The Diashow Paradox: Stronger 3D-Aware Representations Emerge from Image Sets, Not Videos

Anonymous ICCV submission

Paper ID 9

Abstract

Image-based vision foundation models (VFMs) have demonstrated surprising 3D geometric awareness, despite no explicit 3D supervision or pre-training on multi-view data. While image-based models are widely adopted across a range of downstream tasks, video-based models have so far remained on the sidelines of this success. In this work, we conduct a comparative study of video models against image models on three tasks encapsulating 3D awareness: multi-view consistency, depth and surface normal estimation. To enable a fair and reproducible evaluation of both model families, we develop AnyProbe*, a unified framework for probing network representations. The results of our study reveal a surprising conclusion, which we refer to as* the diashow paradox. *Specifically, video-based pre-training does not provide any consistent advantage on downstream tasks involving 3D understanding over image-based pre-training. We formulate two hypotheses to explain our observations, which underscore the need for high-quality video datasets and highlight the inherent complexity of video-based pre-training. AnyProbe will be publicly released to streamline evaluation of image- and video-based VFMs in a mutually consistent fashion.*

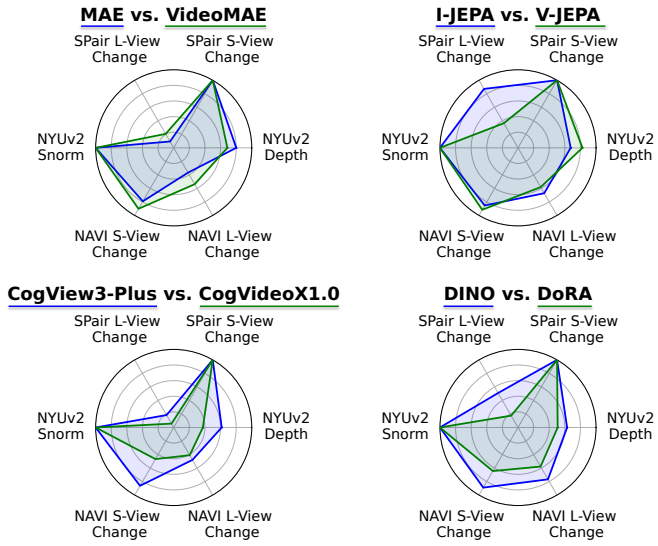

Figure 1. **The diashow paradox.** Streamlining the evaluation protocol, we develop *AnyProbe*, a framework to systematically compare video and image models on depth, surface normals (NYUv2 [22]), and multi-view correspondences [16, 21]. Contrary to a common-sense expectation, we find that video-based vision foundation models (green) do not provide a consistent advantage over image-based models (blue) on 3D awareness tasks.

1. Introduction

Vision foundation models (VFMs), such as DINO [5, 23], have found widespread use in the computer vision community [*e.g.* 29, 30]. These VFMs encode a surprising degree of 3D spatial reasoning, despite using image sets for training [10]. By contrast, video models, such as V-JEPA [4] and VideoMAE [26], learn from image *sequences* providing multiple views on the same scene. Therefore, it should naturally follow that video models will exhibit enhanced 3D awareness in comparison to the image-based VFMs. Consolidating the scattered empirical evidence in the literature [20, 31], this short paper aims at illuminating an apparent paradox encountered in practice: the representations learned by video models [4, 26] tend to exhibit *a lower level of 3D awareness* than their image-based counterparts.

We implement an evaluation framework, *AnyProbe*, which allows us to benchmark video- and image-based models in a unified protocol. Specifically, AnyProbe can train task-specific probes on top of frozen VFMs, including video-based models. In our study, we focus on three key tasks of 3D spatial awareness: monocular depth estimation, surface normal estimation, and multi-view correspondence. Our main contributions are:

- We develop *AnyProbe*, an evaluation framework designed to provide fair and reproducible comparisons of VFMs across diverse downstream tasks with a minimal engineering overhead. We will release AnyProbe to the research

community under the MIT license.

