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

Duc Nguyen[1] Anna Sonnweber[1] Mark Weber[1,2] Nikita Araslanov[1,2] Daniel Cremers[1,2]

[1] TU Munich [2] MCML

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 yet to share in 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 offers no consistent advantage over image-based pre-training on downstream tasks involving 3D understanding. 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.*

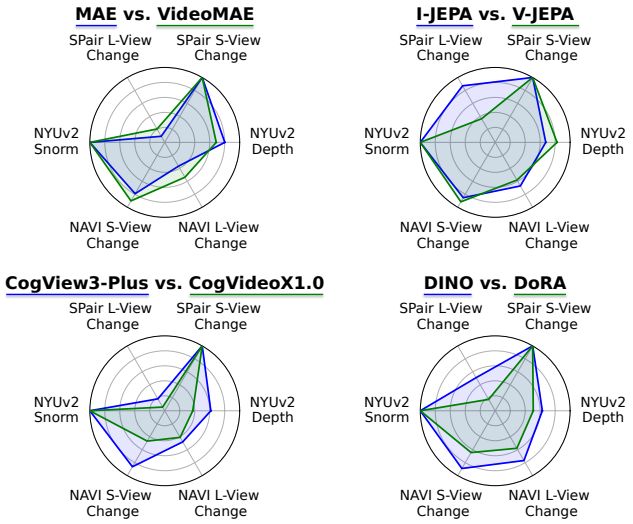

Figure 1. **The diashow paradox.** Streamlining the evaluation protocol, we develop *AnyProbe*, a framework to systematically compare video and image models on depth, surface normals (NYUv2 [26]), and multi-view correspondences [18, 23]. Contrary to a common-sense expectation, we find that video-based 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 [7, 24], have found widespread use in the computer vision community [*e.g.* 31, 32]. Despite using image sets for training, these VFMs encode a surprising degree of 3D spatial reasoning [5]. By contrast, video models, such as V-JEPA [6] and VideoMAE [28], learn from image *sequences*, which can provide multiple views of 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 [22, 33], we illuminate an apparent paradox encountered in practice: the representations learned by video models [6, 28] 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 shared 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 release AnyProbe to the research community under the MIT license.[1]

[1] AnyProbe repository: https://github.com/tum-vision/anyprobe

- 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 currently learn from a "slideshow" of the visual 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 [5], DepthCues [8], and Lexicon3D [22] 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 [15] reconstructs masked image patches; I-JEPA [2] predicts latent codes for missing regions; and the DINO models [7, 24] leverage self-distillation with Vision Transformers for robust feature learning. These methods have naturally been adapted to video input: VideoMAE [28] applies masked auto-encoding to video sequences, V-JEPA and V-JEPA 2 [3, 6] adapt joint-embedding for spatio-temporal data, and DoRA [29], 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.

Model	Data	Res.	Arch.
	Video		
VideoMAE [28]	K400, SSV2	224	B/16, L/16
V-JEPA [3, 6]	VM2M, VM22M	224, 384	H/16
DoRA [29]	WTours	224	S/16
CogvideoX [34]	Video-Text Pairs	768 × 1360	–
	Image		
MAE [15]	IN1K	224	B/16, L/16
I-JEPA [2]	IN1K, IN22K	224, 448	H/16, H/14
DINO [7]	IN1K, WTours	224	S/16
CogView-3Plus [35]	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 [10]	Natural object images	1.28M	1.28M
ImageNet-22K [10]	Fine-grained object images	14.2M	14.2M
VM2M [6]	Video datasets	2M	14B
VM22M [3]	Video + image datasets	22M	90B
WTours [29]	Egocentric city walk videos	10	3.5M
Kinetics-400 [19]	Human action clips	240K	71.8 M
SSV2 [13]	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 [2] *vs.* V-JEPA [6]; MAE [15] *vs.* VideoMAE [28]).

