arXiv:2609.24981v1 [cs.CV] 21 Sep 2026

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

Jiahao Lu[1*] **Minghao Yin**[2,3*] **Wenbo Hu**[2†] **Hengyu Liu**[4]
Wang Zhao[2] **Sai-Kit Yeung**[1] **Ying Shan**[2] **Yuan Liu**[1†]

[1]The Hong Kong University of Science and Technology · [2]ARC Lab, Tencent IEG
[3]The University of Hong Kong · [4]The University of Texas at Austin

[*]Equal contribution [†]Corresponding authors

Project Page



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.

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 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. Visual generation has long separated perceptual compression from generative modeling through variational and vector-quantized tokenizers, with latent diffusion establishing compact continuous codes as the standard generative state (Kingma & Welling, 2014; van den Oord et al., 2017; Esser et al., 2021; Rombach et al., 2022). Representation autoencoders (RAEs) instead use frozen pretrained visual representations as the encoder state, with later work scaling the idea to text-to-image generation and multi-layer features (Zheng et al., 2026; Tong et al., 2026; Singh et al., 2026). REPA aligns denoiser features rather than changing the output state, while latent-diffusability studies show that reconstruction alone does not determine generation quality (Yu et al., 2025a; Zhong et al., 2026a). These methods organize generation around appearance or semantics. GAE instead asks for a compact state that remains natively readable by a geometry foundation model.

Geometry-aware visual generation. Scene-based novel-view synthesis reconstructs radiance fields or Gaussian primitives, while generative methods use pose-conditioned diffusion to synthesize unobserved 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 similarly introduces trajectories through conditioning or positional encoding (Wang et al., 2024b; He et al., 2025; Yin et al., 2026). Joint-output methods generate RGB together with depth or normals (Stan et al., 2023; Krishnan et al., 2025; Kwon et al., 2025; Guizilini et al., 2025). Other approaches align video features to geometry representations, condition on geometry features, couple geometry and appearance latents, or post-train with geometric rewards (Wu et al., 2026; Wan et al., 2026; Huang et al., 2026; Mi et al., 2026; Xiang et al., 2026; An et al., 2026). Across these generative 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 rather than adding geometry to an appearance-native state.

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/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 features have also become predictive states for planning, geometry forecasting, and stochastic world modeling (Zhou et al., 2025a; Sun et al., 2026; Porcher et al., 2026). The closest visual-generation precedents directly model geometry-foundation features: GLD cascades selected DA3 or VGGT levels and propagates the remaining hierarchy (Jang et al., 2026), whereas latent Riemannian flow matching jointly evolves VGGT's four normalized levels on their product manifold (Weijler et al., 2026). Both preserve native geometry readout but generate high-dimensional backbone features directly. GAE instead learns one 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

Geometry-Native Autoencoder (GAE) places the 3D inductive bias in the generated state itself. The method trains in two stages, shown in Figure 2: Stage 1 trains a codec that turns the multi-level features of a frozen geometry backbone into one compact latent decodable to both RGB and

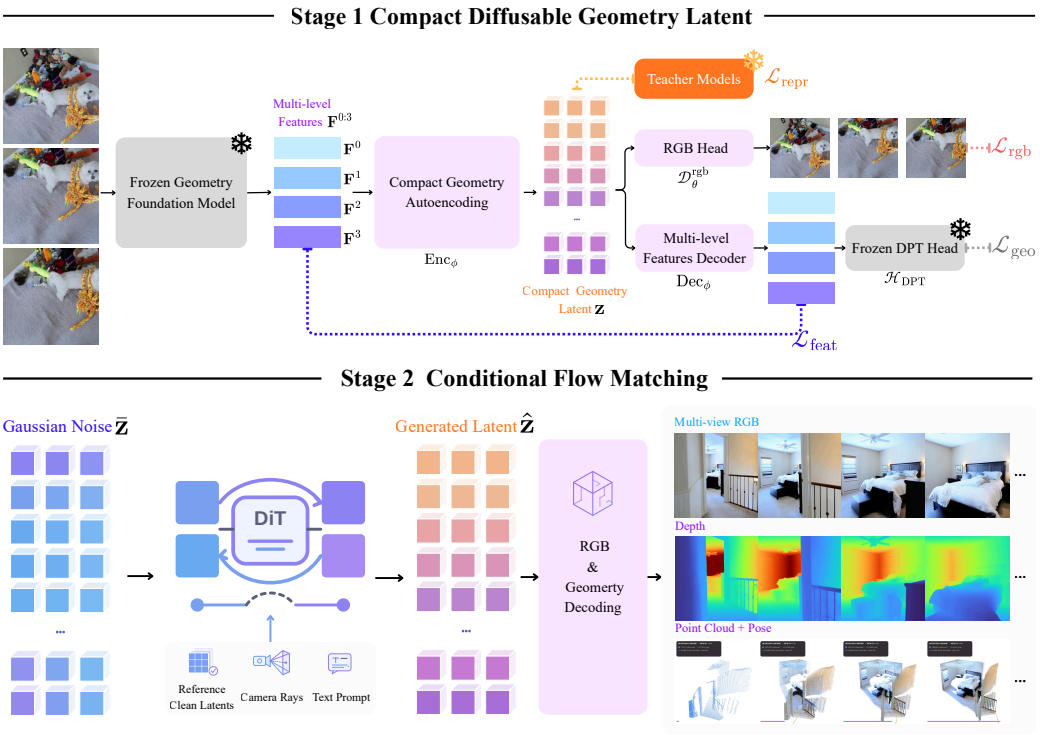


Figure 2: **Overview of Geometry-Native Autoencoder (GAE).** *Stage 1 (codec training)* compresses frozen multi-level geometry foundation features into a compact, diffusable geometry latent; the feature decoder rebuilds the full hierarchy for frozen geometry readout, while a separate learned head decodes RGB. *Stage 2 (flow training)* performs conditional flow matching only over generated-view latents, while clean reference latents, camera rays, and text act as controls.

geometry, and Stage 2 freezes that codec and trains a conditional flow model in its standardized latent space.

