# How Much 3D Do Video Foundation Models Encode?

Zixuan Huang[1*] Xiang Li[1*] Zhaoyang Lv[2] James M. Rehg[1]
[1]University of Illinois at Urbana-Champaign, [2]Impossible, Inc.

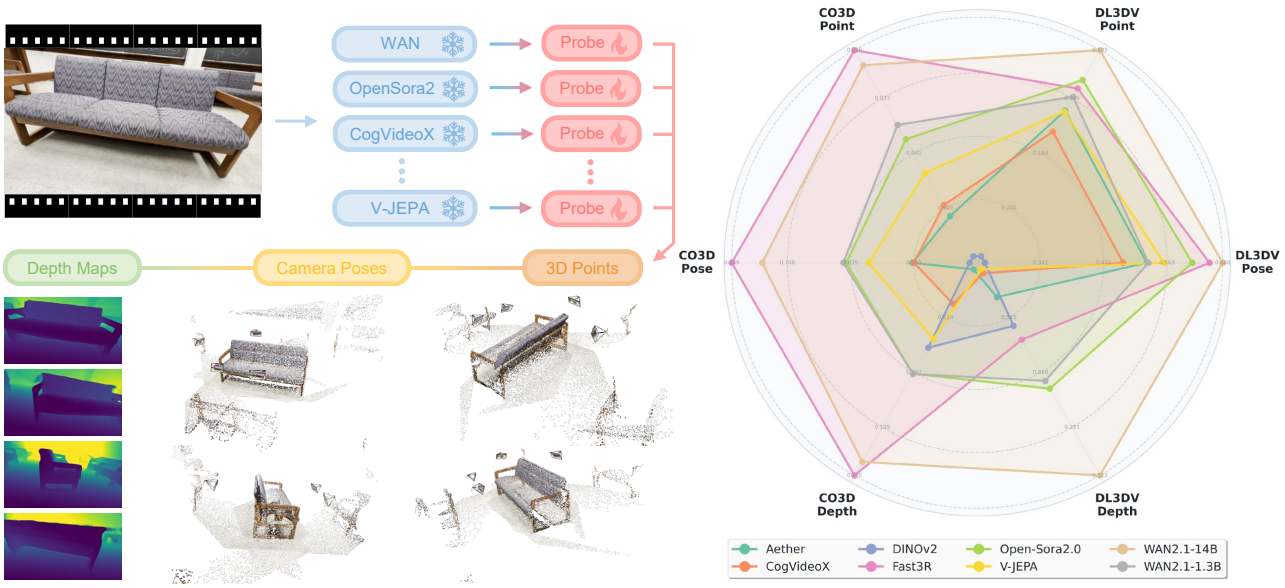



Figure 1. We study the emergence of 3D in video foundation models by probing their features with 3D reconstruction tasks. Our study reveals state-of-the-art video generators develop strong 3D understanding even compared to 3D experts, despite only trained on 2D data.

## Abstract

*Videos are continuous 2D projections of 3D worlds. After training on large video data, will global 3D understanding naturally emerge? We study this by quantifying the 3D understanding of existing Video Foundation Models (VidFMs) pretrained on vast video data. We propose the first model-agnostic framework that measures the 3D awareness of various VidFMs by estimating multiple 3D properties from their features via shallow read-outs. Our study presents meaningful findings regarding the 3D awareness of VidFMs on multiple axes. In particular, we show that state-of-the-art video generation models exhibit a strong understanding of 3D objects and scenes, despite not being trained on any 3D data. Such understanding can even surpass that of large expert models specifically trained for 3D tasks. Our findings, together with the 3D benchmarking of major VidFMs, provide valuable observations for building scalable 3D models.*

## 1. Introduction

Recovering 3D structure from 2D visual observations is a long-standing research problem in computer vision, with broad applications in AR/VR and embodied AI. Despite significant progress, the availability of high-quality 3D data at scale remains the bottleneck for current data-driven approaches. This fundamentally limits the scaling of 3D foundation models and makes it questionable whether we can learn truly generalizable models primarily from 3D data.

Compared to native 3D assets, videos are much easier to acquire at scale, with multiple large curated datasets already available [1, 4, 8, 35]. The diversity and complexity of video data, with the fact that videos are 2D projections of 3D worlds, lead to a promising pathway for scalable 3D learning. Recent works study how to utilize video models for 3D, either by adding 3D control [3, 17, 18, 44, 56] or by producing 3D caches/estimations [15, 21, 22, 25, 29, 32, 33, 41, 47, 55, 60, 63, 64] along with the original frame synthesis target. These works suggest that video priors are useful for 3D, but 3D-inconsistency artifacts, the requirement of 3D fine-tuning, and task-specific engineering leave

[*]Both authors contributed equally to this work.

it unclear whether video data alone can induce strong 3D awareness in a general-purpose setting. These confounds motivate a direct, model-agnostic evaluation.

In this paper, we present the first model-agnostic framework to probe the 3D awareness of video foundation models (VidFMs) pretrained on large-scale video data. We ask whether VidFMs develop internal representations of 3D structure and ego-motion and, if so, how strong and practically useful these representations are. We operationalize this question along four axes: 1) **Extent**: how does the 3D awareness of VidFMs compare to that of image models or specialized 3D models? 2) **Factor**: Which factors impact 3D awareness? Here, we focus on the effects of temporal reasoning, 3D finetuning and model scaling. 3) **Localization**: In which network layers, and at which timesteps in diffusion models, is this 3D information most concentrated? 4) **Implication**: Under limited resources (3D data and compute), can VidFM features be practically useful for 3D reconstruction tasks?

We posit that if a video model understands 3D worlds, it should be feasible to extract accurate 3D properties using shallow readout modules in a feedforward manner, without any post-optimization or fine-tuning of the base model. Unlike prior works that evaluate image models using depth and cross-view consistency [12], or per-scene optimization with off-the-shelf initialization [9], our proposed shallow feedforward readouts that estimate different 3D attributes from VidFMs' feature space are a more direct probe of globally consistent 3D properties from pretrained video models.

Specifically, we extract frozen spatialtemporal features from VidFMs, and design a probe model that predicts 3D points, camera poses and depth maps from these features. The probe model is a shallow VGGT [51]-like transformer, consisting of four alternating-attention layers and three read-out heads: two dense prediction heads for 3D points and depth maps, and one camera head. We train the probe model on top of various video features, including features extracted from self-supervised video models and video generation models of different performance and sizes. We measure the performance of point, camera and depth reconstruction as indicators of 3D awareness.

