sigma}_{\text{total}}$ be the covariance over the complete evaluation set. We compute

$$\rho = \frac{\operatorname{tr}(\mathbf{\Sigma}_{\text{within}})}{\operatorname{tr}(\mathbf{\Sigma}_{\text{total}})}. \tag{14}$$

This normalized variance ratio accompanies the velocity-ambiguity analysis of Zhong et al. (2026). A low value indicates that samples sharing a class or scene occupy a compact portion of the overall latent distribution and therefore induce less ambiguous transport. Because both terms are traces, the ratio is dimensionless and is not artificially improved by appending inactive channels.

Condition number κ and effective rank (lower/higher is better). We flatten the spatial positions and estimate the channel covariance of the latent. If its eigenvalues are $\lambda_1 \geq \cdots \geq \lambda_C$, then

$$\kappa = \frac{\lambda_1}{\lambda_C}, \qquad \text{effrank} = \frac{\left(\sum_i \lambda_i\right)^2}{\sum_i \lambda_i^2}. \tag{15}$$

κ is the covariance condition number; a large value indicates severe anisotropy or near-rank deficiency. Effective rank is the spectral participation ratio. It equals r for a perfectly flat rank-r spectrum but remains a continuous concentration statistic for realistic spectra. Reporting both quantities distinguishes nominal channel count from the number of directions that carry substantial variance.

Latent neighbor consistency LNC@5 (higher is better). Following the semantic-separability diagnostic of Zhong et al. (2026), we foreground-pool each ImageNet-S latent into one descriptor. We retrieve its five nearest descriptors using cosine similarity and average the fraction whose ImageNet label matches the query. LNC@5 therefore measures local semantic organization; it does not by itself measure reconstruction fidelity or patch-level geometry.

Latent distance structure LDS (higher is better). LDS measures whether nearby image locations remain more similar than distant locations in latent space. For each image, we average token cosine similarity over spatially local pairs and subtract the average over far pairs:

$$\mathrm{LDS} = \mathbb{E}_{(i,j)\in\mathcal{N}}[\langle \hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j \rangle] - \mathbb{E}_{(i,j)\in\mathcal{F}}[\langle \hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j \rangle]. \tag{16}$$

We follow the released LDS implementation of Singh et al. (2026; 2025). Its local and far neighborhoods are expressed in common image coordinates (approximately 24 and 96 pixels in our evaluation), so the comparison is not tied to a particular token-grid resolution. A high score means that latent similarity decays meaningfully with spatial distance instead of collapsing to a nearly uniform descriptor.

Spatial relative similarity score SRSS (higher is better). SRSS uses the ImageNet-S foreground mask to test region-level organization. For each foreground anchor, it compares similarity to nearby foreground tokens with similarity to distant background tokens:

$$\mathrm{SRSS} = \mathbb{E}_{i\in\mathrm{FG}}\left[\mathbb{E}_{j\in\mathcal{P}(i)}\langle \hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j \rangle - \mathbb{E}_{k\in\mathcal{B}(i)}\langle \hat{\mathbf{z}}_i, \hat{\mathbf{z}}_k \rangle\right], \tag{17}$$

where $\mathcal{P}(i)$ contains nearby foreground positives and $\mathcal{B}(i)$ contains spatially distant background negatives, as defined by the ImageNet-S masks. We follow the released SRSS implementation of Zhong et al. (2026); Singh et al. (2025). A high score indicates that foreground regions are locally coherent while remaining distinguishable from background content.

Cross-view latent neighbor consistency xLNC* (higher is better). This is our multi-view extension of LNC. We first use the UFM-Refine dense correspondence model (Zhang et al., 2025) to establish matches between two views. For every matched query token, we retrieve its nearest token in the other view by latent cosine similarity and count the retrieval as correct when it falls within the matched target neighborhood. We chance-normalize the resulting accuracy,

$$\mathrm{xLNC}^* = \frac{a - a_{\mathrm{chance}}}{1 - a_{\mathrm{chance}}}, \tag{18}$$

where a is the measured retrieval accuracy. Thus, 0 denotes chance-level retrieval and 1 denotes perfect correspondence. The main table evaluates view pairs with sufficient visual overlap. Unlike LDS and SRSS, which characterize within-image organization, xLNC* directly measures whether a latent retrieves the same scene content across changes in viewpoint.