The two stages are summarized by

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

Here $\mathcal{E}$ is the frozen DA3 encoder and $\mathcal{P}$ normalizes and fuses its four feature levels. Stage 1 trains Enc_ϕ and Dec_ϕ to compress and rebuild $\mathbf{X}_{1:V}$; once frozen, its posterior mean standardized by Std supplies the flow targets $\bar{\mathbf{z}}^{\mathrm{gt}}_{1:V}$. Flow 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 latent tokens encoded from reference views. Only the target-view latents evolve as ODE states; all other signals remain fixed controls. Both outputs are read from the sampled latent:

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

The trainable codec $\mathcal{A}_\phi = (\mathrm{Enc}_\phi, \mathrm{Dec}_\phi)$ reconstructs the full DA3 hierarchy, which the frozen geometry head $\mathcal{H}_{\mathrm{DPT}}$ reads as depth, rays, and point maps. A separate learned head $\mathcal{D}_\theta^{\mathrm{rgb}}$ renders RGB from the same latent. Thus, unlike a pixel VAE, the state remains geometry-decodable; unlike raw-feature diffusion, it is compact enough to be modeled by a single flow network. We write $\mathbf{z}$ for the codec latent, $\bar{\mathbf{z}}$ for its standardized form used by flow matching, and $\bar{\mathbf{z}}^{\mathrm{ref}}$ for clean reference evidence.

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 Section 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 level preserves fine detail and cross-view correspondence, while the deeper levels provide complementary inputs required by the frozen geometry decoder. This distinction does not imply stronger semantic or spatial structure at greater depth; our diagnostics show the opposite for L3 (Table 2).

Before fusion, we normalize each level using fixed, 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:

$$(\boldsymbol{\mu}_v, \log \boldsymbol{\sigma}_v^2) = \mathrm{Enc}_\phi(\mathbf{X}_v), \tag{4}$$

$$\mathbf{z}_v = \boldsymbol{\mu}_v + \boldsymbol{\epsilon}_v \odot \exp\left(\tfrac{1}{2} \log \boldsymbol{\sigma}_v^2\right), \qquad \boldsymbol{\epsilon}_v \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{5}$$

$$\hat{\mathbf{F}}_v^{0:3} = \mathrm{Dec}_\phi(\mathbf{z}_v). \tag{6}$$

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 returns $\hat{\mathbf{F}}_v^{0:3} = \{\hat{\mathbf{f}}_{v,\ell}\}_{\ell=0}^{3}$. 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:

$$\hat{\mathbf{I}}_v = \mathcal{D}_\theta^{\mathrm{rgb}}(\mathbf{z}_v), \qquad \hat{\mathbf{G}}_v = \mathcal{H}_{\mathrm{DPT}}(\hat{\mathbf{F}}_v^{0:3}) = \mathcal{H}_{\mathrm{DPT}}(\mathrm{Dec}_\phi(\mathbf{z}_v)), \tag{7}$$

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:

$$\mathcal{L}_{\mathrm{codec}} = \mathcal{L}_{\mathrm{feat}} + \lambda_{\mathrm{kl}} \mathcal{L}_{\mathrm{kl}} + \mathcal{L}_{\mathrm{rgb}} + \mathcal{L}_{\mathrm{geo}} + \mathcal{L}_{\mathrm{repr}}. \tag{8}$$

Here $\mathcal{L}_{\mathrm{feat}}$ 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. The final term $\mathcal{L}_{\mathrm{repr}}$ shapes how information is organized inside the bottleneck rather than what it reconstructs; the next section motivates this term.

3.2 ORGANIZING THE LATENT FOR GENERATION

Table 1: **Three-step latent shaping on DA3-GIANT.** Fusion and reconstruction repair the conditioning of the raw hierarchy; token alignment improves transport and semantic organization but collapses relational structure; $\mathcal{L}_{\text{struct}}$ 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
$\mathcal{L}_{\text{tok}}$	0.700	170	53.0	0.570	450	26.0	0.490	0.020	0.030	0.580
$\mathcal{L}_{\text{tok}}+\mathcal{L}_{\text{struct}}$	**0.674**	227	51.3	**0.564**	**356**	**45.2**	**0.498**	**0.444**	**0.553**	**0.591**

Reconstruction determines what the bottleneck must preserve, but not how that information should be organized. A reconstruction-only codec can therefore recover RGB and geometry while still producing a latent that is difficult for a flow model to learn. Table 1 summarizes these representation diagnostics; full metric definitions are provided in the Sec. 4.1 and supplementary material. The results make this gap visible: although the reconstruction baseline is compact and well-conditioned, its transport smoothness and semantic neighborhoods remain weak.

We first add position-wise representation supervision. Let $\boldsymbol{\mu}_{v,i}$ be the posterior-mean token at location i in view v. We align a projected token with the co-located feature from a frozen C-RADIO teacher (Heinrich et al., 2025). This token-level objective improves transport smoothness (ρ) and semantic neighborhood consistency (LNC). However, it treats each location independently. It can therefore align individual tokens while destroying the relations among them, which sharply reduces LDS and SRSS and slightly weakens cross-view correspondence.

We address this failure with a complementary relational objective. A frozen DINOv2 teacher (Oquab et al., 2024) provides pairwise similarities among spatial locations, and we match those similarities directly in the raw posterior space. This constrains the neighborhood geometry omitted by position-wise alignment without requiring the student and teacher channel dimensions to match. Table 1 summarizes this three-step progression.