Our study leads to the following novel findings:

- **Extent**: Frontier video generation models exhibit great understanding of 3D objects and scenes, and can be close to or better than models trained with 3D data;
- **Factor #1:** Temporal reasoning is critical to the formation of global 3D understanding;
- **Factor #2:** Finetuning video generation models with 3D objectives improves 3D awareness on in-domain data, but may hurt generalization beyond data domains;
- **Factor #3:** Model scaling leads to mixed impact on 3D awareness, with WAN2.1-14B performing significantly better than WAN2.1-1.3B, while CogVideoX-5B is slightly worse than CogVideoX-2B.
- **Localization:** The best layer and timestep to extract 3D-aware features are surprisingly consistent across all tested video diffusion models: mid-layer features with early-but-not-first timesteps lead to the highest 3D awareness.
- **Implication:** We implement and train a VGGT model using frozen VidFM (WAN2.1-14B) features. On CO3Dv2 and DL3DV, it consistently outperforms the standard DINO-based VGGT, indicating VidFM features may be better suited for 3D reconstruction under limited 3D data.

In summary, we conduct the first systematic model-agnostic evaluation on the 3D awareness of VidFMs and conclude with meaningful findings across multiple axes that the prior work has not surfaced well. Our findings are based on a comprehensive benchmark that compares 3D awareness of various video models, which can benefit the development of VidFMs by enabling the evaluation of their emerging 3D properties in a general-purpose way.

## 2. Related Works

**Video foundation models (VidFMs)** are deep models trained on massive video data that achieve strong performance across various downstream tasks. Early works in the field employ self-supervised learning [5, 48, 52] or contrastive pretraining [57] paradigms to learn discriminative representations of video inputs. More recently, with the tremendous success of diffusion models, there is growing interest in learning generative models that exploit large-scale video priors. Such models achieve strong video synthesis results, exemplified by Sora [7] and numerous follow-ups [20, 26, 38, 50, 59]. While these models demonstrate very strong pixel synthesis performance, their internal representations are not well understood. In our work, we present a comprehensive study to understand how much 3D understanding these models possess.

**3D from Video** is a fundamental task in computer vision, traditionally tackled via Structure from Motion (SfM) [16, 37, 43] and Multi-view Stereo (MVS) [14] techniques. These classical methods rely on feature matching and cannot handle challenging cases (e.g. textureless regions, repetitive patterns, or wide baselines) well. Recent work instead turns to data-driven methods and propose strong feedforward models for direct 3D prediction. This wave of research begins with pairwise models [27, 53] and further extends to multi-view settings that improve accuracy and efficiency for large scenes [46, 51, 58]. These methods achieve better performance and robustness than classical approaches, yet it remains challenging for them to further scale up and to generalize to dynamic or real-world cluttered scenes given their reliance on annotated data.

To address this limitation, recent work considers using video priors from large video generative models. Exist-

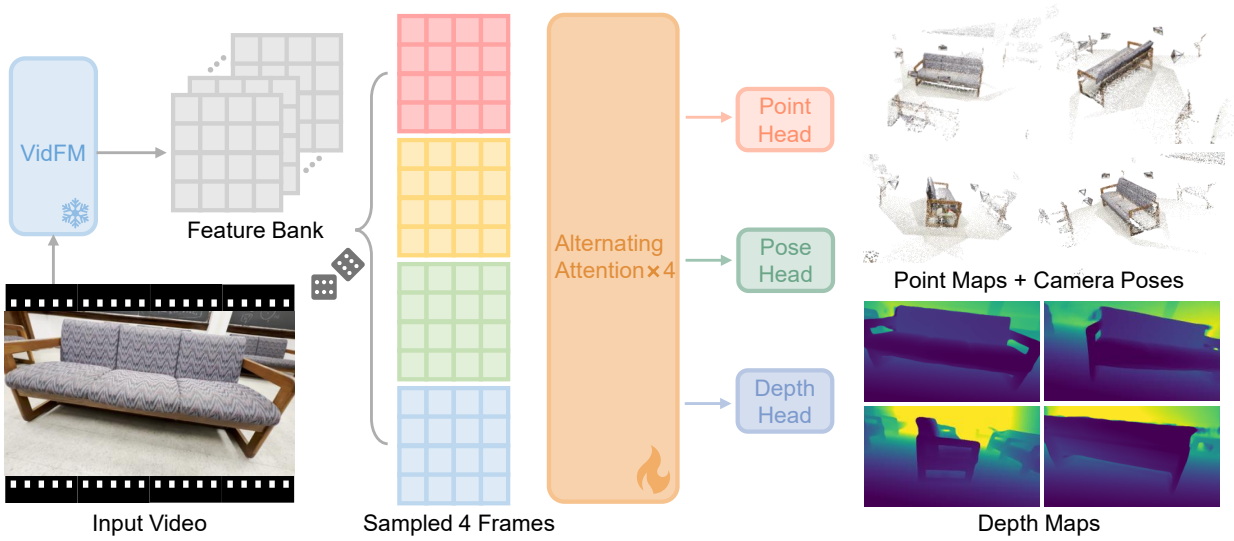



Figure 2. **Overview of the Probe.** We extract video features using various video foundation models and keep the features frozen. We sample four frames from the original video clip and fetch the corresponding feature maps from the video features. We train the probe by taking these spatial features as input, and task the probe to estimate point maps, depth maps and camera poses. Our probe model consists of a shallow transformer and three readout heads. We measure the estimation errors as the main indicators of 3D awareness.

ing work directly finetunes video generation models on 3D data, to achieve 3D control [3, 17, 18, 56] or to simultaneously output 3D attributes [21, 25, 32, 33, 47, 63]. Despite progress, small-scale finetuned video models still exhibit major artifacts of 3D inconsistency, especially on the data distinct from the fine-tuning data. To mitigate this, prior and concurrent works consider the usage of 1) explicit 3D memory [15, 22, 41, 55, 60]; 2) post 3D optimization [10, 31, 44]; or 3) feedforward models on generations [29, 64] to enhance 3D consistency. These efforts demonstrate the utility of video priors in sparse/single-view regimes by using video generators as frame extrapolators, yet the extent of 3D information already encoded in base video models remains unclear in quantitative terms. Answering this question requires a model-agnostic framework that evaluates various models with quantitative metrics, which is what we pursue in this work.

**Quantifying 3D awareness of visual foundation models** is an important research direction which helps the understanding of learned features and guides the development of scalable 3D world models. Early work in this area studies large image models and demonstrates the emergence of semantic correspondence in the feature space [2, 19, 54, 62]. To directly quantify 3D understanding, more recent works use 3D semantic (e.g. 3D-VQA, multi-view object recognition, semantic segmentation) or coarse estimation (e.g. relative depth) tasks to test the 3D understanding of visual foundation models [6, 13, 28, 34, 39, 42, 61, 66]. Other work, such as VBench [23, 24, 65] and WorldScore [11], focuses on benchmarking video generators and evaluates the 3D consistency of generated videos using off-the-shelf priors. Instead of using coarse-grained or model-specific evaluation, Probe3D [12] and Feat2GS [9] consider dense probes to evaluate the 3D awareness of deep models and are more relevant to our work. However, their probes target image models, and their evaluation mainly focuses on depth/normal or multi-view consistency. In our study, we present a comprehensive study on video models by directly probing them with 3D attributes. We additionally show that indirect probes such as depth and multi-view consistency are not necessarily the best metrics to evaluate 3D awareness across different families of models.