C Additional quantitative and qualitative results

C.1 Complete latent-shaping ablation

This table supplements the view-averaged latent-shaping summary in main-paper Table 1, providing the complete single- and multi-view transport and conditioning metrics and effective rank.

Table 8: **Three-step latent shaping on DA3-GIANT.** Supplement to Table 1; Fusion and reconstruction repair the conditioning of the raw hierarchy; token alignment improves transport and semantic organization but collapses relational structure; $\mathcal{L}_{\mathrm{struct}}$ restores spatial and cross-view structure while retaining gains.

	Diffusability (single-view)			Diffusability (multi-view)			Structure (single-view)			Structure (cross-view)
Representation objective	$\rho\downarrow$	$\kappa\downarrow$	effrank↑	$\rho\downarrow$	$\kappa\downarrow$	effrank↑	LNC@5↑	LDS↑	SRSS↑	xLNC* ↑
Raw geometry (DA3-GIANT L0)	0.876	2.8×10^{8}	11.3	0.793	3.9×10^{8}	9.3	0.353	0.102	0.123	0.642
Raw geometry (DA3-GIANT L3)	0.769	6.6×10^{16}	11.5	0.539	2.0×10^{11}	9.0	0.054	0.054	0.070	0.380
None	0.820	**130**	**58.0**	0.630	580	20.0	0.310	0.240	0.300	0.590
$\mathcal{L}_{\mathrm{tok}}$	0.700	170	53.0	0.570	450	26.0	0.490	0.020	0.030	0.580
$\mathcal{L}_{\mathrm{tok}} + \mathcal{L}_{\mathrm{struct}}$	**0.674**	227	51.3	**0.564**	**356**	**45.2**	**0.498**	**0.444**	**0.553**	**0.591**

C.2 Complete appearance-generation metrics

This table supplements the selected FVD and PSNR results in main-paper Table 5, adding the complete FID, LPIPS, and SSIM results for the same controlled variants.

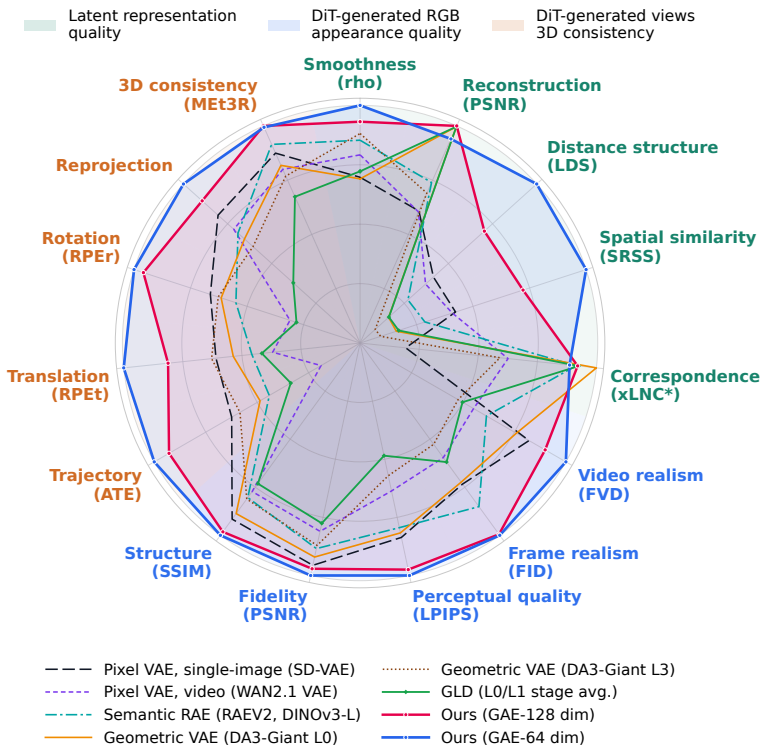

Figure 6: **Latent representation quality and matched DiT generation.** Normalized overview of representation (green), RGB generation (blue), and 3D consistency (orange).