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

$$\mathcal{L}_{\text{tok}} = \frac{1}{VN}\sum_{v,i}\left[1-\left\langle \widehat{g_\eta(\boldsymbol{\mu}_{v,i})}, \hat{\mathbf{c}}_{v,i}\right\rangle\right], \tag{9}$$

$$\mathcal{L}_{\text{struct}} = \frac{1}{VN(N-1)}\sum_{v}\sum_{i\neq j}\left(\langle\hat{\boldsymbol{\mu}}_{v,i},\hat{\boldsymbol{\mu}}_{v,j}\rangle - \left\langle\hat{\mathbf{d}}_{v,i},\hat{\mathbf{d}}_{v,j}\right\rangle\right)^2, \tag{10}$$

$$\mathcal{L}_{\text{repr}} = \lambda_{\text{repa}}\left(\mathcal{L}_{\text{tok}}+\lambda_\mu\mathcal{L}_{\text{struct}}\right), \qquad \lambda_{\text{repa}}=0.25, \quad \lambda_\mu=8.0. \tag{11}$$

Here g_η maps posterior tokens to the C-RADIO feature dimension and hats denote ℓ_2 normalization. C-RADIO supplies token-wise semantic organization, while DINOv2 restores relational structure directly in the latent used by the flow model. 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 discarded after codec training.

After training, we freeze the codec, use the posterior mean $\mathbf{z}_v := \boldsymbol{\mu}_v$, and standardize each channel with training-set statistics:

$$\bar{\mathbf{z}}_v = \frac{\mathbf{z}_v - \mathbf{m}_z}{\mathbf{s}_z+\epsilon}. \tag{12}$$

3.3 CONDITIONAL FLOW MATCHING AND GENERATION

Flow matching. After freezing the codec, we model only its standardized posterior means. Given target latents $\bar{\mathbf{z}}^{\text{gt}}$ and noise $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$, we use the linear path $\bar{\mathbf{z}}_t=(1-t)\bar{\mathbf{z}}^{\text{gt}}+t\boldsymbol{\epsilon}$ and train a conditional transformer by

$$\mathcal{L}_{\text{flow}} = \mathbb{E}\left[\rho(t)\left\|\hat{\mathbf{u}}_\psi(\bar{\mathbf{z}}_t,t,\mathcal{C})-(\boldsymbol{\epsilon}-\bar{\mathbf{z}}^{\text{gt}})\right\|_2^2\right]. \tag{13}$$

Following RAEv2 (Singh et al., 2026), the network uses clean-latent prediction, converted to the velocity in Eq. 13. A single transformer jointly models all target views, allowing cross-view interaction directly in the compact geometry-native state. Parameterization and sampling details are given in the supplementary material.

Reference, camera, and text conditioning. Geometry encoders are set-conditioned: the feature of a reference image encoded jointly with target views differs from that of the same image encoded alone. Placing a full-set reference latent directly in the flow state therefore introduces a train/inference mismatch, since the full-set reference features available during training cannot be constructed at test time. Instead, GAE jointly encodes the K observed reference views using only the observed references as context, and prepends the resulting clean latents $\bar{\mathbf{z}}^{\mathrm{ref}}$ as conditioning tokens. All V output slots, including those at reference camera poses, remain noisy flow variables. The clean reference tokens $\bar{\mathbf{z}}^{\mathrm{ref}}$ are assigned timestep $t = 0$ and attend jointly with all V noisy state tokens, but are removed before the decoder prediction head. This design avoids clamping reference slots in the ODE, yielding a single standard Euler sampler without a denoising-level mismatch. Camera control is provided by metric Plücker-ray embeddings in query/key self-attention (Yin et al., 2026), while frozen language-model features enter through cross-attention. Condition dropout supports text-only, camera-controlled, and reference-conditioned generation with the same weights. Precise token and ray construction is deferred to the supplementary material.

Training and inference. Stage 1 trains the codec and RGB decoder on mixed single- and multi-view data while keeping the geometry backbone and DPT head frozen. Stage 2 freezes the complete codec and trains the DiT flow model on both text-to-image (T2I) and view-conditioned generation, with variable view/reference counts and condition dropout. At inference, Gaussian target-view latents are integrated from $t = 1$ to 0, denormalized, and decoded into RGB and geometry through Eq. 2. Thus, one flow model generates a shared latent state without a hierarchy-level cascade or a separate geometry estimator. Architecture, optimization, and solver details appear in the supplementary material.

4 EXPERIMENTS

We evaluate the representation before testing the generator built on it. We first analyze latent-space properties (Section 4.1) and codec reconstruction (Section 4.2). We then evaluate camera-controllable RGB and geometry generation with matched flow models (Section 4.3), followed by ablations of the codec and conditioning designs (Section 4.4).

Experimental setup. 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. Every controlled variant is trained as a matched flow model on RealEstate10K (Zhou et al., 2018) and DL3DV (Ling et al., 2024), using the same flow-model family, camera conditioning, training budget, and sampling protocol, and is evaluated on a shared held-out pool of 64 scenes with nine views per scene at 252^2, one reference view, 50 Euler steps, and CFG $= 2$. Section 4.3.4 instead shows a separately trained final model, using larger-scale data, higher resolution, and more frames. Those results are qualitative and are not part of the comparison.

Ranking convention. Blue cells mark the best and second-best controlled variants, and displayed ties are ordered using unrounded values unless a table states otherwise. Gray rows are external systems that are excluded from this ranking, because they are complete pipelines rather than latents dropped into our shared flow model: GLD (Jang et al., 2026) generates a cascade of DA3 feature levels, and Gen3R (Huang et al., 2026) is a pretrained geometry-generation system.

4.1 LATENT-SPACE STUDY

We measure transport complexity with ρ (rho) (Zhong et al., 2026b), spectral conditioning with κ and effective rank, and representation structure with LNC@5 (Zhong et al., 2026b), LDS (Singh et al., 2026; 2025), SRSS (Zhong et al., 2026b; Singh et al., 2025), and cross-view retrieval xLNC*. Lower is better for ρ and κ; higher is better otherwise. Full definitions are in the supplementary material.

Figure 3 provides a compact overview of intrinsic representation quality and matched DiT generation; Table 2 reports the raw diagnostics. 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.