## 3. Approach

We probe the 3D awareness of various VidFMs in this work. We define 3D awareness as the extent to which the underlying 3D structure and ego-motion can be recovered using frozen features extracted from 2D video. Under a fixed probe capacity and training set, stronger 3D awareness implies that a shallow readout attains a lower reconstruction error. Our method has two stages. First, we extract per-frame spatial feature maps by running each VidFM on video clips while keeping the VidFM parameters fixed. Second, we train a lightweight feedforward probe on these features to predict, for each sampled frame, (i) a dense 3D point map that represents the 3D coordinates of visible pixels in the coordinate system of the first frame, (ii) a dense depth map at a consistent scale with other frames, and (iii) the camera pose of each frame relative to the first frame; only the probe

is optimized, not the VidFM. We primarily evaluate popular frontier video generation models [38, 47, 50, 59], and also include a self-supervised video encoder, V-JEPA [5], and two control models, DINOv2 [36] and Fast3R [58], to contextualize our results.

### 3.1. Feature Extraction

Given a video $\mathbf{V} \in \mathbb{R}^{T_v \times 3 \times H_v \times W_v}$, we run each VidFM in frozen mode and extract, for every frame at $t$, a spatial feature map $\mathbf{F}_t \in \mathbb{R}^{C \times H_f \times W_f}$. For diffusion-based video generators, we extract features similar to DIFT [45]: we choose a denoising timestep $\tau$, add noise, perform a single denoising step, and read hidden activations from a specified network layer as features. We use an empty text embedding, and for image-to-video models we condition on the first frame. The layer index and $\tau$ are treated as hyperparameters and are fixed across experiments. For V-JEPA, DINOv2 and Fast3R, we run a standard forward pass and take the last-layer spatial features, which we empirically find to be the best-performing layer.

Different VidFMs often operate on different clip lengths. Several models we investigate are trained on fixed small context windows. To test them on longer videos, we split the input video $\mathbf{V}$ into short chunks for these models, by subsampling at fixed strides from beginning. We prepend the first frame to each chunk, so all chunks share the same first-frame reference. We also maintain a frame-to-feature index map $\pi(t)$ that records, for each raw frame at $t$, the corresponding chunk and local index. At probe time, based on the input frame indices $\{t_i\}_{i=1}^{S}$ and $\pi(t)$, we can gather the corresponding features $\{\mathbf{F}_{t_i}\}_{i=1}^{S}$ for all $S$ input frames.

### 3.2. 3D Awareness Probe

**Architecture.** We use a shallow transformer probe with alternating attention and three readout heads. For each input video, we take $S$=4 frames: the first video frame as the reference and three additional frames sampled with a minimum temporal gap of 5 frames. From the corresponding feature maps $\{\mathbf{F}_{t_i}\}_{i=1}^{4}$, we obtain per-frame tokens and apply four blocks of alternating-attention on top. Each alternating-attention block consists of a frame attention that mixes tokens within each frame and a global attention that mixes tokens across frames; this mirrors the VGGT design [51] but is much shallower. Three heads decode 3D outputs: two DPT heads produce dense point maps $\hat{\mathbf{X}}_{t_i} \in \mathbb{R}^{H_v \times W_v \times 3}$ (in the coordinate system of the first frame) and depth maps $\hat{\mathbf{D}}_{t_i} \in \mathbb{R}^{H_v \times W_v}$, similar to Probe3D [12] and VGGT. A camera head predicts the pose of each frame relative to the first frame.

**Loss.** We train the probe with a multi-task objective similar to VGGT:

$$\mathcal{L} = \lambda_{\text{pmap}}\mathcal{L}_{\text{pmap}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{cam}}\mathcal{L}_{\text{cam}},$$

with $\lambda_{\text{pmap}} = \lambda_{\text{depth}} = \lambda_{\text{cam}} = 1$ unless otherwise stated. For $\mathcal{L}_{\text{pmap}}$ and $\mathcal{L}_{\text{depth}}$, we use confidence-weighted $\ell_2$ losses between predicted point/depth maps and groundtruth point/depth maps. Note that the groundtruth scenes are normalized before loss calculation to remove scale ambiguity. For camera poses, we use a Huber loss between the predicted poses and groundtruth poses.

## 4. Experiments

### 4.1. Evaluation

**Datasets.** We perform experiments on CO3Dv2 [40] and DL3DV [30]. CO3Dv2 is an object-centric dataset consisting of turntable-type videos. We curate CO3Dv2 by filtering out sequences with heavy truncation or portrait-oriented videos that prevent forming border-less horizontal object-centric crops. The filtered split contains 11k videos in total. From each video, we sample consecutive frames as inputs to the feature extraction pipeline and use features from the first 76 frames during training. We adopt a 9:1 train test split at the video level and additionally create an ablation subset of 10 diverse categories (2.7k videos total) for ablation study. On the other hand, DL3DV contains large, cluttered scenes and is generally more challenging than CO3D. We use the first 6k splits and a 9:1 train test split by video. For both datasets, we run VGGT [51] to generate ground truth for every frame: dense point maps, depth, and camera poses. For point and depth maps, we also save the confidence maps, which are used in our losses. Unlike probe time where we only sample 4 frames from the video clips, we use all frames to generate the groundtruth. This leads to much more accurate annotations than the groundtruth originally provided by the datasets.

**Metrics.** The main metrics to evaluate 3D awareness are errors of point map, pose, and depth predictions. For point maps, we first normalize each scene to remove global scale, then align the predicted and ground-truth point clouds with the Umeyama algorithm [49], and report the mean $\ell_2$ error. For depth, we report the mean $\ell_2$ error after the same scene normalization. For camera pose, we compute relative pose errors over all frame pairs: rotation error $e_R$ is the geodesic angle on $SO(3)$, and translation error $e_T$ is the angle between translation directions. Accuracy at threshold $\theta$ is defined jointly as $\Pr[\max(e_R, e_T) \leq \theta]$, i.e., both rotation and translation must be within $\theta$. Following [51], we report AUC@$\Theta$, the area under this joint accuracy curve as $\theta$ sweeps from $0^\circ$ to $\Theta^\circ$ (e.g., $\Theta \in \{5, 30\}$).