- Using AnyProbe, we conduct a study to compare video models to image models on 3D awareness tasks. The results, illustrated in Fig. 1, lead to a surprising observation: video models do not provide a consistent advantage over their image-based counterparts on 3D understanding tasks. We dub this phenomenon as *the diashow paradox*, recognizing that state-of-the-art models essentially learn from a "slide show" of the world rather than a continuous video stream – in contrast to humans.
- Lastly, we propose two hypotheses to explain the observed results: *(i)* a lack of visual diversity in existing video datasets, and *(ii)* a deceptively high non-triviality of 3D-aware representation learning from videos.

The results of our study challenge the conventional understanding that using video training data per se implies improved 3D-aware representations in comparison to image-based pre-training. Our findings illuminate two open avenues for future research: curating high-quality video datasets and 3D-aware pre-training strategies on videos.

2. Related Work

Probing VFMs. Recent studies such as Probe3D [10], DepthCues [6], and Lexicon3D [20] compare visual foundation models on core spatial reasoning tasks, including depth estimation, surface normal prediction, multiview consistency, and 3D scene understanding. These works provide a foundation for our AnyProbe framework in assessing the geometric capabilities of VFMs. Specifically, our work considers the impact of video pre-training: we compare video- and image-based models with a similar architecture to isolate the effect of motion in pre-training on videos.

Image- and video-based foundation models. Image-based foundation models leverage large-scale datasets and self-supervised learning techniques to extract visual representations that transfer well across diverse downstream tasks. MAE [14] reconstructs masked image patches; I-JEPA [1] predicts latent codes for missing regions; and the DINO models [5, 23] leverage self-distillation with Vision Transformers for robust feature learning. These methods have naturally been adapted to video input: VideoMAE [26] applies masked auto-encoding to video sequences, V-JEPA and V-JEPA 2 [2, 4] adapt joint-embedding for spatio-temporal data, and DoRA [27], drawing inspiration from DINO, leverages object continuity to learn robust representations from long, unlabeled videos.

3. Methodology

We introduce AnyProbe, a unified foundation probing framework that enables systematic evaluation across diverse datasets, tasks, architectures, and probing protocols. For example, AnyProbe supports lightweight integration of video-based VFMs for monocular image tasks, allowing for a comparative analysis of image- and video-based models on 3D-awareness tasks. We will open-source AnyProbe.

Model	Data	Res.	Arch.
	Video		
VideoMAE [26]	K400, SSV2	224	B/16, L/16
V-JEPA [2, 4]	VM2M, VM22M	224, 384	H/16
DoRA [27]	WTours	224	S/16
CogvideoX [32]	Video-Text Pairs	768 × 1360	–
	Image		
MAE [14]	IN1K	224	B/16, L/16
I-JEPA [1]	IN1K, IN22K	224, 448	H/16, H/14
DINO [5]	IN1K, WTours	224	S/16
CogView-3Plus [33]	LAION-2B[25]	512	–

Table 1. **VFMs in our study.** Models are categorized by architecture (Arch.), training datasets (Data) and resolution (Res.).

Dataset	Content	Scenes	Frames
ImageNet-1K [8]	Natural object images	1.28M	1.28M
ImageNet-22K [8]	Fine-grained object images	14.2M	14.2M
VM2M [4]	Video datasets	2M	14B
VM22M [2]	Video + image datasets	22M	90B
WTours [27]	Egocentric city walk videos	10	3.5M
Kinetics-400 [17]	Human action clips	240K	71.8 M
SSV2 [12]	Object interaction videos	160K	19.4 M
LAION-2B [25]	Image–text pairs	2.26B	2.26B

Table 2. **Overview of pre-training datasets**. Each dataset is specified by its number of scenes and frames. For video datasets, frame counts are estimated based on the total duration and frame rate.

Model overview. Our study evaluates a diverse set of image- and video-based VFMs summarized in Tab. 1. We contextualize the models with statistics of the underlying training data in Tab. 2. We select official checkpoints and use comparable architectures and training strategies available for both input modalities, images and videos (*e.g.* I-JEPA [1] *vs.* V-JEPA [4]; MAE [14] *vs.* VideoMAE [26]).