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 [6] and VideoMAE [28], 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:

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

This formulation is equivalent to the practice of duplicating the image along the temporal dimension [3], but is more computationally efficient. To adapt the 3D positional embedding, we use only the components of the first frame and discard the rest. As a result, video-based models retain their

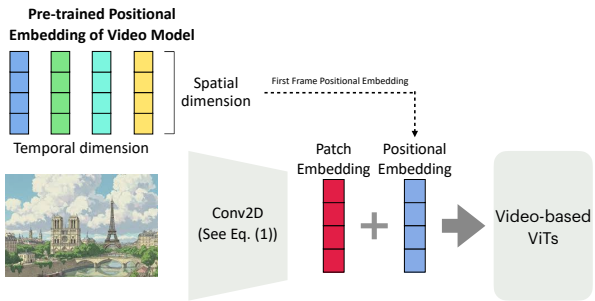

Figure 2. **Adapting video-based ViT models for image tasks**. We tokenize the image using a 2D patch-embedding convolution adapted for the video model (Eq. (1)), as in image ViTs, and use only the spatial positional embeddings of the first frame.

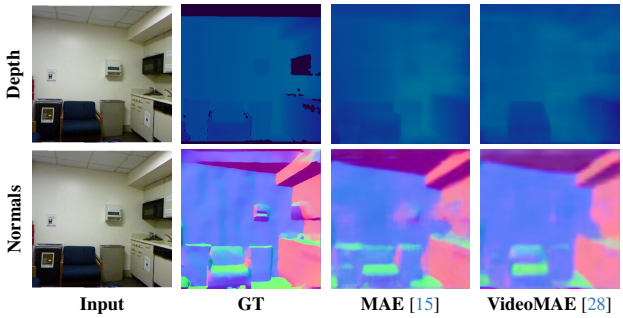

Figure 3. **Qualitative examples.** We visualize depth and surface normal maps derived from running AnyProbe with a DPT [36] head on frozen MAE and VideoMAE on NYUv2 [26]. MAE produces consistently sharper, more accurate geometry.

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 [36]. For depth, we use a scale-invariant sigmoid depth loss [11] and a depth-gradient matching loss [16]; for normal estimation, we use an uncertainty-aware angular loss [4]. 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 [20] with an initial learning rate of $5e-4$, a cosine decay schedule [21] and batch size of 8.

Multi-view correspondence. We follow the previous metrics and protocols for evaluating multi-view correspondence [5], 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** [23] 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** [18] samples dense points on the object mask. We extract dense features, perform nearest-neighbor matching with Lowe's ratio test [5], 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 [26] for our experiments with depth and surface normal estimation. For multi-view correspondence, we use SPair-71k [23] and NAVI [18]. We rely on AnyProbe to obtain the results, summarized in Tab. 3. Notably, DINOv2 [24], 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:

I. VideoMAE [28] exhibits higher multi-view consistency than MAE [15], but inferior accuracy of depth and surface normals. Within the autoencoder family, MAE consistently outperforms VideoMAE [28] 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 [7] outperforms both MAE variants substantially on both SPair-71k and NAVI.

II. V-JEPA [6] and its sequel V-JEPA 2 [3] yield inferior accuracy than their image-based counterpart, I-JEPA [2], 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.

III. An image model, DINO [7], pre-trained on videos is only marginally worse than a video model, DoRA [29]. However, a DINO model pre-trained on IN1K consistently outperforms DoRA 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 consistency (*e.g.* 25.66 *vs.* 11.36 recall on SPairs-71k).

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.

IV. An image-based diffusion model, CogView-3Plus [35], substantially outperforms its video-based counterpart, CogVideoX [34]. 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. More technically, the gradients in video pre-training exhibit high correlation [14]. 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 known to contain low-resolution sequences and compression artifacts. While a "golden bullet" dataset exists for image pre-training (ImageNet) [17], 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 in the long term.

