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 [51]	**1.051**	**0.323**	**0.660**	**1.317**	**0.374**	**0.567**
Fast3R [59]	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-

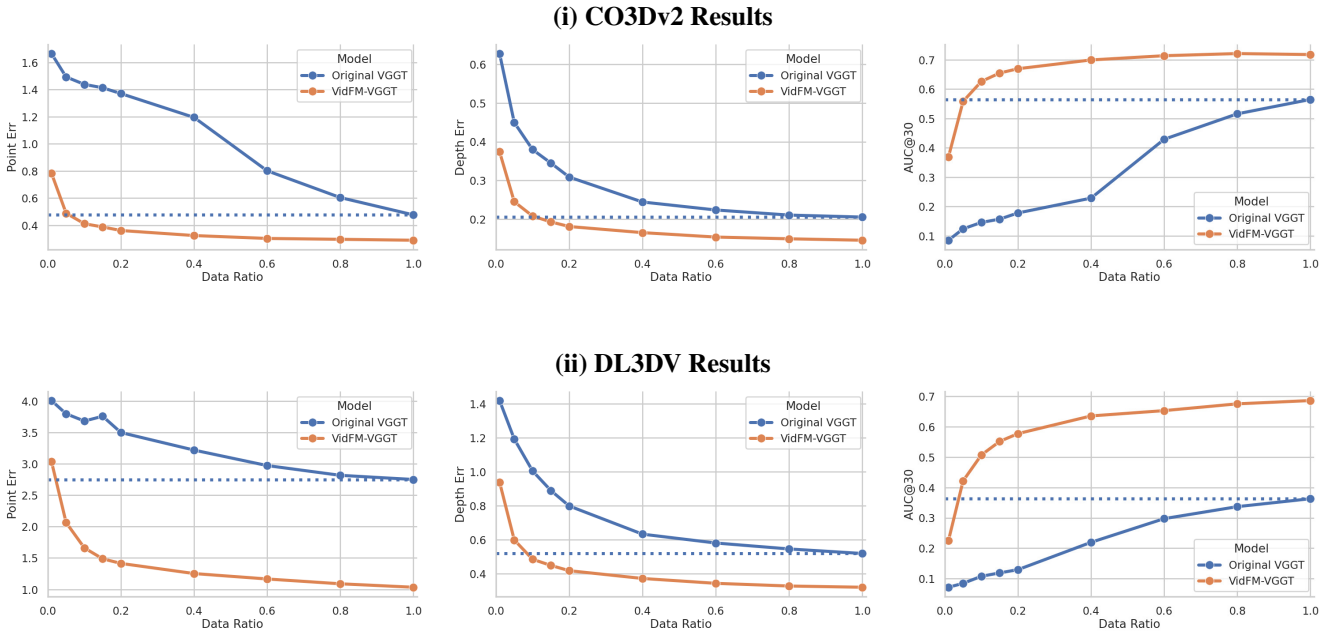

Figure 6. **Data scaling on CO3Dv2 and DL3DV.** For each dataset we report point map error, depth error, and AUC@30 against the fraction of data used to train the model. The horizontal dashed line denotes the performance of the original VGGT trained with 100% of the 3D data. VidFM VGGT typically outperforms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

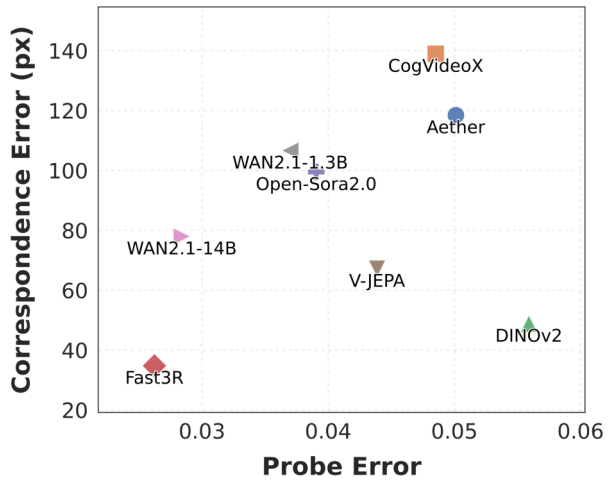

Figure 7. **3D awareness vs. multi-view consistency.** Scatter plot of *3D Probe Error* (lower is better) versus *Cross-view Correspondence Error* (lower is better). Within the family of video diffusion models, the 3D probe error positively correlates with the multi-view correspondence error. DINOv2 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-

Model	Params (M)	FLOPs (T)	Runtime (s/clip)	Estim. Pretraining Data
DINOv2	304.4	96.0	0.93	142M images
V-JEPA	637.5	209.0	4.61	$\sim$ 2M videos
CogVideoX	5,786.1	803.2	3.49	2B images + 35M videos
Aether	5,787.4	871.4	4.14	CogVideoX + $\sim$ 40K videos
Open-Sora2.0	12,137.9	942.1	8.65	$\sim$ 70M video
WAN2.1-14B	14,415.4	1,833.2	25.61	Undisclosed
Fast3R	647.6	114.5	0.46	Undisclosed

Table 4. Summary of backbone compute, runtime, and estimated pretraining data.