**VidFM Baselines.** We evaluate various VidFM baselines, including video generators and self-supervised encoders. For generative models, WAN [50] and Open-Sora2.0 [38] are the strongest open-weight generators we probe. We also

| Probed Feature | CO3Dv2 | | | | DL3DV | | | |
|---|---|---|---|---|---|---|---|---|
| | Point Err(↓) | Depth Err(↓) | AUC@5 (↑) | AUC@30 (↑) | Point Err(↓) | Depth Err(↓) | AUC@5 (↑) | AUC@30 (↑) |
| DINOv2 [36] | 0.559 | 0.209 | 0.051 | 0.508 | 2.814 | 0.534 | 0.013 | 0.245 |
| V-JEPA [5] | 0.439 | 0.214 | 0.076 | 0.619 | 1.576 | 0.613 | 0.076 | 0.558 |
| CogVideoX [59] | 0.485 | 0.231 | 0.051 | 0.569 | 1.748 | 0.608 | 0.061 | 0.486 |
| Aether [47] | 0.501 | 0.249 | 0.054 | 0.571 | 1.566 | 0.574 | 0.067 | 0.527 |
| Open-Sora2.0 [38] | 0.391 | 0.196 | 0.096 | 0.643 | 1.306 | 0.445 | 0.115 | 0.607 |
| WAN2.1-14B [50] | 0.284 | 0.151 | 0.200 | 0.736 | **1.051** | **0.323** | **0.136** | **0.660** |
| Fast3R [58] | **0.262** | **0.145** | **0.272** | **0.769** | 1.379 | 0.514 | 0.134 | 0.637 |

Table 1. **3D awareness benchmark results on CO3Dv2 and DL3DV.** We evaluate **video generators**, **self-supervised video encoders**, **3D experts**, and **per-frame image models**. State-of-the-art video generators such as WAN2.1-14B and Open-Sora2.0 exhibit strong 3D awareness and outperform Fast3R on scenes. Point map errors have been multiplied by 10 for clarity.


Figure 3. **CO3Dv2 qualitative results.** For each scene, we show input frames and the unprojected 3D points prediction. Fast3R, WAN2.1-14B, and Open-Sora2.0 best preserve intricate details (e.g., the truck's gripper) and reconstruct the overall structure.

probe CogVideoX [59], an earlier work than WAN/Open-Sora2.0, and Aether [47], which fine-tunes CogVideoX with 3D-aware objectives. All generative models here are latent diffusion models, which consist of a VAE that maps between raw videos and latents, and a denoiser that denoises the latents. For self-supervised models, we evaluate V-JEPA [5], a large self-supervised video encoder.

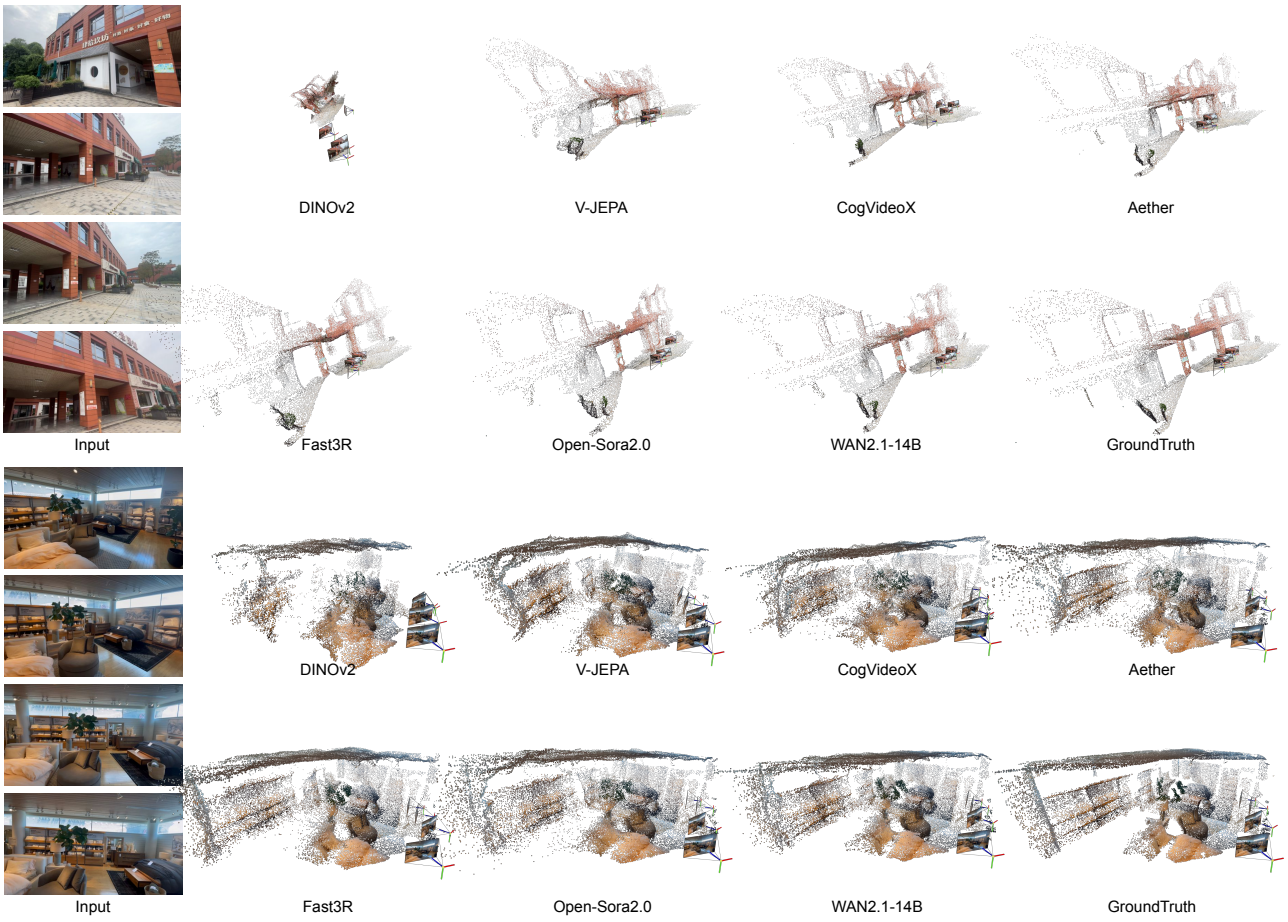


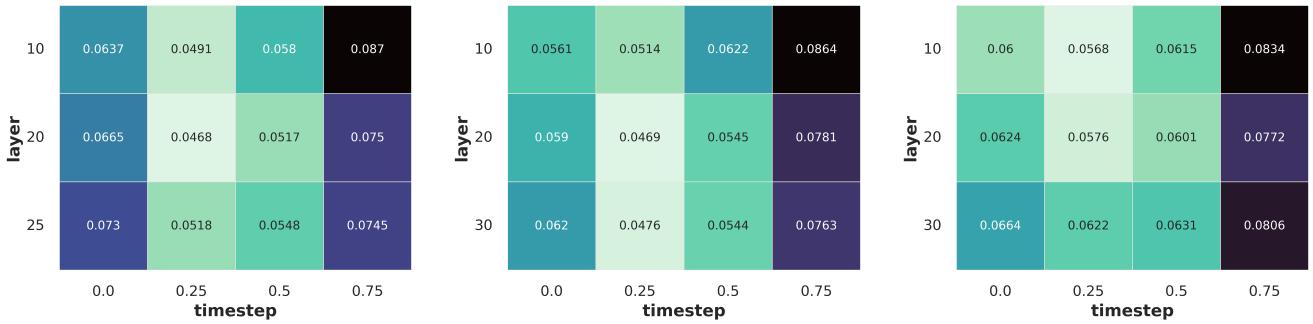

Figure 4. **DL3DV qualitative results.** On this more challenging dataset, DINOv2 sometimes fails catastrophically. Top video generators often retain coherent geometry, where WAN2.1-14B produces the sharpest and most accurate point clouds overall.