Learning 3D-aware representations from videos is deceptively non-trivial. The photographic bias leveraged by self-supervised methods on image datasets [7, 15] might be less prominent in videos, where strong frame-to-frame correlations and camera motion introduce additional challenges [9, 12, 14, 30]. 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*. 3]. 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 of video-based models.

Acknowledgments. DN is supported by the DAAD program Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the German Federal Ministry of Education and Research.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep ViT features as dense visual descriptors. *arXiv:2112.05814*, 2021. 2

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael G. Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 15619–15629. IEEE, 2023. 2, 3

[3] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew J. 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

[4] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 13117–13126. IEEE, 2021. 3

[5] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas J. Guibas, Justin Johnson, and Varun Jampani. Probing the 3D awareness of visual foundation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 21795–21806. IEEE, 2024. 1, 2, 3

[6] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Trans. Mach. Learn. Res.*, 2024. 1, 2, 3

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9630–9640. IEEE, 2021. 1, 2, 3, 4

[8] Duolikun Danier, Mehmet Aygün, Changjian Li, Hakan Bilen, and Oisin Mac Aodha. DepthCues: Evaluating monocular depth perception in large vision models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 20049–20059. Computer Vision Foundation / IEEE, 2025. 2, 3

[9] Srijan Das and Michael S. Ryoo. ViewCLR: Learning self-supervised video representation for unseen viewpoints. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pages 5562–5572. IEEE, 2023. 4

[10] 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 the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2

[11] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2366–2374, 2014. 3

[12] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 7191–7200. Computer Vision Foundation / IEEE, 2025. 4

[13] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, 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 *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5843–5851. IEEE Computer Society, 2017. 2

[14] Tengda Han, Dilara Gokay, Joseph Heyward, Chuhan Zhang, Daniel Zoran, Viorica Patraucean, João Carreira, Dima Damen, and Andrew Zisserman. Learning from streaming video with orthogonal gradients. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 13651–13660. Computer Vision Foundation / IEEE, 2025. 4

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15979–15988. IEEE, 2022. 2, 3, 4

[16] Julia Hornauer, Amir El-Ghoussani, and Vasileios Belagiannis. Revisiting gradient-based uncertainty for monocular depth estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(6):4395–4408, 2025. 3

[17] Mi-Young Huh, Pulkit Agrawal, and Alexei A. Efros. What makes ImageNet good for transfer learning? *arXiv:1608.08614*, 2016. 4

[18] Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, André Araújo, 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 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 1, 3

[19] 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. *arXiv:1705.06950*, 2017. 2

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 3

[21] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 3

[22] Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert, Liangyan Gui, and Yu-Xiong Wang. Lexicon3D: Probing visual foundation models for complex 3D scene understanding. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 1, 2

[23] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv:1908.10543*, 2019. 1, 3

[24] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido 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. *Trans. Mach. Learn. Res.*, 2024. 1, 2, 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 *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 2

[26] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *Computer Vision 12th European Conference on Computer Vision, ECCV 2012, Florence, Italy, October 7-13, 2012, Proceedings, Part V*, pages 746–760. Springer, 2012. 1, 3

[27] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 2

[28] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 1, 2, 3

[29] 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 *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 2, 3

[30] Jinpeng Wang, Yuting Gao, Ke Li, Yiqi Lin, Andy J. Ma, Hao Cheng, Pai Peng, Feiyue Huang, Rongrong Ji, and Xing Sun. Removing the background by adding the background: Towards background robust self-supervised video representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11804–11813. Computer Vision Foundation / IEEE, 2021. 4

[31] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotný. VGGT: Visual geometry grounded transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 5294–5306. Computer Vision Foundation / IEEE, 2025. 1

[32] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 10510–10522. Computer Vision Foundation / IEEE, 2025. 1

[33] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling foundation models for multimodal video understanding. In *Computer Vision 18th European Conference, ECCV 2024, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXV*, pages 396–416. Springer, 2024. 1