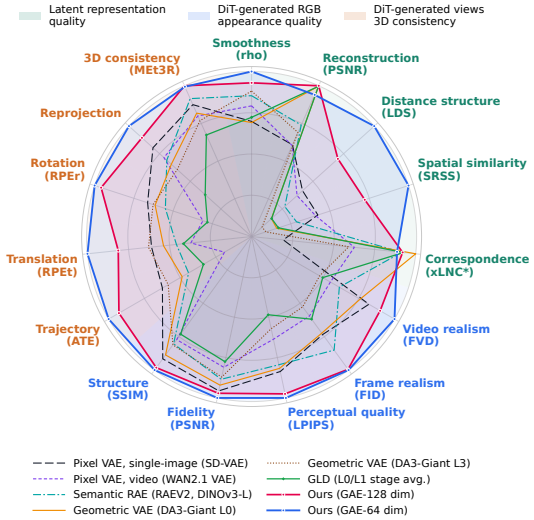


Figure 3: **Latent representation quality and matched DiT generation.** Normalized overview of representation (green), RGB generation (blue), and 3D consistency (orange).

Table 2: **Latent-space diagnostics.** GAE combines smooth transport and good conditioning with strong spatial and cross-view structure while retaining useful semantic organization.

		Diffusability (single-view)			Diffusability (multi-view)			Structure (single-view)			Structure (cross-view)
Latent	$C\times H\times W$	$\rho\downarrow$	$\kappa\downarrow$	effrank↑	$\rho\downarrow$	$\kappa\downarrow$	effrank↑	LNC@5↑	LDS↑	SRSS↑	xLNC* ↑
Pixel VAE, single-image (SD-VAE)	$4\times32\times32$	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)	$16\times32\times32$	0.847	127	5.3	0.609	160	4.3	0.084	0.234	0.313	0.402
Semantic RAE (RAEV2)	$1024\times16\times16$	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)	$3072\times18\times18$	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)	$3072\times18\times18$	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)	$128\times18\times18$	0.674	227	51.3	0.564	356	45.2	0.498	0.444	0.553	0.591
Ours (GAE-64)	$64\times18\times18$	**0.634**	53.1	37.0	**0.519**	117	33.2	0.488	**0.631**	**0.767**	0.569

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. Table 3 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 (Table 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 3: **RGB reconstruction.** Single-view uses ImageNet; multi-view averages five video datasets.

	Single-view			Multi-view		
Latent	PSNR↑	LPIPS↓	rFID↓	PSNR↑	LPIPS↓	rFVD↓
Pixel VAE, single-image (SD-VAE)	18.49	0.140	10.1	19.97	0.121	37.7
Pixel VAE, video (WAN2.1 VAE)	18.34	0.125	8.9	19.83	0.163	46.7
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

Table 4 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. Only δ_1 is higher-is-better. 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-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.

Table 4: **Geometry reconstruction from encoded latents.** Depth and point maps are compared with Pi3 (Wang et al., 2026c) on the real target frames; camera poses are compared with the input real trajectory. Displayed ties between GAE-64 and GAE-128 are ordered using unrounded values.

	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

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

Table 5 evaluates distributional quality with FID/FVD and paired fidelity with LPIPS, PSNR, and SSIM. GAE-64 gives the strongest overall results on both datasets, reducing FVD by 12.7% on RealEstate10K and 23.1% on DL3DV relative to the best non-GAE controlled latent. It also leads the paired metrics, while GAE-128 gives the best RealEstate10K FID. Figure 4 provides the corresponding RGB comparison on two RealEstate10K scenes. Under the same reference image and camera trajectory, GAE-64 better preserves object identity and fine structure as the viewpoint moves away from the reference, consistent with the metrics in Table 5.

Table 5: **Camera-conditioned appearance generation.** All controlled latent variants use the matched protocol described above.

	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



Figure 4: **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.

4.3.2 3D CONSISTENCY

Table 6 evaluates camera and multi-view consistency using VGGT (Wang et al., 2025a), MEt3R (Asim et al., 2025), and a DA3-GIANT (Lin et al., 2026) cross-check. For camera recovery, ATE measures global trajectory error after Sim(3) alignment, while RPEt and RPEr measure relative translation and rotation errors, respectively, with RPEr in degrees. For reprojection error (Jang et al., 2026), we Sim(3)-align reconstructed point maps to the dataset camera frame and measure the normalized 2D round-trip pixel error between high-confidence points across views, retaining the top 50% confidence pixels. MEt3R (Asim et al., 2025) directly measures symmetric feature disagreement between reconstructed view pairs with visibility and occlusion masking. Across both datasets, the two GAE variants achieve the two best controlled VGGT results. GAE-64 reduces ATE by 52.8% on RealEstate10K and 23.3% on DL3DV, while GAE-64 and GAE-128 achieve the best controlled MEt3R on RealEstate10K (0.1208) and DL3DV (0.1347), respectively. Together, these results show that the appearance gains do not compromise camera control or multi-view consistency. Figure 5 further shows that GAE yields a more coherent 3D scene and more faithful camera motion under the same reference view and prescribed trajectory.

4.3.3 GEOMETRY GENERATION

Table 7 evaluates geometry decoded from sampled latents using the metrics defined in Section 4.2.2. We directly decode depth, camera rays, and point maps from each latent. Depth and point maps are compared with Pi3 (Wang et al., 2026c) applied to the real target frames, while camera trajectories are compared with the input dataset poses. This common reference avoids evaluating each method against geometry estimated from its own generated images.

On DL3DV, GAE-64 leads all six geometry metrics among the controlled geometry-native methods. On RealEstate10K, it gives the best depth accuracy and pose estimates, while GAE-128 gives lower Chamfer and point-map errors. Because the depth and point references come from Pi3 and the pose reference comes from the dataset cameras, these gains do not rely on DA3-GIANT as the evaluator. Figure 6 focuses on comparisons with external baselines. By showing multiple frames,

Table 6: **Generated-view 3D consistency**. VGGT is independent of all latent backbones; DA3-GIANT is reported as a backbone-overlap cross-check.