**Control groups.** A potential risk in our probe is that some 3D properties can already be estimated from raw videos. While relative rankings among VidFMs are still informative, if their probe performance is on par with direct 3D estimation from image features, the practical meaning of such rankings is compromised. To contextualize the results, we include two control baselines. *Per-frame Image control:* we probe DINOv2 features extracted from each frame of the video. Since the features are extracted in isolation, any global 3D understanding of the video (e.g. 3D points in a common coordinate frame) must be induced by the probe rather than supplied by the backbone itself. To make the task well-posed, we append a reference-frame indicator token that marks the first frame; all losses, schedules, and hyperparameters mirror the VidFM setting. *Native 3D control:* we probe features from Fast3R [58], a state-of-the-art model trained directly to predict 3D point maps from multi-view images. Because this model is optimized for the same target as our probe, probing it under the same probe architecture and supervision provides a strong reference. Meanwhile, CO3D is part of Fast3R's training sets but not DL3DV; this allows us to study the generalization behavior of its features. Together, the per-frame control (lower reference) and native-3D control (upper reference) contextualize VidFM results and ground our conclusions.

### 4.2. 3D Awareness Benchmark

We now present and analyze our results along the four axes defined earlier. We additionally analyze the relationship between our direct 3D probe and the multi-view evaluation from prior works in the supplementary material.

**Extent: how does the 3D awareness of VidFMs compare to that of image models or specialized 3D models?** *Strong video diffusion models exhibit great 3D awareness even compared to 3D experts.* On CO3Dv2, WAN2.1-14B is second only to Fast3R across all metrics (e.g., Point 0.284 vs. 0.262, Depth 0.151 vs. 0.145, AUC@30 0.736 vs. 0.769, Table 1 (left)). On DL3DV, which lies outside Fast3R's training distribution, WAN2.1-14B surpasses




(a) WAN2.1 (b) Open-Sora2.0 (c) CogVideoX

Figure 5. **Layer–timestep ablations.** We show point-map error (lower is better) on the ablation data when probing different diffusion layers and denoising time steps. Best results are consistently from mid layers and early-but-not-first time steps.

Fast3R on all metrics (Point 1.051 vs. 1.379, Depth 0.323 vs. 0.514, AUC@30 0.660 vs. 0.637, Table 1 (right)). Open-Sora2.0 is consistently strong as well, supporting the observation that state-of-the-art video generators yield features with universally strong 3D awareness across data domains.

**Factor #1: how does temporal reasoning impact 3D awareness?** *Effective temporal reasoning is critical to global 3D understanding.* Per-frame DINOv2 attains competitive depth on CO3Dv2 (0.209) but is significantly worse on global 3D understanding (Point 0.559, AUC@30 0.508) than all video models, including the self-supervised V-JEPA (Point 0.439, AUC@30 0.619). The key difference between image and video models is that the latter allows information exchange along the time axis. This gap in global 3D estimation widens on DL3DV (DINOv2 Point 2.814, AUC@30 0.245 vs. V-JEPA Point 1.576, AUC@30 0.558), whereas the depth estimation of DINOv2 remains competitive. The radar plots in Figure 1 mirror this pattern: methods with explicit temporal reasoning produce polygons that expand along *Point* and *Pose*, not just *Depth*.

**Factor #2: how does 3D finetuning impact 3D awareness?** *3D-aware fine-tuning does not always benefit.* Aether (fine-tuned from CogVideoX with 3D-aware objectives and conditions) indeed improves 3D awareness over CogVideoX on DL3DV (Point 1.566 vs. 1.748, Depth 0.574 vs. 0.608, AUC@30 0.527 vs. 0.486; Table 1 (right)). However, on object-centric data, it is slightly worse than its base model (Point 0.501 vs. 0.485, Depth 0.249 vs. 0.231; Table 1 (left)). Such discrepancy likely relates to the training data of Aether, which are mostly large synthetic scenes from games/simulators. This result suggests that 3D generative fine-tuning does have the potential to significantly improve 3D awareness, but how to avoid degraded generalization remains an interesting research direction.

**Qualitative analysis.** Figure 3 and Figure 4 align well with the ranking of 3D awareness in the quantitative tables. On CO3Dv2, Fast3R, WAN2.1-14B, and Open-Sora2.0 yield the most faithful and consistent reconstructions: thin structures and fine details (e.g., the gripper of the truck, the armrests and legs of the chair) remain sharp after unprojection, whereas other models exhibit noisy reconstructions and clear artifacts due to inconsistencies. On DL3DV, DINOv2 can fail catastrophically (e.g. the first building example, where the first view and the remaining views scarcely overlap), while top video generators often produce coherent point clouds. Overall, WAN2.1-14B delivers the sharpest and most accurate reconstructions, matching its lead in Table 1 (right). Similarly, Aether demonstrates a clear improvement over CogVideoX qualitatively. Across both datasets, most failure cases concentrate around object boundaries.

### 4.3. Ablations

**Factor #3: how does model size impact 3D awareness?** On the ablation set, we further study whether models at larger scales produce more 3D-aware features. Given the limited availability of open-source checkpoints, we study the scaling of WAN and CogVideoX. For WAN, scaling the model from 1.3B to 14B parameters significantly reduces point-map error from $0.0468$ to $0.0360$ on the ablation set (relatively $-23\%$). In contrast, CogVideoX slightly worsens in 3D awareness as parameters increase from 2B to 5B (from $0.0576$ to $0.0590$, relatively $+2\%$). This result suggests that parameter count alone does not guarantee stronger 3D awareness. We hypothesize that additional training data likely plays an important role here [1].

**Localization: in which network layers, and at which timesteps in diffusion models, is 3D information most**

[1] Unlike CogVideoX that mainly scales the architecture, WAN includes additional high-quality high-resolution data when scaling up.

| Method | CO3Dv2 | | | | DL3DV | | | |
|---|---|---|---|---|---|---|---|---|
| | Point Err(↓) | Depth Err(↓) | AUC@5 (↑) | AUC@30 (↑) | Point Err(↓) | Depth Err(↓) | AUC@5 (↑) | AUC@30 (↑) |
| Original VGGT [51] | 0.476 | 0.205 | 0.076 | 0.565 | 2.751 | 0.518 | 0.058 | 0.363 |
| VidFM-VGGT | **0.289** | **0.145** | **0.178** | **0.718** | **1.034** | **0.319** | **0.183** | **0.686** |

Table 2. **VidFM vs. DINO in VGGT.** Comparison between VGGT [51] (DINO features) and our VidFM-based variant using frozen WAN2.1-14B features on CO3Dv2 and DL3DV. Our model substantially improves all metrics, highlighting the advantage of video foundation model features for feedforward 3D reconstruction under limited 3D data.