[34] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. 2, 4

[35] Wendi Zheng, Jiayan Teng, Zhuoyi Yang, Weihan Wang, Jidong Chen, Xiaotao Gu, Yuxiao Dong, Ming Ding, and Jie Tang. CogView3: Finer and faster text-to-image generation via relay diffusion. In *Computer Vision 18th European Conference, ECCV 2024, Milan, Italy, September 29-October*

4, 2024, Proceedings, Part LXXVII, pages 1–22. Springer, 2024. 2, 4

[36] Shuangquan Zuo, Yun Xiao, Xiaojun Chang, and Xuanhong Wang. Vision transformers for dense prediction: A survey. *Knowl. Based Syst.*, 253:109552, 2022. 3

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

– Supplementary Material –

A. Adapting Video ViTs for Image Tasks

In contrast to image-based VFMs, video-based VFMs have spatio-temporal positional embeddings. To establish an effective inference mode of video-based VFMs on dense image tasks, we evaluate and compare multiple inference modes of pre-trained V-JEPA and VideoMAE on dense geometric tasks: depth and surface normal estimation.

A straightforward strategy to adapt video-based VFMs for static image tasks is to duplicate the input image N times, where N is the number of frames used during video pre-training. We obtain the image features by concatenating the output tokens at corresponding spatial positions across all replicated frames. These temporally aggregated features are then passed through a probing head for downstream evaluation. Fig. 4 illustrates the aggregation process used for probing. One downside of this approach is the computational cost. An inference pass on a single image now scales with the number of frames divided by the temporal dimension of the tubelet. In practice, this overhead amounts to a multi-fold increase in FLOPs compared to image-based counterparts, such as VideoMAE/MAE or V-JEPA/I-JEPA. To avoid the computational overhead, we adopt the Conv2D head from Eq. (1) and apply three different strategies of handling the temporal dimension of the positional embeddings as shown in Fig. 5. In total, we evaluate four adaptation strategies for video ViTs on dense image tasks:

- **Temporal Duplication and Aggregation:** Replicate the image N times and concatenate tokens across the temporal axis before probing.
- **Temporal Interpolation:** In this strategy, we utilize the adapted Conv2D as in Eq. (1) and interpolate the temporal dimension of the pre-trained positional embeddings to single-image input.
- **Temporal Frame Alignment:** Using the Conv2D head from Eq. (1), we realign the temporal positional embeddings to a single input image at inference/probing, matching the model's temporal prior to a static input.
- **Representation Replication:** To disentangle gains from representation ability versus parameter count, we replicate the *Temporal Frame Alignment* features to match the *Temporal Duplication and Aggregation* tensor shape before decoding.

We evaluate four adaptation strategies of the temporal positional embeddings with two video-based VFMs. We use NYUv2 depth estimation as the evaluation benchmark. Tab. 4 presents the quantitative results.

While *Temporal Duplication and Aggregation* achieves the lowest RMSE across both architectures (0.3805 for VideoMAE-b16 and 0.2649 for V-JEPA-h16), the performance gains over *Temporal Frame Alignment* are modest,

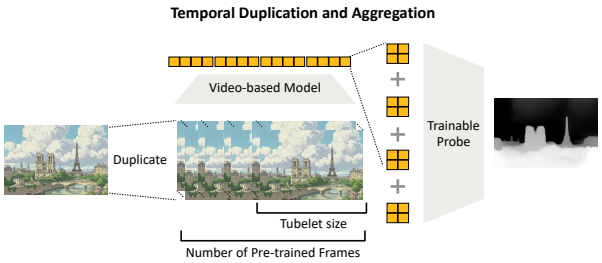

Figure 4. **Input duplication and temporal aggregation.** Input image is replicated N times; tokens at corresponding spatial positions across frames are concatenated along the temporal axis before probing.