	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

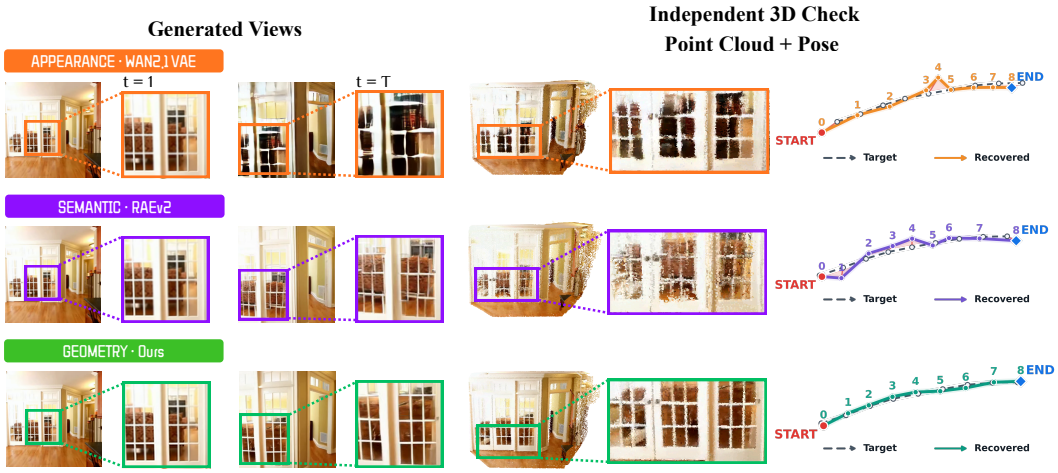


Figure 5: **Independent 3D check across latent families.** The appearance-native Wan2.1 VAE, semantic-native RAEv2, and geometry-native GAE are conditioned on the same reference image and prescribed camera trajectory. Generated views, detail crops, VGGT-reconstructed point clouds, and recovered camera trajectories highlight the impact of the latent representation on cross-view visual consistency and camera control.

generated point clouds, and recovered camera poses for the same scenes, it makes cross-frame drift and geometric distortions readily visible. GAE closely follows the prescribed camera trajectory and preserves the overall scene structure, while Gen3R and GLD exhibit larger pose deviations and more severe point-cloud distortions.

4.3.4 Additional qualitative capabilities

Long-rollout showcase. The controlled comparisons above use nine views. Figure 7 instead shows an 81-view rollout at 672×378, sampled from one reference view with 50 Euler steps and CFG $= 2$. We decode each sampled latent once into RGB and geometry. The point cloud simply

Table 7: **Geometry generation from sampled latents.** Depth and point maps are compared with Pi3 on the real target frames; camera poses are compared with the real trajectory.

	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	0.164	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**

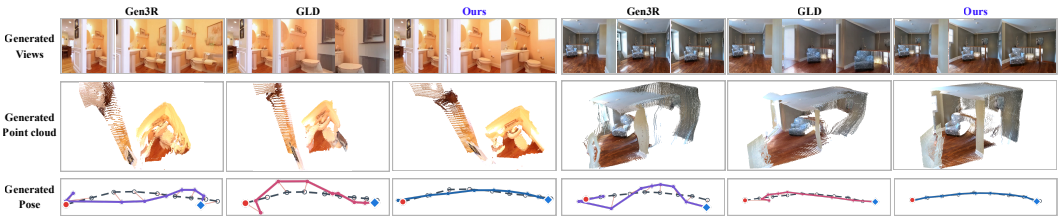


Figure 6: **Multi-frame RGB, geometry, and camera-pose comparison across RealEstate10K scenes.** Gen3R, GLD, and Ours receive the same reference image and prescribed camera trajectory. For each scene, early, middle, and final generated views are paired with the generated point cloud and recovered pose. The multi-scene comparison exposes cross-frame structure, depth-layer distortions, and trajectory drift that are less apparent in a single final view.

concatenates the 81 per-view depth maps after unprojection with cameras recovered from the decoded rays; it uses no cross-view fusion, external reconstruction, or test-time optimization. Detailed training settings and dataset information are provided in the supplementary material.

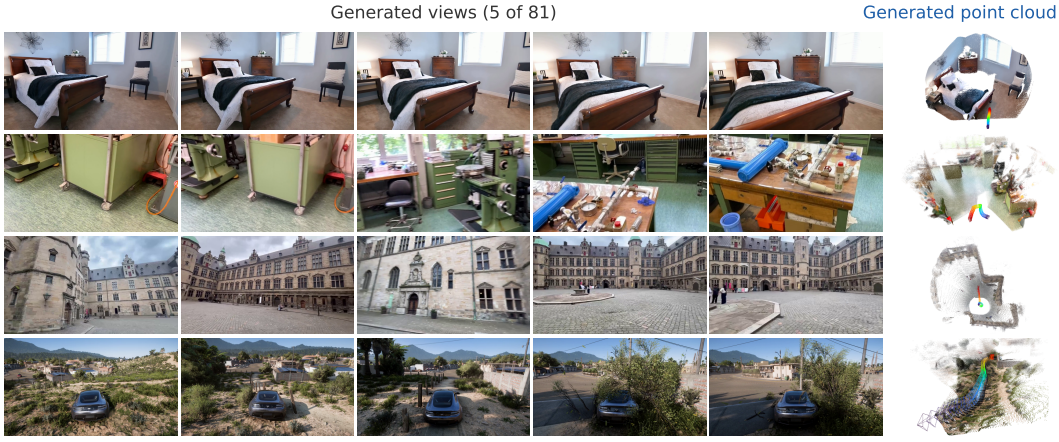


Figure 7: **Long-rollout showcase.** Five views from each 81-view rollout and the point cloud decoded from the same latents.

Text-conditioned generation. The same flow checkpoint used for the long-rollout showcase also supports direct text-conditioned RGB and geometry generation. Figure 8 pairs each generated image with depth decoded from the same latent for a compact set of diverse prompts. We treat these samples as a qualitative capability check rather than a standalone text-to-image benchmark.

4.4 Ablation Studies

We isolate representation shaping, codec design, and conditioning while fixing the flow architecture and training recipe. The representation study in Table 1 shows that token alignment improves transport but weakens relational structure, which $\mathcal{L}_{\text{struct}}$ restores.