**concentrated?** We ablate which diffusion *layer* and *timestep* yield the most 3D-aware features by sweeping over three network layers and four denoising timesteps. Across all the models we study, the optimum is consistent: *mid-network layers* combined with an *early-but-not-first* time step, are significantly better than other layers and time steps. For the choice of layers, the observation of *mid-network layers* outperforming early or late layers is intuitive: in diffusion models, late layers are specialized to the per-frame RGB synthesis task, which suppresses high-level 3D-related features; whereas in too early layers, high-level features might not have formed yet. For the choice of time steps, in diffusion models, earlier time steps correspond to less noise added to the data or encoded feature. Considering the task of denoising, either too little or too much noise would lead to the degeneration of the task (i.e. either too easy or too hard) and make the features less useful. Comparing between early and late timesteps, early steps work better because the input signal is less corrupted by the noise. Overall, mid-layer and moderately early features strike a balance, retaining global 3D cues while being less influenced by the large noise added for denoising.

### 4.4. VidFM Features for Feedforward 3D

Building on our previous analysis, we observe that features from video foundation models (VidFMs), especially video generative models such as WAN, are highly effective for 3D reconstruction. This raises a natural question: since current state-of-the-art feedforward 3D reconstructors like VGGT [51] rely on DINO features, how does the model perform with VidFM features such as WAN?

We investigate this question in our relatively small-data regime including DL3DV and CO3Dv2. We follow the same dataset split as in the previous experiments, and train (i) the original VGGT model from scratch, with DINO features optimized end-to-end, and (ii) an otherwise identical variant in which we replace DINO with frozen WAN2.1-14B features. Under a matched compute budget, we train both models to convergence and report the results in Table 2. On these benchmarks, our VidFM-based VGGT consistently outperforms the original VGGT by a large margin across all metrics. These results suggest that, when high-quality 3D supervision is limited to small datasets such as CO3Dv2 and DL3DV, it is preferable to use video model features rather than DINO features for feedforward 3D reconstruction.

### 4.5. Limitations

Our study relies on publicly released checkpoints rather than models trained under controlled conditions. Compute and data constraints prevent us from training video generators with precisely controlled variations at scale, so we cannot strictly attribute 3D-awareness differences to several factors of interest (e.g., data, training strategy). In particular, to the best of our knowledge, there are no open-source models that provide multiple versions of checkpoints *only* differing in the scale of training data; as a result, we cannot isolate the effect of data scale. Meanwhile, due to resource constraints, we are unable to train large-scale 3D reconstruction models from scratch on massive datasets with VidFM features—an interesting direction for future work.

## 5. Conclusion

In this paper, we study the 3D awareness of video foundation models. Unlike prior work that focuses on image models and relies on 2.5D or optimization-based proxies, we probe video models using direct 3D prediction tasks. We find that state-of-the-art video generators exhibit strong, generalizable 3D awareness—even compared to domain experts. Our experiments demonstrate the importance of temporal reasoning for 3D understanding, and we examine how 3D fine-tuning, model scaling, and diffusion feature-extraction choices impact 3D awareness. Our experiments also show the promise of using VidFM features for 3D reconstruction in the limited-data regime. Beyond analysis, our work presents a 3D evaluation protocol and benchmark for existing video foundation models. We will publicly release our code, data, and weights, and we hope this work provides a solid step toward understanding and building scalable 3D world models.

## References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 1

[2] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021. 3

[3] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22875–22889, 2025. 1, 3

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738, 2021. 1

[5] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-jepa: Latent video prediction for visual representation learning. 2023. 2, 4, 5, 12

[6] Tyler Bonnen, Stephanie Fu, Yutong Bai, Thomas O'Connell, Yoni Friedman, Nancy Kanwisher, Joshua B Tenenbaum, and Alexei A Efros. Evaluating multiview object consistency in humans and image models. *Advances in Neural Information Processing Systems*, 37:43533–43548, 2024. 3

[7] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 2

[8] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1

[9] Yue Chen, Xingyu Chen, Anpei Chen, Gerard Pons-Moll, and Yuliang Xiu. Feat2gs: Probing visual foundation models with gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6348–6361, 2025. 2, 3

[10] Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3d: Video diffusion models are effective 3d generators. *arXiv preprint arXiv:2403.06738*, 2024. 3

[11] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2025. 3

[12] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21795–21806, 2024. 2, 3, 4, 12

[13] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024. 3

[14] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 2

[15] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–12, 2025. 1, 3

[16] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2

[17] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 1, 3

[18] Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. *arXiv preprint arXiv:2503.10592*, 2025. 1, 3

[19] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. *arXiv*, 2023. 3

[20] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2

[21] Chun-Hao Huang, Niloy J. Mitra, Hyeonho Jeong, Jae Shin Yoon, and Duygu Ceylan. Jog3r: Towards 3d-consistent video generators. In *BMVC*, 2025. 1, 3

[22] Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson WH Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *arXiv preprint arXiv:2506.04225*, 2025. 1, 3

[23] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[24] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024. 3

[25] Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Geo4d: Leveraging video genera-

tors for geometric 4d scene reconstruction. *arXiv preprint arXiv:2504.07961*, 2025. 1, 3

[26] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2

[27] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 2

[28] Xiang Li, Zirui Wang, Zixuan Huang, and James M. Rehg. Cue3d: Quantifying the role of image cues in single-image 3d generation. In *Advances in Neural Information Processing Systems*, 2025. 3

[29] Hanwen Liang, Junli Cao, Vidit Goel, Guocheng Qian, Sergei Korolev, Demetri Terzopoulos, Konstantinos N Plataniotis, Sergey Tulyakov, and Jian Ren. Wonderland: Navigating 3d scenes from a single image. *arXiv preprint arXiv:2412.12091*, 2024. 1, 3

[30] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 4

[31] Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. Reconx: Reconstruct any scene from sparse views with video diffusion model. *arXiv preprint arXiv:2408.16767*, 2024. 3

[32] Yuanxun Lu, Jingyang Zhang, Tian Fang, Jean-Daniel Nahmias, Yanghai Tsin, Long Quan, Xun Cao, Yao Yao, and Shiwei Li. Matrix3d: Large photogrammetry model all-in-one. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11250–11263, 2025. 1, 3

[33] Jinjie Mai, Wenxuan Zhu, Haozhe Liu, Bing Li, Cheng Zheng, Jürgen Schmidhuber, and Bernard Ghanem. Can video diffusion model reconstruct 4d geometry? *arXiv preprint arXiv:2503.21082*, 2025. 1, 3

[34] Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert, Liangyan Gui, and Yu-Xiong Wang. Lexicon3d: Probing visual foundation models for complex 3d scene understanding. *Advances in Neural Information Processing Systems*, 37:76819–76847, 2024. 3