Table 9: **Complete camera-conditioned appearance generation.** Supplement to Table 5; full FID, FVD, LPIPS, PSNR, and SSIM results for all controlled variants.

	RealEstate10K					DL3DV				
Latent	FVD↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	FVD↓	FID↓	LPIPS↓	PSNR↑	SSIM↑
Pixel VAE, single-image (SD-VAE)	258.6	36.4	0.171	19.32	0.666	373.2	57.1	0.232	17.05	0.495
Pixel VAE, video (WAN2.1 VAE)	362.9	41.0	0.219	16.52	0.575	596.7	77.3	0.313	14.31	0.407
Semantic RAE (RAEV2)	379.4	31.4	0.191	17.60	0.585	453.4	49.2	0.231	16.01	0.434
Raw geometry (DA3-GIANT L0)	298.6	35.5	0.177	18.31	0.643	376.5	60.2	0.235	16.68	0.481
Raw geometry (DA3-GIANT L3)	488.9	50.0	0.273	16.82	0.579	584.8	80.6	0.314	16.22	0.447
Ours (GAE-128)	233.4	**25.6**	0.146	19.54	0.701	345.2	44.2	0.200	17.39	0.541
Ours (GAE-64)	**225.7**	27.1	**0.143**	**20.02**	**0.711**	**287.0**	**41.3**	**0.194**	**18.00**	**0.553**
GLD	445.1	44.0	0.337	14.91	0.518	587.4	66.4	0.356	14.51	0.406
Gen3R	269.7	25.7	0.196	17.79	0.621	580.5	48.3	0.270	15.62	0.440

Qualitative RGB comparisons. Figures 11 and 12 show all nine views for one RealEstate10K scene and one DL3DV scene from the shared 64-scene evaluation pools. The displayed PSNR is computed over the eight generated views of that scene, not averaged across the dataset; these qualitative examples complement the quantitative appearance results in Table 5.

C.3 Complete 3D-consistency results

This table supplements the selected ATE, Reproj., and MEt3R results in main-paper Table 5, and includes the full VGGT evaluation, MEt3R, and the DA3-GIANT cross-check.

Independent 3D check. Figure 8 compares appearance-native, semantic-native, and geometry-native generated states on two scenes. Alongside the generated views, an independent reconstruction pipeline recovers a point cloud and camera trajectory from each output clip. This separates visual



Figure 7: **Qualitative RGB comparison on RealEstate10K.** Columns are methods and rows show one reference and two target views for each scene under identical camera conditioning. Insets report per-scene PSNR over the two displayed target views; Ours denotes the C=64 model.

Table 10: **Complete generated-view 3D consistency.** Supplement to Table 5; VGGT provides the independent evaluation, DA3-GIANT the backbone-overlap cross-check, and MEt3R the pairwise feature consistency.