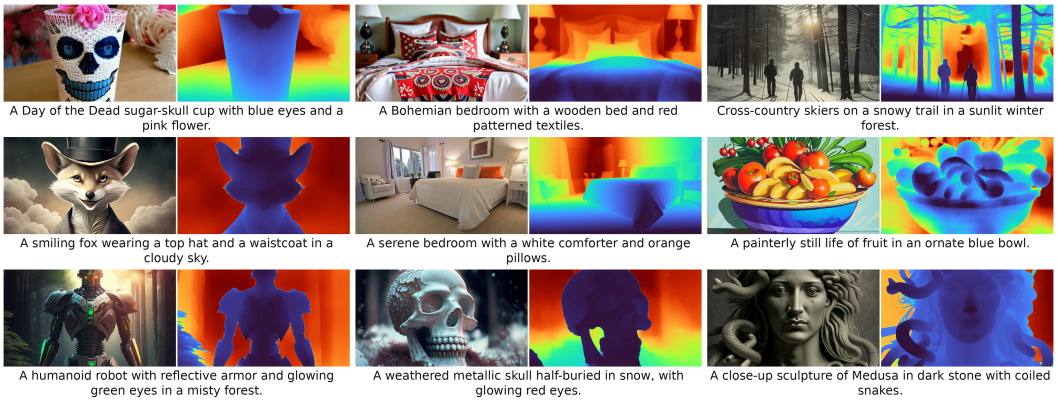


Figure 8: **Text-conditioned RGB and geometry generation.** Each pair shows an RGB sample and depth decoded from the same latent. These selected examples demonstrate capability rather than benchmark text-to-image quality.

Taken together, the complete $\mathcal{L}_{\text{repr}}$ improves the latent's transport, semantic organization, and relational structure. To verify that this improved latent quality transfers to downstream generation, we evaluate matched flow models on RealEstate10K in Table 8. Adding $\mathcal{L}_{\text{repr}}$ to DA3-GIANT consistently improves appearance, cross-view geometry, and camera recovery. At fixed latent width and objective, DA3-GIANT improves appearance and pose over DA3-LARGE, while LARGE retains a small MEt3R advantage. Thus representation shaping and backbone capacity provide complementary gains rather than optimizing only one metric.

Table 8: **Codec and conditioning ablations on RealEstate10K.** Diag. denotes ATE, point-cloud gap, scale-sensitive SE(3) ATE, or LPIPS by block. Except PSNR, lower is better.

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 $\mathcal{L}_{\text{repr}}$	**233.4**	**19.54**	0.1213	**0.0041**
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**
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**
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**

The same table isolates three conditioning choices. Clean reference tokens keep observed evidence outside the ODE state, removing the set-encoding train–inference mismatch and reducing the reference–target point-cloud gap. Metric ray conditioning preserves translation scale discarded by per-scene pose normalization, substantially improving scale-sensitive camera recovery as well as appearance and view consistency. Finally, text-to-image co-training supplies a broader appearance prior, improving FVD and LPIPS without sacrificing MEt3R. Together, these results support the complete conditioning recipe across appearance, geometry, and camera motion.

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.

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.

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.

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.

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*, 2026a.

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*, 2026b.

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

This section expands Sections 3.1–3.3 with codec, flow, conditioning, training, and evaluation details omitted from the main paper.

A.1 Geometry-native codec details

DA3-GIANT (Lin et al., 2026) is frozen throughout codec training. We collect its four output levels, normalize each level with fixed training-set channel statistics, reshape the patch tokens to their spatial grids, and concatenate the levels along the channel dimension. The codec preserves this patch grid and compresses only the channel dimension to $C_z \in \{64, 128\}$. The encoder and decoder are convolutional pyramids with spatial self-attention. During codec training we sample from the posterior; for latent statistics, flow training, and inference we use the posterior mean.

The codec reconstructs the normalized feature hierarchy, RGB, and the outputs of the frozen DA3 dense-prediction head. The geometry targets are therefore pseudo-targets read from the original DA3 features rather than dataset depth labels. The frozen head is never adapted to reconstructed features, so geometry quality must remain readable through the original DA3 interface. C-RADIO (Heinrich et al., 2025) and DINOv2 (Oquab et al., 2024) teachers, together with the projection head used by $\mathcal{L}_{\mathrm{repr}}$, are required only during codec training and are discarded afterward.

A.2 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

$$\bar{\mathbf{z}}_t = (1-t)\bar{\mathbf{z}}^{\mathrm{gt}} + t\bar{\mathbf{z}}^{(1)}, \qquad t \in [0,1], \tag{14}$$

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,

$$\hat{\bar{\mathbf{z}}}^{\mathrm{gt}} = \mathbf{f}_\psi(\bar{\mathbf{z}}_t, t, \mathcal{C}), \tag{15}$$

which is converted to velocity by

$$\hat{\mathbf{u}}_t = \frac{\bar{\mathbf{z}}_t - \hat{\bar{\mathbf{z}}}^{\mathrm{gt}}}{\max(t, t_\epsilon)}, \qquad t_\epsilon = 0.05. \tag{16}$$

We then optimize Eq. 13, 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.3 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:

$$\bar{\mathbf{z}}^{\mathrm{ref}} = \mathrm{Std}\Big[\mathrm{Enc}^{\mu}_{\phi}\big[\mathcal{P}\big(\mathcal{E}(\mathbf{I}^{\mathrm{ref}}_{1:K})\big)\big]\Big]. \tag{17}$$

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 target 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

$$\mathbf{d}, \qquad \hat{\mathbf{m}} = \frac{\mathbf{m}}{\|\mathbf{m}\| + \epsilon}, \qquad s = \log(\|\mathbf{m}\| + \epsilon). \tag{18}$$

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.4 Two-stage training and sampling

Stage 1 jointly trains $\mathcal{A}_{\phi}$ and $\mathcal{D}^{\mathrm{rgb}}_{\theta}$ on mixed single- and multi-view data while freezing $\mathcal{E}$ and $\mathcal{H}_{\mathrm{DPT}}$. Stage 2 freezes all Stage 1 modules and trains only the conditional flow model. In both

settings below, we sample Gaussian noise for all requested output slots, encode reference images once as clean conditions, integrate the velocity field from $t = 1$ to $t = 0$, and denormalize and decode the resulting latents through Eq. 2.