[35] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019. 1

[36] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4, 5, 12

[37] Onur Özyeşil, Vladislav Voroninski, Ronen Basri, and Amit Singer. A survey of structure from motion*. *Acta Numerica*, 26:305–364, 2017. 2

[38] Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, Yuhui Wang, Anbang Ye, Gang Ren, Qianran Ma, Wanying Liang, Xiang Lian, Xiwen Wu, Yuting Zhong, Zhuangyan Li, Chaoyu Gong, Guojun Lei, Leijun Cheng, Limin Zhang, Minghao Li, Ruijie Zhang, Silan Hu, Shijie Huang, Xiaokang Wang, Yuanheng Zhao, Yuqi Wang, Ziang Wei, and Yang You. Open-sora 2.0: Training a commercial-level video generation model in $200k. *arXiv preprint arXiv:2503.09642*, 2025. 2, 4, 5

[39] Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. Sat: Dynamic spatial aptitude training for multimodal language models, 2025. 3

[40] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021. 4

[41] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6121–6132, 2025. 1, 3

[42] Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, David A Forsyth, and Anand Bhattad. Shadows don't lie and lines can't bend! generative models don't know projective geometry... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 28140–28149, 2024. 3

[43] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[44] Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 2024. 1, 3

[45] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 4

[46] Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5283–5293, 2025. 2

[47] Aether Team, Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, et al. Aether: Geometric-aware unified world modeling. *arXiv preprint arXiv:2503.18945*, 2025. 1, 3, 4, 5

[48] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners

for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 2

[49] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 13(4): 376–380, 2002. 4

[50] 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. 2, 4, 5, 12

[51] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 2, 4, 8

[52] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023. 2

[53] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 2

[54] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022. 3

[55] Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *arXiv preprint arXiv:2506.05284*, 2025. 1, 3

[56] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024. 1, 3

[57] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 2

[58] Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21924–21935, 2025. 2, 4, 5, 6, 12

[59] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2, 4, 5

[60] 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. *arXiv preprint arXiv:2409.02048*, 2024. 1, 3

[61] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. A general protocol to probe large vision models for 3d physical understanding. *Advances in Neural Information Processing Systems*, 37:43468–43498, 2024. 3

[62] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. 2023. 3

[63] Qihang Zhang, Shuangfei Zhai, Miguel Angel Bautista Martin, Kevin Miao, Alexander Toshev, Joshua Susskind, and Jiatao Gu. World-consistent video diffusion with explicit 3d modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21685–21695, 2025. 1, 3

[64] Songchun Zhang, Huiyao Xu, Sitong Guo, Zhongwei Xie, Hujun Bao, Weiwei Xu, and Changqing Zou. Spatialcrafter: Unleashing the imagination of video diffusion models for scene reconstruction from limited observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27794–27805, 2025. 1, 3

[65] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. 3

[66] Yiming Zuo, Karhan Kayan, Maggie Wang, Kevin Jeon, Jia Deng, and Thomas L Griffiths. Towards foundation models for 3d vision: How close are we? In *2025 International Conference on 3D Vision (3DV)*, pages 1285–1296. IEEE, 2025. 3

# How Much 3D Do Video Foundation Models Encode?

## Supplementary Material

| | Original Probe | | | Smaller Probe | | |
|---|---|---|---|---|---|---|
| Probed Feature | Point Err(↓) | Depth Err(↓) | AUC@30 (↑) | Point Err(↓) | Depth Err(↓) | AUC@30 (↑) |
| DINOv2 [36] | 2.814 | 0.534 | 0.245 | 3.344 | 0.623 | 0.163 |
| V-JEPA [5] | 1.576 | 0.613 | 0.558 | 1.707 | 0.657 | 0.505 |
| WAN2.1-14B [50] | **1.051** | **0.323** | **0.660** | **1.317** | **0.374** | **0.567** |
| Fast3R [58] | 1.379 | 0.514 | 0.637 | 1.551 | 0.572 | 0.549 |

Table 3. **Ablation on probe sizes.** We compare the 3D awareness evaluation results using our original probe against a smaller probe on DL3DV. The relative rankings and our conclusions remains unchanged despite the change of probe sizes.

This supplementary material presents additional experiments and analyses on the 3D awareness of VidFMs. In Sec. A, we study how probe size affects measured 3D awareness and show that our main conclusions are robust across probe capacities. In Sec. B, we extend our study in Sec. 4.4 by showing how performance scales with the amount of 3D training data. We demonstrate that strong video generator features are especially beneficial for feed-forward 3D reconstruction with limited 3D data or in challenging learning scenarios. Finally, in Sec. C, we analyze the relationship between 3D probe performance and multi-view feature consistency. We find that cross-view correspondence alone can be a biased proxy for true 3D awareness, especially when comparing different model families.

## A. Ablation on Probe Size

In our main experiments, we employ shallow probes with 4 layers and 1024 channels. Here we evaluate whether our conclusions remain valid under even smaller probes. We follow the same experimental protocol as the main paper, but use a significantly smaller probe by halving the model width from 1024 to 512. Table 3 presents 3D awareness results for different-sized probes on DL3DV. We observe the relative performance remain stable across probe sizes; using a smaller probe does not affect our conclusions: features from state-of-the-art video generation models, e.g., WAN2.1-14B, exhibit strong 3D awareness compared to other model categories.

## B. Data Scaling for VidFM VGGT

In Table 2 of the main paper, we compare the original VGGT with our variant that uses VidFM features. We show that using VidFM features significantly benefits feed-forward 3D models under limited resources: under the same training data, VidFM-VGGT outperforms the original VGGT by a large margin. We now extend this experiment by studying how performance changes with the amount of available training data. The scaling behaviors of both VGGT variants on CO3Dv2 and DL3DV are shown in Figure 6. In each plot, the dotted line denotes the performance of the original VGGT trained with 100% of the 3D training data. Our VidFM-VGGT typically surpasses the full-data baseline with only less than 10% of the training data across all metrics. Such contrast suggests that it is possible to induce strong 3D understanding from video features with a tiny fraction of 3D data, especially when compared to the commonly used image features. Thus, strong video generator features are particularly valuable in low-data settings. The gap is especially large on DL3DV, where the scenes are much more diverse and cluttered. This indicates that strong video generator features substantially benefit 3D learning in diverse and challenging data domains. Due to the availability of compute and data, we are not able to extrapolate our curves to the scale of original VGGT's training set, which pools most available 3D data. Such extrapolation will be an interesting future direction.

## C. Analysis on Multi-view Consistency

We study how a model's 3D probe performance relates to multi-view consistency, which prior works often consider as a proxy for 3D awareness [12].