Adapting video ViT models for image tasks. To fairly compare video to image models, we adapt video models for image tasks. The video models included in our study, V-JEPA [4] and VideoMAE [26], tokenize videos by embedding tubelets of size $T \times H \times W$ with a 3D convolution. To adapt these models for per-image processing, we define a 2D convolution kernel k_{2D}, which derives from its 3D counterpart k_{3D} by summing over the temporal dimension:

$$k_{\text{2D}}(i, j) = \sum_{t=1}^{T} k_{\text{3D}}(t, i, j). \quad (1)$$

This formulation is equivalent to the practice of duplicating the image along the temporal dimension (*cf*. [2]), but is more computationally efficient. To adapt the 3D positional

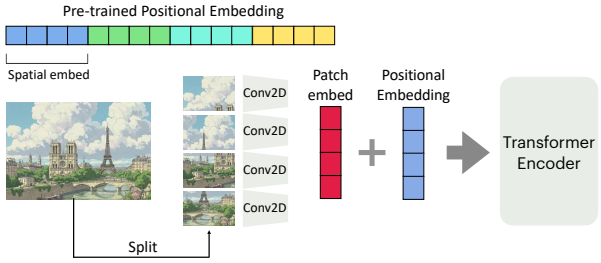

Figure 2. **Adapting video-based ViT models for image tasks**. We extract token embeddings from image patches using 2D convolutions adapted for video models, see Eq. (1), and use only the spatial positional embeddings of the first frame.

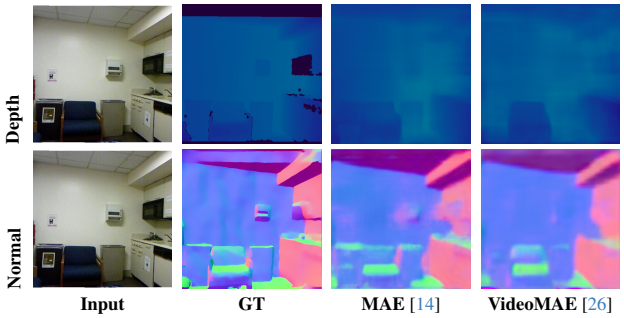

Figure 3. **Qualitative examples.** We visualize depth and surface normal maps derived from running AnyProbe with a DPT [24] head on top of frozen MAE and VideoMAE on NYUv2 [22]. MAE produces consistently sharper and more accurate geometry.

embedding, we use only the components of the first frame and discard the rest. As a result, video-based models retain their architecture, but can operate on image tasks without any computational overhead, as visualized in Fig. 2.

Depth and surface normal estimation. We extract pixel-level features and optimize a DPT probe [24]. For depth, we use a scale-invariant sigmoid depth loss [9] and a depth-gradient matching loss [15]; for normal estimation, we use an uncertainty-aware angular loss [3]. To ensure the fairness of our evaluation, we resize input images to the original pre-training resolution. This approach avoids interpolating the positional embedding, which we empirically confirmed to have a negative effect on the downstream accuracy. Instead, we downsample the input to the pre-training resolution. We resize the predicted maps produced by the DPT probe (depth or normals) to the original image size for evaluation. Specifically, we optimize the DPT probe for 10 epochs using AdamW [18] with an initial learning rate of $5e-4$, a cosine decay schedule [19] and batch size of 8.

Multi-view correspondence. We follow the previous metrics and protocols for evaluating multi-view correspondence [10], with adaptations for video models, as shown in Fig. 2. We compute the recall, the fraction of the matches with an error below a threshold. Given two views of the same object or scene, we extract dense feature maps from a frozen backbone and establish feature correspondences based on their cosine similarity. We consider two evaluation scenarios. **SPair-71k** [21] provides class-level keypoints on object instances. We evaluate only the annotated points and perform a single nearest-neighbor lookup in feature space. We calculate 2D pixel-level error in normalized image coordinates and report the aggregate recall at threshold 0.10. We further break down the recall by a relative change in viewpoint levels $d \in \{0, 1, 2\}$ (with $d = 0$ indicating small and $d = 2$ indicating large view change). **NAVI** [16] samples dense points on the object mask. We extract dense features, perform nearest-neighbor matching with Lowe's ratio test [10], and retain the top 1000 correspondences. We then compute the 3D Euclidean error between each match and its ground-truth back-projected point, and report the recall at a fixed 2cm threshold, grouped by relative viewpoint bins $\theta \in \{[0, 30), [30, 60), [60, 90), [90, 120]\}$.