	VGGT (independent)				DA3-GIANT (cross-check)				
Latent	ATE↓	RPEt↓	RPEr° ↓	Reproj.↓	ATE↓	RPEt↓	RPEr° ↓	Reproj.↓	MEt3R↓
RealEstate10K									
Pixel VAE, single-image (SD-VAE)	0.0072	0.0182	0.648	0.0093	0.0046	0.0191	0.338	0.0018	0.1385
Pixel VAE, video (WAN2.1 VAE)	0.0225	0.0252	0.746	0.0105	0.0135	0.0176	0.751	0.0019	0.1441
Semantic RAE (RAEV2)	0.0085	0.0233	0.420	0.0104	0.0066	0.0219	0.507	0.0019	0.1263
Raw geometry (DA3-GIANT L0)	0.0085	0.0202	0.435	0.0096	0.0047	0.0167	0.358	0.0020	0.1420
Raw geometry (DA3-GIANT L3)	0.0112	0.0211	0.516	0.0124	0.0036	0.0148	0.355	0.0019	0.1575
Ours (GAE-128)	0.0041	0.0173	**0.332**	0.0075	**0.0026**	0.0143	0.304	0.0017	0.1213
Ours (GAE-64)	**0.0034**	**0.0060**	0.380	**0.0068**	**0.0026**	**0.0141**	**0.225**	**0.0013**	**0.1208**
GLD	0.0124	0.0258	0.886	0.0133	0.0088	0.0221	0.950	0.0038	0.1660
Gen3R	0.0090	0.0213	0.423	0.0072	0.0063	0.0211	0.387	0.0014	0.1157
DL3DV									
Pixel VAE, single-image (SD-VAE)	0.0073	0.0098	0.432	0.0108	0.0055	0.0076	0.358	0.0024	0.1542
Pixel VAE, video (WAN2.1 VAE)	0.0243	0.0288	1.419	0.0258	0.0206	0.0251	1.016	**0.0019**	0.1777
Semantic RAE (RAEV2)	0.0104	0.0143	0.638	0.0130	0.0088	0.0113	0.585	0.0031	0.1549
Raw geometry (DA3-GIANT L0)	0.0110	0.0136	0.634	0.0129	0.0079	0.0102	0.482	0.0042	0.1721
Raw geometry (DA3-GIANT L3)	0.0097	0.0120	0.603	0.0151	0.0052	0.0073	0.380	0.0039	0.1754
Ours (GAE-128)	**0.0056**	0.0072	0.294	0.0080	0.0043	**0.0053**	0.260	0.0029	**0.1347**
Ours (GAE-64)	**0.0056**	**0.0066**	**0.293**	**0.0077**	**0.0038**	**0.0053**	**0.254**	0.0028	0.1370
GLD	0.0136	0.0164	1.242	0.0241	0.0104	0.0136	1.048	0.0087	0.2183
Gen3R	0.0096	0.0133	0.678	0.0143	0.0096	0.0131	0.608	0.0065	0.1662

plausibility from whether the generated frames jointly support a coherent 3D scene and camera path, complementing the quantitative consistency results in Table 5.

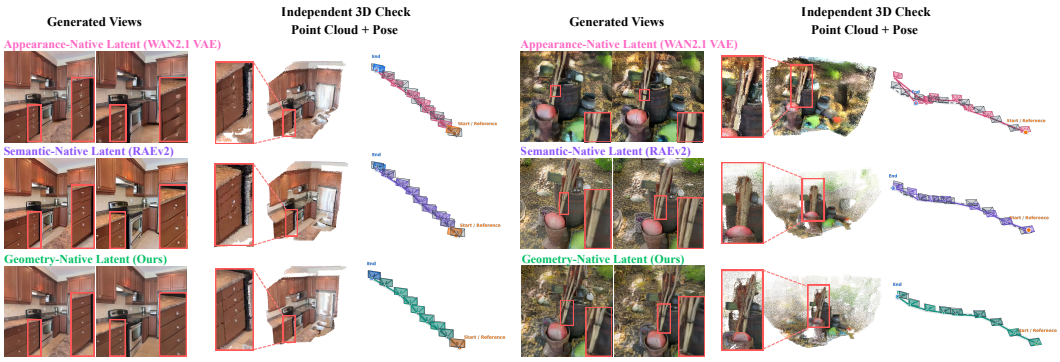

Figure 8: **Independent 3D check of generated views.** Supplement to main-paper Figure 3. The appearance-native WAN2.1 VAE, semantic-native RAEV2, and our geometry-native latent receive the same reference and camera controls. The right-hand panels show point clouds and camera trajectories reconstructed from the generated RGB clips by an independent geometry model VGGT.

C.4 Appearance and 3D consistency on more scenes

Tables 11 and 12 extend the camera-conditioned evaluation to larger scene pools using the same model checkpoints as the main comparison. For each dataset, we randomly select 256 scenes. The frame interval is 10 for RealEstate10K and 1 for DL3DV. Each example contains nine views with one clean reference. Sampling uses 50 steps and a classifier-free guidance scale of 2.

Appearance metrics are computed from the saved generated videos, and 3D consistency is evaluated using the same estimators and metrics as above. On RealEstate10K, GAE-128 achieves the best appearance results across all five metrics, while GAE-64 leads on DL3DV. Both variants also perform strongly in 3D consistency, although SD-VAE achieves the second-lowest VGGT relative-rotation error on RealEstate10K. The best and second-best distinct displayed values in each column are highlighted; equal displayed values receive the same rank.