Controlled latent comparison. To isolate the choice of latent representation, all compared backends are inserted into the same DDT-style flow architecture and trained with a matched protocol. Training clips comprise nine 252×252 views sampled from an equal mixture of RealEstate10K (Zhou et al., 2018) and DL3DV (Ling et al., 2024), with one to four prefix views used as references. The transformer has 28 encoder blocks of width 768 and six decoder blocks of width 2048, uses clean-latent prediction with an eight-block auxiliary base head, and conditions on frozen Qwen3-0.6B (Yang et al., 2025) text features through cross-attention. The shared flow backbone contains approximately 0.93B parameters; the total parameter count varies slightly across backends because their latent shapes require different input and output projections. Each controlled model is trained on eight GPUs. SD-VAE (Rombach et al., 2022), the WAN2.1 VAE (Wan et al., 2025), and RAEV2 (Singh et al., 2026) natively produce 256×256 images; for evaluation, we resize their generated outputs to 252×252 to match the target resolution used by the other backends and the ground-truth clips. 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 is warmed up for 2k steps to 10^{-4} and then cosine-decayed 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, and a single-view text-to-image batch is interleaved every three multi-view updates. For quantitative inference, every backend uses the same nine-view setting with one clean reference, 50 Euler steps, and a CFG scale of 2. The frame interval is 10 for RealEstate10K and 1 for DL3DV, and evaluation uses the shared held-out sets described below. Thus, the flow architecture, data mixture, optimization schedule, and sampling protocol are held fixed; only the latent encoder and decoder, together with the input and output projections required by each latent shape, are changed. For every backend, the same latent normalization statistics used during training are also applied at inference; architecture- or conditioning-specific changes are made only in the corresponding ablations.

Final model. The qualitative final model is trained separately for longer, higher-resolution generation and is not included in the controlled ranking. It uses a frozen space–time codec built on DA3-GIANT features at 672×378 and trains the flow on 81-view clips with one to four prefix reference views. Its multi-domain training mixture spans 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), and several internally curated datasets. We train the 0.93B-parameter flow transformer on 80 GPUs in BF16 using AdamW with a base learning rate of 5×10^{-5}, 500 warmup steps, cosine decay to 10^{-6} by 60k steps, gradient clipping at 1.0, and an EMA decay of 0.9995. This parameter count covers 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 co-training stream. For the showcased results, all 81 requested views are sampled jointly from one reference using 50 Euler steps and a CFG scale of 2, while camera conditioning is retained in the unconditional CFG branch.

A.5 Controlled evaluation protocol

The main comparison uses shared held-out pools of 64 scenes from RealEstate10K and DL3DV. Each example contains nine views at 252^2 with one reference view; all target views are sampled jointly using 50 Euler steps and CFG $= 2$. FID and FVD measure frame- and clip-level distributional quality. LPIPS, PSNR, and SSIM use the paired target frames under the prescribed camera trajectory. Blue table cells rank only the seven latent variants trained with the shared flow recipe; GLD and Gen3R are shown as external system-level references.

For camera metrics, VGGT or DA3-GIANT reconstructs the complete real and generated nine-view clips. The recovered trajectories are aligned by Umeyama Sim(3) before ATE, relative translation error (RPEt), and relative rotation error (RPEr) are computed. VGGT is independent of all evaluated latent backbones; DA3-GIANT is included only as a backbone-overlap cross-check.

The reported reprojection metric (Jang et al., 2026) uses the recovered world points after Sim(3) alignment to the dataset camera frame. High-confidence points are transferred between views and projected back to the source view; the round-trip 2D pixel error is divided by the image diagonal. We report Reproj@50, which retains the top half of points by estimator confidence. MEt3R is computed with the official implementation (Asim et al., 2025). For every pair of generated views it recovers pairwise geometry, upsamples visual features with FeatUp, renders the features into the opposite view with visibility and z-buffer occlusion masks, and computes the symmetric cosine-feature disagreement. We average over all $\binom{V}{2}$ pairs. Both Reproj and MEt3R are lower-is-better.

For native geometry evaluation, depth and point maps decoded from the latent are compared with Pi3 (Wang et al., 2026c) applied to the real target frames. Depth uses per-view median-scale alignment; point clouds use the Sim(3) estimated from the camera trajectories. Camera ATE and RPEr are evaluated against the input dataset trajectory. This protocol separates native geometry readout from the post-hoc RGB reconstruction used for the consistency table.

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. (2026a), 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); Wang et al. (2026a), 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 $\mathbf{\Sigma}_{\mathrm{within}}$ 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 $\mathbf{\Sigma}_{\mathrm{total}}$ be the covariance over the complete evaluation set. We compute

$$\rho = \frac{\mathrm{tr}(\mathbf{\Sigma}_{\mathrm{within}})}{\mathrm{tr}(\mathbf{\Sigma}_{\mathrm{total}})}. \tag{19}$$

This normalized variance ratio accompanies the velocity-ambiguity analysis of Zhong et al. (2026a). 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

$$\kappa = \frac{\lambda_1}{\lambda_C}, \qquad \mathrm{effrank} = \frac{\left(\sum_i \lambda_i\right)^2}{\sum_i \lambda_i^2}. \tag{20}$$

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. (2026a), 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:

$$\mathrm{LDS} = \mathbb{E}_{(i,j)\in\mathcal{N}}[\langle \hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j \rangle] - \mathbb{E}_{(i,j)\in\mathcal{F}}[\langle \hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j \rangle]. \tag{21}$$

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:

$$\mathrm{SRSS} = \mathbb{E}_{i\in\mathrm{FG}}\left[\mathbb{E}_{j\in\mathcal{P}(i)}\langle \hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j \rangle - \mathbb{E}_{k\in\mathcal{B}(i)}\langle \hat{\mathbf{z}}_i, \hat{\mathbf{z}}_k \rangle\right], \tag{22}$$

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. (2026b); 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,

$$\mathrm{xLNC}^* = \frac{a - a_{\mathrm{chance}}}{1 - a_{\mathrm{chance}}}, \tag{23}$$

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 qualitative comparisons

C.1 Independent 3D check

Figure 9 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 plausibility from whether the generated frames jointly support a coherent 3D scene and camera path.