4. Experiments

We use NYUv2 [22] for our experiments with depth and surface normal estimation. For multi-view correspondence, we use SPair-71k [21] and NAVI [16]. We rely on AnyProbe to obtain the results, summarized in Tab. 3. Notably, DINOv2 [23], provided for reference, surpasses all video models in both single-image estimation and multi-view consistency. Comparing video and image models with the same architecture, we make the following observations: **VideoMAE [26] exhibits higher multi-view consistency than MAE [14], but inferior accuracy of depth and surface normals.** Within the autoencoder family, MAE consistently outperforms VideoMAE [26] across all architectures in depth and surface normal estimation, achieving lower RMSE scores on ViT-L16 (0.277 *vs.* 0.323 RMSE for depth; 24.79 *vs.* 26.53 RMSE for normals). The results with ViT-B16 align with these trends. While VideoMAE shows improved multi-view consistency over MAE, this advantage is rather marginal: DINO [5] outperforms both MAE variants substantially on both SPair-71k and NAVI.

V-JEPA [26] and its sequel V-JEPA 2 [2] yield inferior accuracy than their image-based counterpart, I-JEPA [1], across all benchmarks. Specifically, I-JEPA models (H16/H14) significantly outperform V-JEPA-H16 and V-JEPA 2 in view consistency on SPairs-71k, achieving the average recall of 44.81/35.74 *vs.* 25.13/28.4, respectively.

An image model, DINO [5], pre-trained on videos is only marginally worse than a video model, DoRA [27]. However, a DINO [5] model pre-trained on IN1K consistently outperforms DoRA [27] pre-trained on the World Tours dataset in terms of depth (0.353 *vs.* 0.437 RSME), surface normals (28.49 *vs.* 31.60 RSME), and multi-view consis-

Arch	Model	Dataset	Res.	SPair-71k ↑				NAVI ↑				NYUv2 ↓	
				$d=0$	$d=1$	$d=2$	all	θ^{0}_{30}	θ^{30}_{60}	θ^{60}_{90}	θ^{90}_{120}	Depth RMSE	Snorm RMSE
B14	DINOv2	LVD-142M	518x518	62.55	51.40	48.79	56.26	94.27	69.04	39.65	25.85	0.222	24.79
B16	DINO	IN1K	224x224	28.83	24.40	24.90	26.10	86.33	55.79	30.65	22.69	0.334	28.01
	MAE	IN1K	224x224	9.60	6.24	4.91	7.92	76.91	40.15	19.68	11.96	**0.317**	**25.95**
	VideoMAE	K400	224x224	8.98	**6.45**	**5.47**	7.73	88.32	**48.32**	**20.38**	**11.58**	0.384	28.37
	VideoMAE	SSV2	224x224	**13.03**	6.21	4.21	**10.03**	**89.14**	46.69	19.56	11.10	0.413	29.17
L16	MAE	IN1K	224x224	8.22	6.00	4.57	7.17	74.58	37.92	14.55	11.39	**0.277**	**24.79**
	VideoMAE	K400	224x224	**20.11**	**12.68**	**10.07**	**16.51**	**85.02**	**49.72**	**21.32**	**12.12**	0.323	26.53
H16	I-JEPA	IN1K	448x448	**51.35**	**38.47**	**42.52**	**44.81**	**81.58**	**52.57**	**27.42**	**17.72**	**0.246**	**23.43**
	V-JEPA	VM2M	224x224	29.00	17.47	17.75	23.34	86.39	52.56	23.03	13.79	0.269	26.09
	V-JEPA	VM2M	384x384	30.92	18.74	19.60	25.13	77.19	45.41	21.23	15.03	0.270	24.88
	V-JEPA 2	VM22M	224x224	32.48	25.78	24.46	28.40	76.01	48.11	23.12	13.15	0.259	25.05
H14	I-JEPA	IN22K	224x224	**43.18**	**31.28**	32.12	**37.45**	**83.07**	**49.84**	23.89	**14.88**	**0.294**	**25.64**
	I-JEPA	IN1K	224x224	40.48	30.79	**32.75**	35.74	82.51	49.73	**24.99**	14.83	0.329	27.69
S16	DINO	IN1K	224x224	**27.97**	**24.32**	**24.88**	**25.66**	**83.85**	**53.77**	**30.42**	**22.46**	**0.353**	**28.49**
	DINO	WTours	224x224	8.98	9.05	6.63	8.41	54.76	32.75	21.51	14.81	0.477	33.70
	DoRA	WTours	224x224	12.45	10.28	8.55	11.36	60.74	36.71	22.93	16.37	0.437	31.60
3B	CogView-3Plus	LAION-2B	512x512	**20.02**	**12.04**	**8.95**	**15.91**	**81.34**	**45.66**	**19.02**	10.31	**0.361**	**29.32**
2B	CogVideoX1.0	Video-Text Pairs	768x1360	3.17	2.93	2.77	3.05	44.08	28.27	16.26	**11.54**	0.594	34.47