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.

D. Backbone Compute and Scale

To contextualize our findings, we add the table summarizing backbone compute and scale (see Table 4). As we evaluate frozen backbones with identical probes, our study compares the representations in off-the-shelf models; we do not claim improvements are solely attributable to the training objective independent of scale. In fact, the strong scalability of VidFMs is a major benefit. A core contribution of our paper is that we propose the first model-agnostic framework to quantitatively study the emergence of 3D encodings from frontier 2D VidFMs, with direct comparisons to 3D experts. This is not only diagnostic: we show replacing DINO features with VidFM features yields a strong feedforward 3D model under limited 3D data.

Views	WAN2.1		DINOv2	
	AUC@30 (↑)	Point Err (↓)	AUC@30 (↑)	Point Err (↓)
4	0.477	1.190	0.131	2.510
8	0.493	1.130	0.128	2.640
12	0.502	1.090	**0.151**	2.510
16	**0.507**	**1.060**	0.149	**2.450**

Table 5. Ablation on number of input frames.

E. VidFM Feature Extraction for 3D

We study whether merging features across layers or timesteps can lead to better 3D decoding. We explore concatenating different features and projecting them to probe dimensions before the probe module. We tested four configurations: the top-2/all timesteps at the best layer, and the top-2/all layers at the best timestep, on the CO3Dv2 ablation set. None of these improves over using the single best feature. This suggests that many 3D signals are likely already recoverable from one carefully chosen feature, and a shallow projection does not reliably add complementary information. Yet this remains a promising direction and it will be valuable to explore more expressive feature fusion.

F. Implementation Details on Temporal Compression

Most video models compress the video temporally and the feature extraction needs to handle this properly. In our work, for a queried frame s, we select the temporally compressed spatial feature whose window contains s and use it as the probe input. To avoid many-to-one ambiguities (multi-frames mapping to one token), we sample query frames with stride K equal to the LCM of the temporal window sizes across models (e.g., $s \in \{0, 1, 5, 9, ...\}$), so each selected feature consistently corresponds to the first frame of its window. For models with smaller windows, we take all features within the K-frame window and concatenate them channel-wise (e.g. DINOv2 packs 4 features as 1), yielding an apples-to-apples K-frame representation across models.

Probed Feature	Point Err(↓)	Depth Err(↓)	AUC@5 (↑)	AUC@30 (↑)
DINOv2 [36]	2.511	0.785	0.001	0.131
DINOv3 [44]	2.365	0.776	0.003	0.151
WAN2.1-14B [51]	**0.906**	**0.431**	**0.070**	**0.610**

Table 6. **DINOv2 vs DINOv3 on DL3DV-1K mini subset.** DINOv3 shows modest improvements over DINOv2 across all metrics, but both per-frame models remain far lower than video models. Point map errors have been multiplied by 10 for clarity.

G. Ablation on Number of Input Frames

In Table 5, we show probing results for varying frame numbers (trained on DL3DV 1K subset). Global 3D estimations become more accurate with more views, yet the change is relatively small compared to the gap across models and does not affect our conclusions.

H. Original GT vs. COLMAP GT

To further validate our findings, we additionally train and evaluate all probes using COLMAP reconstructions as ground truth on a DL3DV-1K subset. COLMAP provides classical SfM-based 3D structure that is independent of any learned model, but is inherently sparse ($\sim$1–2% of pixels have valid 3D correspondences). As a result, point and depth metrics on COLMAP GT are less informative. On camera metrics, AUC@30, the relative ranking of VFMs is remarkably consistent with Table 1 in the main paper: WAN2.1-14B [51] (0.553), Fast3R [59] (0.549), V-JEPA [5] (0.526), Open-Sora2.0 [38] (0.509), Aether [48] (0.426), CogVideoX [60] (0.413), and DINOv2 [36] (0.124). WAN2.1-14B and Fast3R remain the top two models on camera pose metrics, while DINOv2 remains the weakest. These results confirm the robustness of our findings.

I. DINOv2 vs. DINOv3

We further investigate whether the recently released DINOv3 narrows the gap between self-supervised image encoders and video models (Table 6). DINOv3 shows modest improvements over DINOv2 across all metrics, with AUC@30 increasing from 0.131 to 0.151 (+15%). However, both image models remain far below video generators such as WAN2.1-14B.