Table 11: **Camera-conditioned appearance generation on 256 scenes.** Supplement to Table 5.

	RealEstate10K					DL3DV				
Latent	FVD↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	FVD↓	FID↓	LPIPS↓	PSNR↑	SSIM↑
Pixel VAE, single-image (SD-VAE)	102.8	18.8	0.200	18.04	0.632	96.4	24.5	0.156	20.51	0.622
Pixel VAE, video (WAN2.1 VAE)	142.8	25.4	0.242	15.96	0.568	160.7	30.1	0.214	16.84	0.513
Semantic RAE (RAEV2)	109.3	13.0	0.194	17.56	0.590	150.0	18.3	0.158	18.77	0.539
Raw geometry (DA3-GIANT L0)	111.5	18.7	0.207	17.47	0.619	109.6	26.0	0.155	19.73	0.609
Raw geometry (DA3-GIANT L3)	231.5	24.4	0.283	16.40	0.568	251.4	39.2	0.252	18.00	0.535
Ours (GAE-128)	**76.4**	**12.3**	**0.157**	**18.84**	**0.681**	64.5	14.9	0.120	21.33	0.686
Ours (GAE-64)	78.3	13.0	0.163	18.76	0.677	**61.7**	**13.2**	**0.111**	**21.53**	**0.694**

C.5 Additional qualitative capabilities

The main paper retains one representative row for each capability in Figure 5. Figure 9 supplements panel (a), the long-rollout showcase, while Figure 10 supplements panel (b), the text-conditioned RGB–geometry showcase.

C.6 Additional conditioning ablations

This subsection isolates two conditioning choices described in Section 3.3 and Appendix A.2. Table 13 reports both comparisons in full. All variants use the controlled RealEstate10K setting, the same codec and flow backbone, one reference view, nine output views, 50 Euler steps, and CFG $= 2$; within each block, only the indicated conditioning choice changes.

The reference-conditioning block compares two ways to inject the observed image. In the first variant, the reference latent is included among the ODE state variables. In the second, it remains clean,

Table 12: **Generated-view 3D consistency on 256 scenes.**

	VGGT (independent)				DA3-GIANT (cross-check)				
Latent	ATE↓	RPEt↓	RPEr° ↓	Reproj.↓	ATE↓	RPEt↓	RPEr° ↓	Reproj.↓	MEt3R↓
RealEstate10K									
Pixel VAE, single-image (SD-VAE)	0.0088	0.0151	0.517	0.0137	0.0055	0.0095	0.532	0.0029	0.1679
Pixel VAE, video (WAN2.1 VAE)	0.0201	0.0243	0.931	0.0137	0.0133	0.0167	1.026	0.0031	0.1858
Semantic RAE (RAEV2)	0.0137	0.0204	0.536	0.0156	0.0080	0.0141	0.634	0.0031	0.1604
Raw geometry (DA3-GIANT L0)	0.0129	0.0189	0.776	0.0143	0.0087	0.0112	0.632	0.0038	0.1792
Raw geometry (DA3-GIANT L3)	0.0095	0.0122	0.730	0.0151	0.0058	0.0090	0.613	0.0036	0.1889
Ours (GAE-128)	**0.0050**	0.0109	**0.517**	**0.0115**	0.0045	0.0085	0.462	0.0026	**0.1533**
Ours (GAE-64)	0.0053	**0.0108**	0.534	0.0116	**0.0037**	**0.0084**	**0.388**	**0.0025**	0.1534
DL3DV									
Pixel VAE, single-image (SD-VAE)	0.0047	0.0065	0.323	0.0089	0.0047	0.0062	0.297	0.0011	0.1118
Pixel VAE, video (WAN2.1 VAE)	0.0121	0.0159	1.191	0.0121	0.0123	0.0153	1.023	0.0018	0.1209
Semantic RAE (RAEV2)	0.0073	0.0107	0.474	0.0097	0.0069	0.0093	0.460	0.0014	0.1146
Raw geometry (DA3-GIANT L0)	0.0068	0.0089	0.541	0.0101	0.0069	0.0081	0.513	0.0016	0.1255
Raw geometry (DA3-GIANT L3)	0.0058	0.0079	0.458	0.0112	0.0049	0.0065	0.330	0.0017	0.1324
Ours (GAE-128)	0.0036	0.0052	0.256	0.0064	0.0039	0.0052	0.243	0.0010	0.0995
Ours (GAE-64)	**0.0035**	**0.0050**	**0.247**	**0.0062**	**0.0038**	**0.0051**	**0.238**	**0.0009**	**0.0971**

Table 13: **Reference conditioning and text-to-image co-training on RealEstate10K.** Diag. denotes the reference–target point-cloud gap or LPIPS, respectively.