Table 3. **Image *vs.* video comparison on 3D awareness.** We evaluate all models with AnyProbe, which extracts features for multi-view correspondence and train a DPT probe for depth and surface normal estimation. For multi-view consistency, we use SPair-71k (200 keypoints) and NAVI (1000 keypoints) and report recall (*cf*. Sec. 3). We use NYUv2 for depth and surface normal estimation and report RMSE. Across most architectures, image-based models (highlighted in blue) consistently outperform their video counterparts on 3D tasks. The image-based DINOv2 achieves the best overall results.

tency (*e.g*. 25.66 *vs.* 11.36 recall on SPairs-71k).

An image-based diffusion model, CogView-3Plus [33], substantially outperforms its video-based counterpart, CogVideoX [32]. Consolidating previous observations, we test text-to-image diffusion models, CogView-3Plus and CogVideoX. The video-based CogVideoX shows diminished 3D awareness compared to image pre-training across all metrics, despite sharing similarities in the architecture and the pre-training scheme.

5. Discussion and Conclusion

Our findings highlight an apparent paradox, referred to as *the diashow paradox*. Contrary to the prevailing belief, learning from image sets, not videos, appears to yield more 3D-aware representations. Our study supports this postulation with empirical evidence: VFMs trained on static images consistently match or exceed their video-based counterparts in geometric 3D-awareness tasks (*cf*. Tab. 3). We discuss two hypotheses to explain our observations:

Existing video datasets lack visual diversity. Deep models benefit from data diversity and size. The temporal continuity inherent to videos renders image sequences highly redundant. Recall from Tab. 3 that DINO trained on Walking Tours (a few, but long videos) performs significantly worse than its IN1K variant across all metrics. Furthermore, existing video datasets are well-known to contain low-resolution sequences and compression artifacts. While a "golden bullet" dataset exists for image pre-training (ImageNet), we may have yet to curate such a dataset for the video modality. However, the same curation effort would apply to image sets as well. Amassing diverse image sets may still prove more cost-efficient than curating videos.

Learning 3D-aware representations from videos is deceptively non-trivial. The photographic bias leveraged by self-supervised methods on image datasets [5, 14] might be less prominent in videos, where strong frame-to-frame correlations and camera motion introduce additional challenges [7, 11, 13, 28]. Consequently, learning multi-view consistency and geometric priors from videos becomes a highly non-trivial task (*e.g*. V-JEPA 2 lacks I-JEPA's multi-view robustness despite larger-scale pre-training).

Perhaps as a consequence of the task complexity, a typical evaluation protocol of video models is rarely a 3D-aware task, favoring instead other spatio-temporal benchmarks, such as action recognition and anticipation [*e.g*. 2]. We hope that AnyProbe will encourage more comprehensive evaluation of future video models, including 3D awareness tasks, by simplifying the engineering effort of probing experiments and streamlining cross-modal comparison.