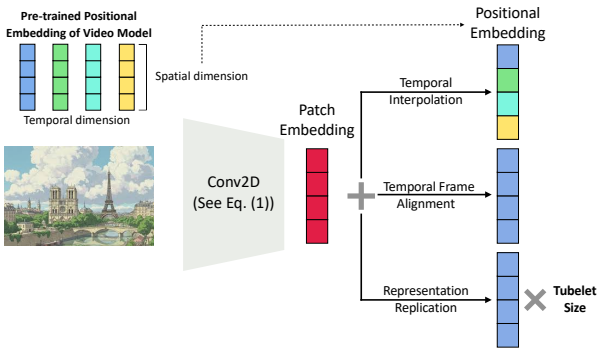

Figure 5. **Temporal adaptation for a single-image input.** Patch embedding from a single image is encoded via adapted Conv2D, the pre-trained positional embeddings are adapted with three strategies: *Temporal Interpolation*, *Temporal Frame Alignment* and *Representation Replication*.

Model	Temporal Pos Emb Strategy	RMSE ↓
VideoMAE-b16	Temporal Interpolation	0.4184
	Temporal Frame Alignment	0.3906
	Temporal Duplication and Aggregation	**0.3805**
	Representation Replication	0.3969
V-JEPA-h16	Temporal Interpolation	0.2866
	Temporal Frame Alignment	0.2685
	Temporal Duplication and Aggregation	**0.2649**
	Representation Replication	0.2677

Table 4. **Adaptation of temporal positional embeddings in video ViTs.** We evaluate different strategies for adapting video ViTs for image tasks on NYUv2 depth estimation with VideoMAE and V-JEPA.

2.6% and 1.3% relative improvement, respectively, while incurring a substantial computational overhead.

To investigate whether these gains stem from increased model capacity rather than the representation capability, we evaluate *Representation Replication*, which matches the spatial dimensionality of the duplication setup. The results (0.3969 for VideoMAE-b16 and 0.2677 for V-JEPA-h16) demonstrate that the benefits of *Temporal Duplication and Aggregation* cannot be attributed solely to increased parameter count. Rather, the performance improvements arise from the aggregation of multiple temporal representations, indicating that ensemble effects contribute significantly to the observed gains.

Temporal Interpolation consistently underperforms across both models, suggesting that interpolation of the temporal positional embeddings is not the best strategy to preserve the learned temporal relationships critical for effective feature extraction from a single image.

Conclusion. *Temporal Frame Alignment* emerges as a strong baseline for adapting video-based VFMs to dense image tasks, achieving competitive performance while maintaining computational efficiency. This approach preserves the learned temporal structure of the model without the computational overhead of frame duplication. These properties make the approach practical for real-world deployment, where computational resources are constrained.

B. Adapting Video ViTs for Multi-View Tasks

We adapt video-based models for inference on multi-view tasks. Here, we focus on the multi-view correspondence task as the main benchmark [1, 5, 27], which has previously been studied in the context of *image* encoders. We explore four different strategies for adapting *video* models:

- **Temporal Duplication and Aggregation**: As in Appendix A, each input image is temporally replicated to fill the expected number of frames for the pre-trained backbone. The resulting features are then concatenated as in Fig. 4.
- **Temporal Frame Alignment:** Following Appendix A, we also align temporal positional embeddings to a single-frame input and extract correspondence features.
- **Joint Temporal Attention:** To capture cross-frame attention from video-based models, we present both input images as temporally adjacent frames. We extract features from the correspondence tokens of each frame, utilizing cross-frame attention in the video backbone.
- **Asymmetric Frame Composition:** Temporal positional embeddings in video models can introduce representational asymmetries between otherwise identical frames, leading to instability in direct feature comparisons. To address this, we adopt an asymmetric composition strategy designed to neutralize position-induced drift: *(1)* image A followed by image B, from which we extract features corresponding to image A; and *(2)* image B followed by image A, from which we derive the features corresponding to image B.