Setting	FVD ↓	PSNR ↑	MEt3R ↓	Diag. ↓
Reference conditioning (Diag. = Ref–tgt PC gap)				
Reference latent in ODE state	257.1	18.91	0.1217	0.1016
Clean conditioning token	**225.7**	**20.02**	**0.1208**	**0.0318**
Text-to-image co-training (Diag. = LPIPS)				
Without T2I co-training	472.9	16.77	0.1262	0.199
With T2I co-training	**225.7**	**20.02**	**0.1208**	**0.143**

is prepended to the noisy all-view tokens, is assigned $t = 0$, and is excluded from the prediction head. The clean-token design yields lower FVD, higher PSNR, lower MEt3R, and a much smaller reference–target point-cloud gap, supporting the fixed-context construction used by the final model.

The text-to-image block removes or restores the single-view text-to-image updates interleaved with multi-view training. Co-training substantially lowers FVD and LPIPS while improving PSNR and MEt3R, showing that the additional single-view stream improves appearance generation under the same camera-conditioned evaluation protocol.

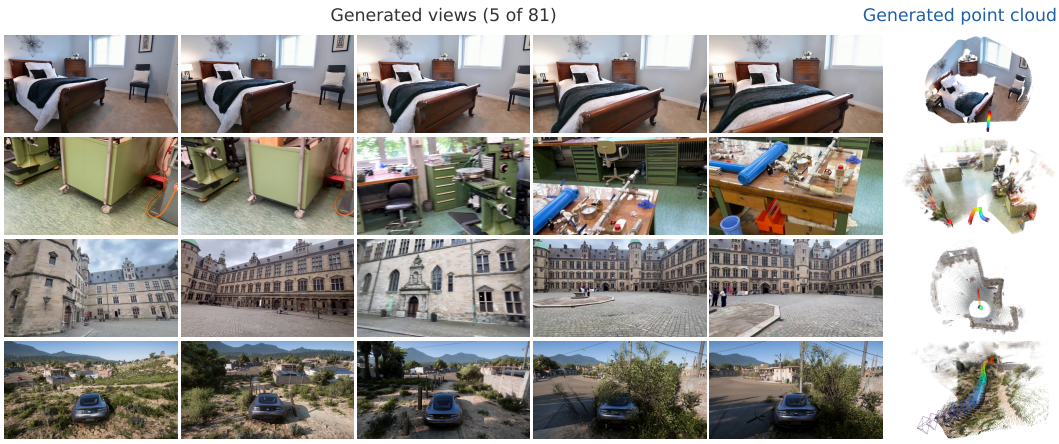

Figure 9: **Complete long-rollout showcase.** Supplement to panel (a) of main-paper Figure 5. Five generated views from each 81-view rollout and the point cloud decoded from the same latents. The rollouts use one reference view, 50 Euler steps, CFG $= 2$, and decode RGB and geometry from each sampled latent without cross-view fusion or test-time reconstruction.