C.2 RGB comparisons

Figures 10 and 11 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.

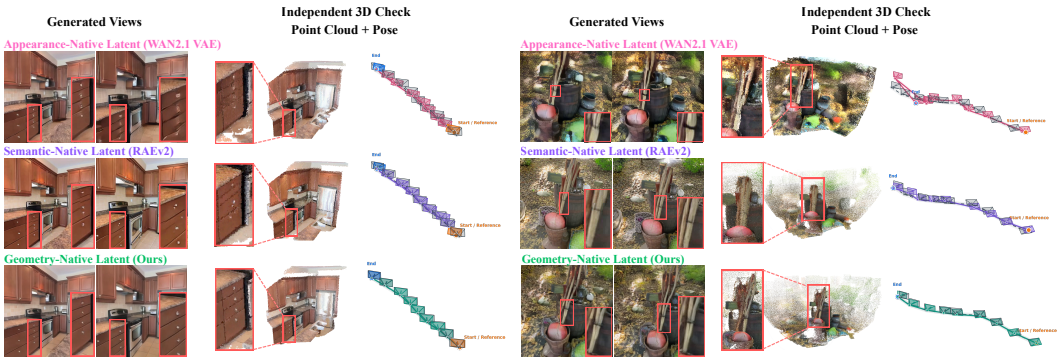


Figure 9: **Independent 3D check of generated views.** The appearance-native WAN2.1 VAE, semantic-native RAEV2, and our geometry-native latent receive the same reference and camera controls. The right-hand panels show point clouds and camera trajectories reconstructed from the generated RGB clips by an independent geometry model VGGT.



Figure 10: **RealEstate10K comparison.** All methods receive the same reference image and target camera trajectory, and the eight remaining views are generated jointly.

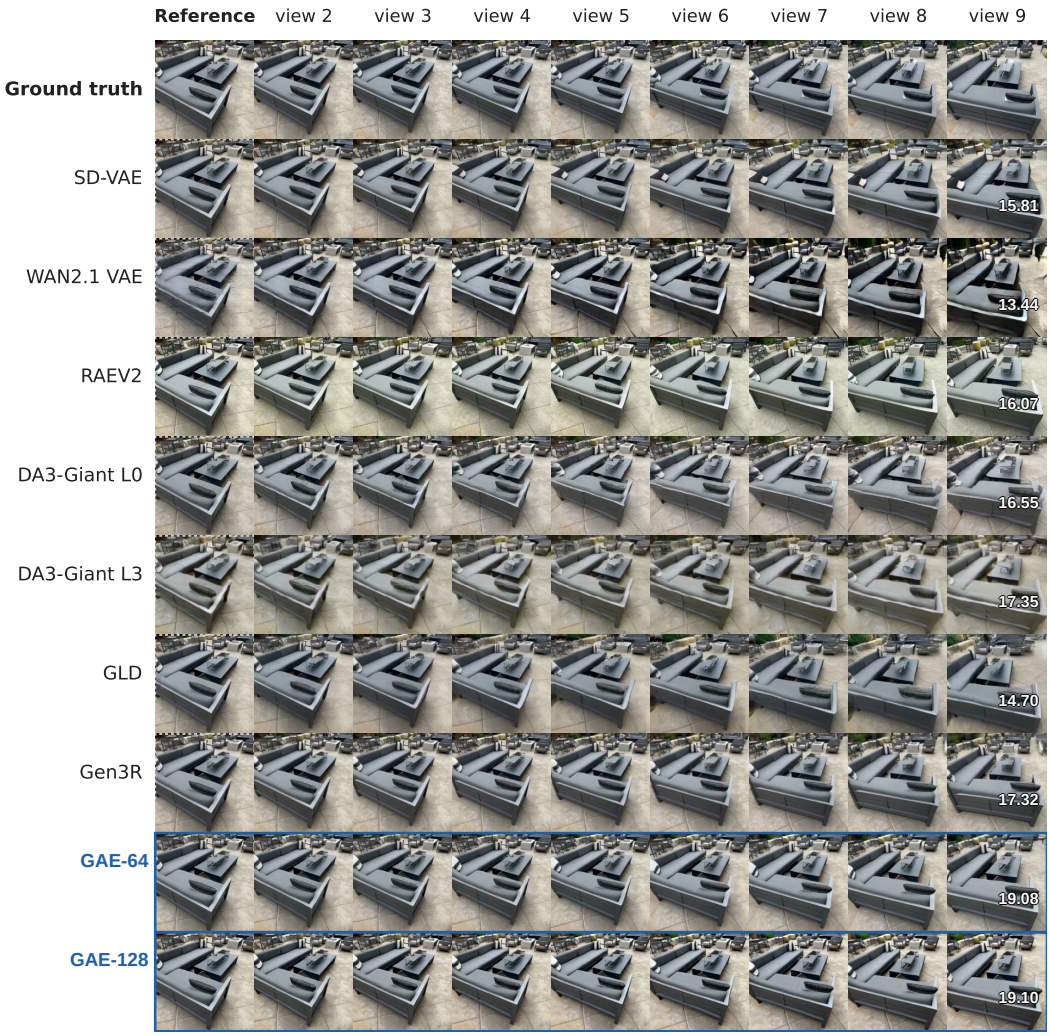


Figure 11: **DL3DV comparison.** All methods receive the same reference image and target camera trajectory, and the eight remaining views are generated jointly.