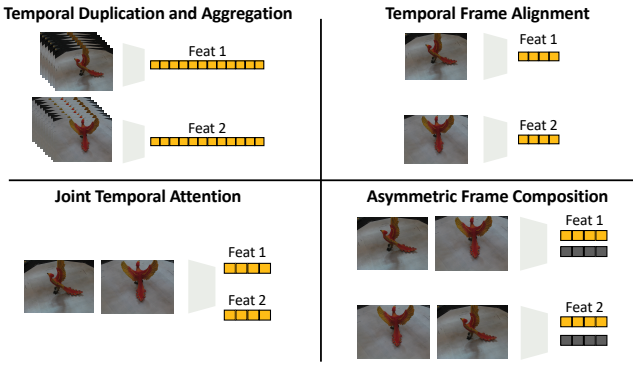

Figure 6. **Adapting video ViTs for multi-view correspondence.** We compare four feature-extraction modes: (i) Temporal Duplication and Aggregation, (ii) Temporal Frame Alignment, (iii) Joint Temporal Attention (two views as adjacent frames), and (iv) Asymmetric Frame Composition to cancel position-induced drift.

Model	Strategies	θ^{0}_{30}	θ^{30}_{60}	θ^{60}_{90}	θ^{90}_{120}
VideoMAE-b16	Temporal Frame Alignment	88.32	**48.32**	**20.38**	11.58
	Asymmetric Frame Composition	68.90	35.04	16.60	**11.78**
	Joint Temporal Attention	12.84	10.93	7.52	7.35
	Temporal Duplication and Aggregation	**89.63**	48.24	18.87	10.20
V-JEPA-h16	Temporal Frame Alignment	**58.22**	**30.71**	**17.48**	**11.25**
	Asymmetric Frame Composition	51.32	27.27	15.53	10.52
	Joint Temporal Attention	11.19	9.23	6.67	6.00

Table 5. **Performance metrics summary for feature extraction strategies.** Comparison of four feature-extraction modes for video pre-trained backbones on NAVI with 1000 keypoints: Temporal Duplication and Aggregation, Temporal Frame Alignment, Joint Temporal Attention, and Asymmetric Frame Composition. Bold indicates the best per column within each model.

Fig. 6 demonstrates the setup and the structural differences among the approaches.

B.1. Results and Analysis

The results reveal key insights into probing video-based ViTs' representations for multi-view correspondence.

Temporal Frame Alignment preserves cross-view consistency in both VideoMAE and V-JEPA and nearly dominates across view changes, where it matches *Temporal Duplication and Aggregation*, while employing the same compute as the image-based ViTs.

Cross-frame attention variants perform poorly compared to other variants. *Joint Temporal Attention* collapses below 13% for all viewpoint changes, while *Asymmetric Frame Composition*, which extracts features from A in context (A, B) and from B in context (B, A), exposes positional asymmetry in otherwise identical frames, yielding large gains over *Joint Temporal Attention* across settings (e.g., 68.90%

vs. 12.84% for VideoMAE at θ^0_{30}). Yet, there remains a substantial gap to *Temporal Frame Alignment*, indicating that generic cross-frame attention does not enforce view-consistent geometry representation.

Conclusion. The results in Tab. 5 indicate that for video-based VFMs, such as VideoMAE and V-JEPA, temporal frame alignment offers a more stable and consistent representation for multi-view correspondence and preserves the spatial semantics of the input.

C. Adapting ViTs for Different Resolutions

Downstream tasks may require inference at a different input resolution than the one used for model training. When the test-time input resolution deviates from the pre-training resolution, we can either *recompute* the positional embeddings on the new grid or *interpolate* them from the pre-trained grid [5, 8]. In our study, we consider three strategies:

- **Recompute Positional Embedding:** We recompute the positional sinusoidal embeddings on the new grid.
- **Interpolate Positional Embedding:** We interpolate pre-trained embeddings to the new grid.
- **Resize Input Images:** We resize the image to the pre-training resolution, which requires no changes to the pre-trained backbone. We then upsample the features to the target size for the probe.