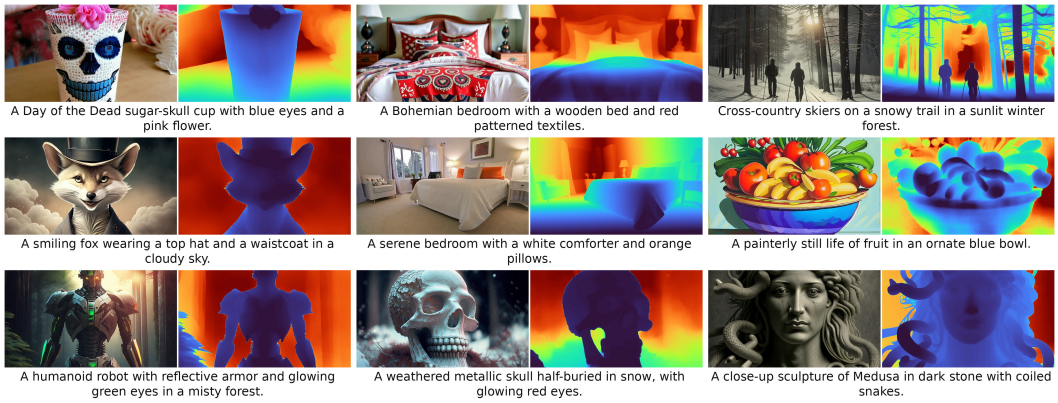

Figure 10: **Text-to-image RGB and geometry generation.** Supplement to panel (b) of main-paper Figure 5. Each prompt is followed by its generated RGB image and color-coded depth map decoded from the same latent.

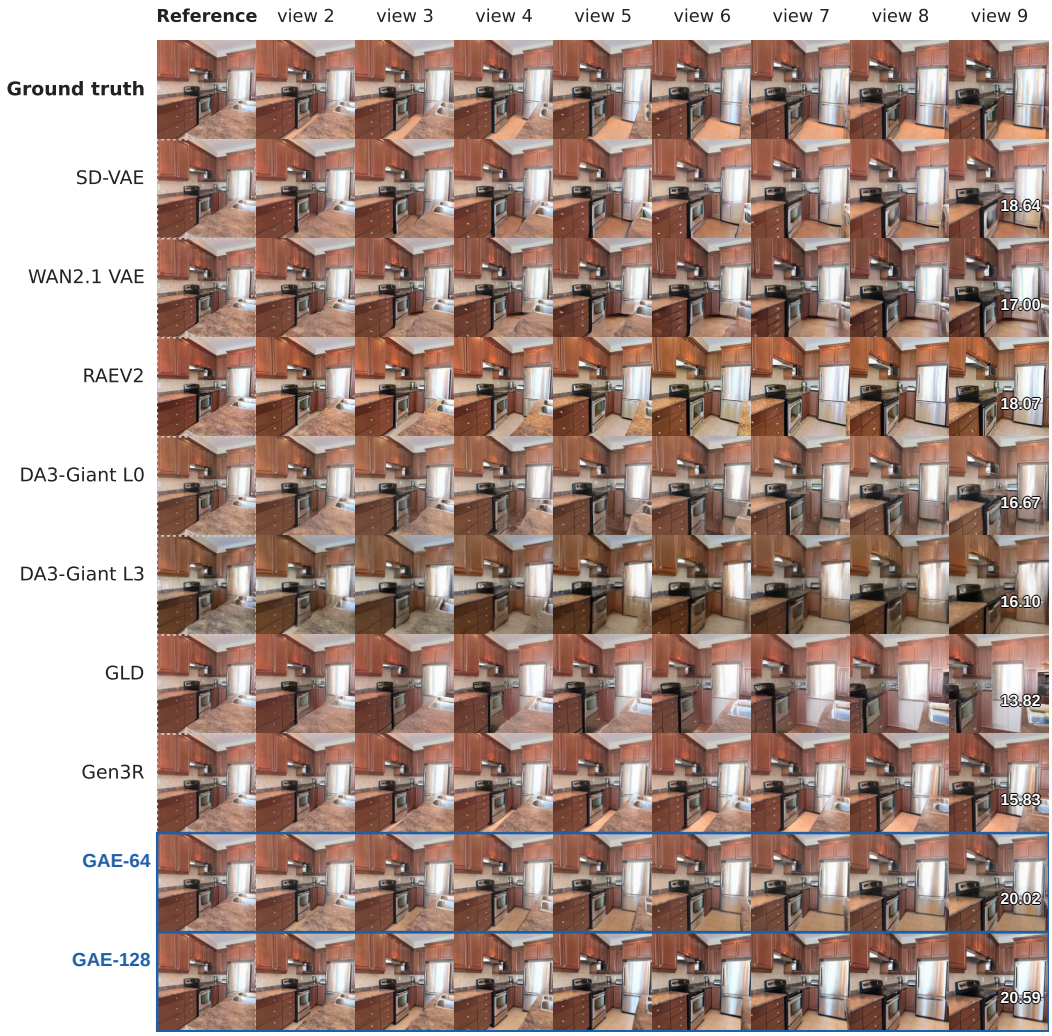

Figure 11: **RealEstate10K comparison.** All methods receive the same reference image and target camera trajectory, and the eight remaining views are generated jointly.