Conclusion. Systematically comparing video to image models, our study exposes *the diashow paradox*: image sets yield surprisingly more 3D-aware representations than videos. By highlighting the paradox, we hope to incentivize the community in future explorations of video curation, and 3D-aware training and evaluation strategies.

References

[1] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[2] Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction, and planning. *arXiv:2506.09985*, 2025. 2, 3, 4

[3] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[4] Adrien Bardes, Quentin Garrido, Jean Ponce, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv:2404.08471*, 2024. 1, 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 3, 4

[6] Duolikun Danier, Mehmet Aygün, Changjian Li, Hakan Bilen, and Oisin Mac Aodha. DepthCues: Evaluating monocular depth perception in large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[7] Srijan Das and Michael S. Ryoo. Viewclr: Learning self-supervised video representation for unseen viewpoints. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5573–5583, 2023. 4

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2

[9] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network, 2014. 3

[10] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 3

[11] Jiayi Gao, Zijin Yin, Changcheng Hua, Yuxin Peng, Kongming Liang, Zhanyu Ma, Jun Guo, and Yang Liu. Conmo: Controllable motion disentanglement and recomposition for zero-shot motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 4

[12] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense . In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Los Alamitos, CA, USA, 2017. IEEE Computer Society. 2

[13] Tengda Han, Dilara Gokay, Joseph Heyward, Chuhan Zhang, Daniel Zoran, Viorica Patraucean, Joao Carreira, Dima Damen, and Andrew Zisserman. Learning from streaming video with orthogonal gradients. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025. 4

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 4

[15] Julia Hornauer, Amir El-Ghoussani, and Vasileios Belagiannis. Revisiting gradient-based uncertainty for monocular depth estimation. *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6), 2025. 3

[16] Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, Andre Araujo, Ricardo Martin-Brualla, Kaushal Patel, Daniel Vlasic, Vittorio Ferrari, Ameesh Makadia, Ce Liu, Yuanzhen Li, and Howard Zhou. NAVI: Category-agnostic image collections with high-quality 3d shape and pose annotations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1, 3

[17] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. 2

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 3

[19] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017. 3

[20] Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert, Liang-Yan Gui, and Yu-Xiong Wang. Lexicon3d: Probing visual foundation models for complex 3d scene understanding. In *Advances in Neural Information Processing Systems*, 2024. 1, 2

[21] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence, 2019. 1, 3

[22] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision (ECCV)*, 2012. 1, 3

[23] 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. In *International Conference on Learning Representations*, 2025. 1, 2, 3

[24] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[25] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: an open large-scale dataset for training next generation image-text models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2022. Curran Associates Inc. 2

[26] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2022. Curran Associates Inc. 1, 2, 3

[27] Shashanka Venkataramanan, Mamshad Nayeem Rizve, João Carreira, Yuki M Asano, and Yannis Avrithis. Is imagenet worth 1 video? learning strong image encoders from 1 long unlabelled video. In *International Conference on Learning Representations*, 2024. 2, 3

[28] Jinpeng Wang, Yuting Gao, Ke Li, Yiqi Lin, Andy J Ma, Hao Cheng, Pai Peng, Rongrong Ji, and Xing Sun. Removing the background by adding the background: Towards background robust self-supervised video representation learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2021. 4

[29] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. to appear. 1

[30] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. to appear. 1

[31] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling foundation models for multimodal video understanding. In *Computer Vision – ECCV 2024, Lecture Notes in Computer Science, vol. 13631*, pages 396–416. Springer, 2024. 1

[32] 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:2408.06072*, 2024. 2, 4

[33] Wendi Zheng, Jiayan Teng, Zhuoyi Yang, Weihan Wang, Jidong Chen, Xiaotao Gu, Dong Yuxiao, Ming Ding, and Jie Tang. *CogView3: Finer and Faster Text-to-Image Generation via Relay Diffusion*. 2024. 2, 4