Motivation. With frozen encoders, changing the positional code or sequence length induces a distribution shift. Interpolation is common at inference, but with relative positions, this change to the spatial/temporal grid can impair performance [5]. If this is not carefully controlled, comparisons become biased by the adaptation strategy, not the underlying representation quality.

In each approach, the input image is divided into patches consistent with the pre-trained configuration. However, the first two methods (*Recomputing Positional Embeddings*, *Interpolating Pre-trained Positional Embeddings*) alter the input size, thereby changing the number of tokens and the sequence length processed by the transformer. This variation can lead to increased computational demands, particularly with high-resolution images.

In probing scenarios, models are typically frozen, making the introduction of new positional embeddings through recomputation potentially inconsistent. The models pre-trained on specific image sizes may not generalize well to different resolutions if we alter their positional embeddings. Conversely, interpolating existing positional embeddings is a widely adopted practice during inference, enabling models to handle varying image sizes effectively without retraining. However, as discussed by Banani et al. [5], both recomputing and interpolating positional embeddings may not be effective with models that use relative positional embeddings and can significantly degrade model performance. In contrast, resizing input images to the pre-trained dimensions allows the model to operate under familiar conditions, providing a more stable foundation for parametric probes and accommodating a broader variety of models.

Model	Strategy	δ_1	δ_2	δ_3	Depth-RMSE
MAE-b16	Recompute Pos Emb	0.818	0.961	0.991	0.4779
	Interpolate Pos Emb	0.883	0.980	0.995	0.3875
	Resize Input Images	**0.932**	**0.988**	**0.997**	**0.3187**
SAM-b16	Interpolate Pos Emb	0.821	0.965	0.991	0.4865
	Resize Input Images	**0.881**	**0.978**	**0.996**	**0.3998**
DeiT-3-b16	Recompute Pos Emb	0.940	0.991	0.998	0.3063
	Resize Input Images	**0.942**	**0.991**	**0.998**	**0.3028**
CLIP-b16	Recompute Pos Emb	0.721	0.932	0.982	0.6082
	Resize Input Images	**0.763**	**0.944**	**0.984**	**0.5589**
MiDaS-l16	Recompute Pos Emb	0.915	0.984	0.996	0.3491
	Resize Input Images	**0.914**	**0.984**	**0.995**	**0.3490**
DoRA-b16	Recompute Pos Emb	0.793	0.953	0.988	0.5227
	Resize Input Images	**0.850**	**0.969**	**0.991**	**0.4373**
DINOv2-b14	Interpolate Pos Emb	**0.975**	**0.997**	**0.999**	**0.2223**
	Resize Input Images	**0.975**	**0.997**	**0.999**	**0.2223**
DINO-s16	Interpolate Pos Emb	0.895	0.980	0.995	0.3790
	Resize Input Images	**0.913**	**0.983**	**0.996**	**0.3528**
DINO-b16	Interpolate Pos Emb	0.909	0.983	0.996	0.3571
	Resize Input Images	**0.920**	**0.986**	**0.996**	**0.3340**

Table 6. **Adapting ViTs to input resolution.** We evaluate depth estimation as positional embeddings are adapted across resolutions via interpolation, resizing, or recomputation.

Tab. 6 (NYUv2, 480×480) shows that *Resizing Input Images* consistently matches or outperforms recomputation/interpolation across MAE, SAM, DeiT-3, CLIP, MiDaS, DoRA, DINO/DINOv2 in terms of depth accuracy (higher δ; lower RMSE). When pre-training includes higher-resolution exposure (*e.g.*, DINOv2), the gap to interpolation vanishes.

Conclusion. For trainable probes on dense image and video tasks, resizing the inputs to the backbone's pre-training resolution appears to be a superior configuration compared to the alternatives. This approach preserves the computational cost and attention statistics, avoids positional drift, and consistently yields the strongest, most stable results.