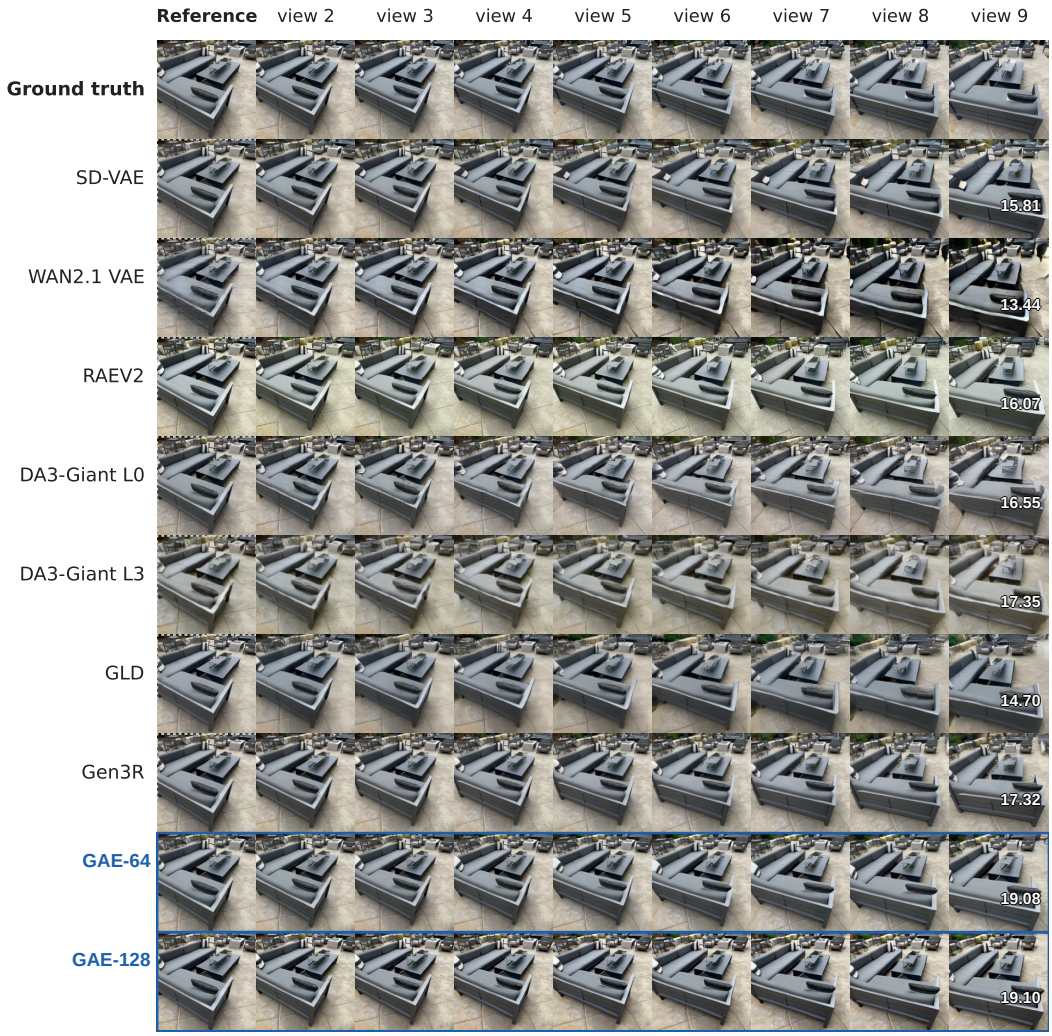

Figure 12: **DL3DV comparison.** All methods receive the same reference image and target camera trajectory, and the eight remaining views are generated jointly.