**Measuring multi-view consistency.** To quantify multi-view consistency, we measure the *cross-view correspondence error* of different VidFM features. Cross-view correspondence error is defined as the pixel distance between the predicted correspondence and groundtruth correspondence. To obtain groundtruth correspondence, we sample a random anchor view $A$ and a set of pixels within this view. We then reproject these pixels to another view $B$ using ground-truth 3D, and record their locations if they are not occluded. To obtain predicted correspondence, we use the standard near-

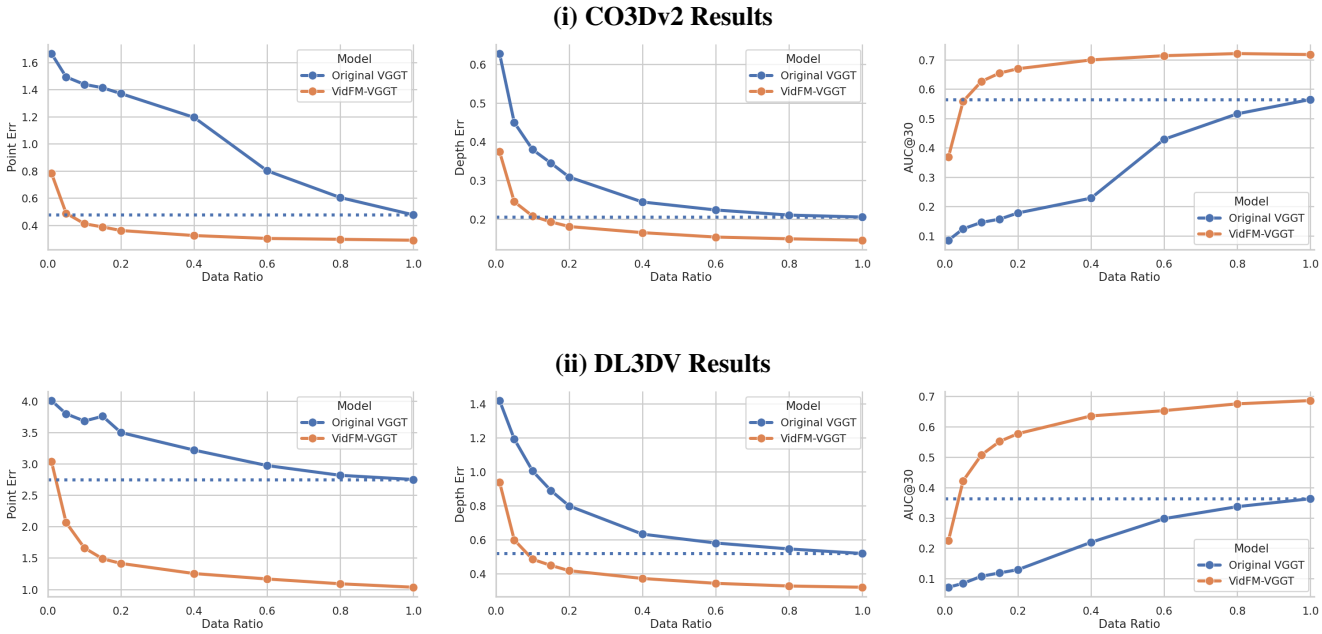



Figure 6. **Data scaling on CO3Dv2 and DL3DV.** For each dataset we report point map error, depth error, and AUC@30 against the fraction of data used to train the model. The horizontal dashed line denotes the performance of the original VGGT trained with 100% of the 3D data. VidFM VGGT typically outperforms this full-data baseline with less than 10% of the 3D training data.

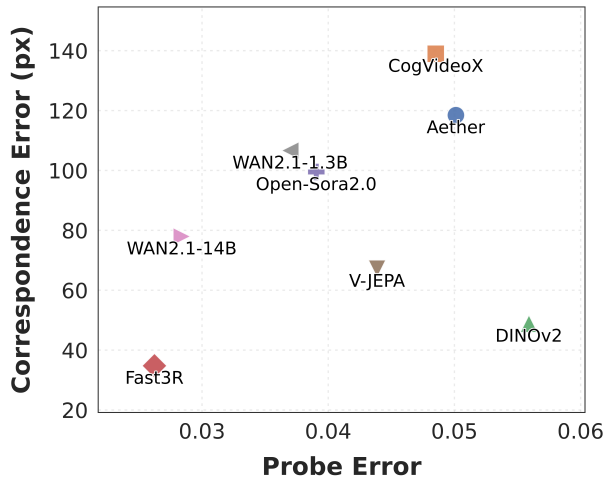



Figure 7. **3D awareness vs. multi-view consistency.** Scatter plot of *3D Probe Error* (lower is better) versus *Cross-view Correspondence Error* (lower is better). Within the family of video diffusion models, the 3D probe error positively correlates with the multi-view correspondence error. DINOv2 and V-JEPA achieve great multi-view correspondence, while performing significantly worse in 3D probing experiments. This suggests that cross-view feature similarity may not be a sufficient proxy for measuring 3D awareness, especially when comparing across families of models.

est neighbor query in feature space: for each anchor points in view $A$, we retrieve the top-1 nearest neighbor in view $B$ based on the VidFM features. We then compute the average Euclidean pixel distance between the predicted correspondence and groundtruth correspondence. We use this mean distance as our measure of multi-view feature consistency, reported as cross-view correspondence error.

**Correlation between 3D probe and multi-view consistency.** Figure 7 plots 3D probe error (x axis; lower is better) against cross-view correspondence error in pixels (y axis; lower is better). We perform this analysis on CO3Dv2, where the probe error is the point error reported in Table 1 in the main paper. Among video diffusion models, we observe a positive correlation, where lower probe error accompanies lower correspondence error. CogVideoX is the worst on both axes, Open-Sora2.0 and WAN2.1-1.3B are intermediate, and WAN2.1-14B is the best (bottom-left). By contrast, feedforward models (Fast3R, V-JEPA, DINOv2) lie *below* the diffusion models. At a comparable probe error, they show better multi-view consistency. Within feedforward models, DINOv2 achieves particularly strong multi-view consistency, yet performs poorly at inferring global 3D properties from its features. We now discuss possible reasons for these observed discrepancies.

**Comparison: diffusion models vs. feedforward models.** Diffusion models exhibit worse multi-view feature consis-

tency than feedforward models at the same level of 3D awareness. This follows from how diffusion features are extracted: noise is injected into the VAE features and a single denoising step is performed to estimate the noise or velocity. This not only makes features noisy at locations where large noise is added, but the underlying representation also includes features specifically tailored to denoising, which is affected by the random noise. Consequently, two pixels corresponding to the same 3D point across frames can carry different features, leading to feature discrepancies that suppress the raw feature consistency even when the underlying 3D structure is well-recoverable by shallow probes.

**Comparison: video models vs. image models.** DINOv2 attains especially strong multi-view consistency, surpassing even the self-supervised video encoder V-JEPA. We hypothesize that in video models some channels correlate with local motions at the current frame; pixels corresponding to the same 3D point may exhibit different local motions across frames. In this way, while video models encode richer temporal information that aids 3D decoding, their features can appear less "consistent" under nearest-neighbor matching. Such factor makes feature consistency alone a potentially biased evaluation for measuring 3D awareness.