ALIGN2REFINE: DIAGNOSING AND REFINING VIDEO REPRESENTATIONS THROUGH SPATIAL LATENT ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Pre-trained video encoders can produce spurious tokens that fail to represent the visual content at their spatial locations. Prior work derives identification criteria from internal activations and token interactions, but extending these analyses to video is complicated by architectural diversity and complex spatiotemporal attention. We approach this challenge through task feedback: the spatial association between prediction errors and local feature fragmentation suggests that task responses can help locate spurious tokens. To obtain this feedback without downstream annotations, we formulate spatial latent alignment. We find that encoder representations capturing high-level semantics or low-level visual details can both serve as diagnostic references for this task. Building on these insights, we introduce **Align2Refine**, a framework for post-hoc refinement of video encoders on unlabeled videos. To make alignment feedback more discriminative for token identification, we use deliberately imbalanced learning to reinforce differences in fitting progress. The resulting token partition guides selective refinement using the original encoder's features as supervision. Across five heterogeneous video encoders, Align2Refine delivers substantial improvements in both spatial prediction and temporal correspondence under frozen-backbone evaluation.

1 INTRODUCTION

Local token representations in pre-trained video encoders (Assran et al., 2025; Wang et al., 2026) are not always faithful to the visual content at their corresponding spatial locations. While recent work such as V-JEPA 2.1 seeks to improve dense representations through dedicated pre-training objectives (Mur-Labadia et al., 2026), we study post-hoc representation refinement of existing encoders. We adapt the encoder on unlabeled videos without repeating large-scale pre-training or relying on downstream task supervision.

In image vision transformers, register-based methods mitigate local representation artifacts (Darcet et al., 2024; Jiang et al., 2025). UniRefiner distinguishes *regular* tokens, which represent content at their corresponding spatial locations, from *spurious* tokens, which fail to do so (Qiu et al., 2026). This distinction guides selective refinement: preserving regular token features while training boundary registers to absorb the features of identified spurious tokens. To identify spurious tokens, prior work derives diagnostic cues from the encoder's forward computation, relating activation and token-interaction patterns to mechanisms such as global information aggregation (Shi et al., 2026). Extending this analysis to video encoders is challenging, as architectural diversity and complex spatiotemporal attention make the mechanisms underlying spurious tokens difficult to disentangle (Assran et al., 2025; Wang et al., 2026; Tang et al., 2026).

These difficulties motivate a complementary perspective: diagnosing token representations through the task responses they support. Comparing linear segmentation predictions with feature visualizations in V-JEPA 2 and OV-Encoder, we observe that prediction errors spatially overlap with local feature fragmentation (Sec. 3). This association suggests that task responses can expose where local content is poorly represented, providing feedback for locating candidate spurious tokens. Direct use of segmentation errors, however, would make diagnosis dependent on ground-truth annotations and a task-specific prediction head.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

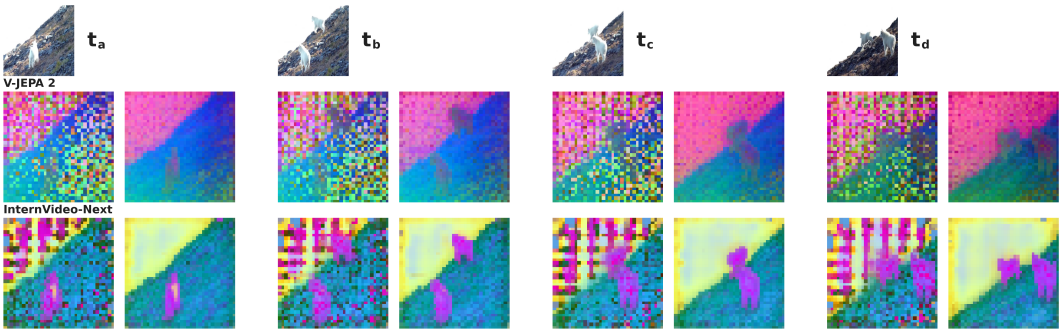

Figure 1: **Video representation refinement.** PCA visualizations of V-JEPA 2 and InternVideo-Next features at four timestamps before and after Align2Refine. Base and Refined features are shown on the left and right of each pair, respectively. For each backbone, the PCA basis and scale are fitted jointly on Base features from all displayed frames and reused for Refined features. The refined features exhibit less local fragmentation across the displayed timestamps.

To obtain such feedback without downstream annotations, we formulate *spatial latent alignment* as a more general diagnostic framework. The task predicts a reference latent from the video token at the same spatial location, using alignment performance to assess how well the token supports a readout of that content. We examine reference latents that emphasize high-level semantics (DINOv3 (Siméoni et al., 2025)) and low-level visual information (SD3.5 VAE (Esser et al., 2024)). Alignment to either target yields feedback for identifying spurious tokens.

Building on these findings, we introduce **Align2Refine**, which uses spatial latent alignment to identify spurious tokens and guide selective refinement. To make alignment feedback more discriminative, we train an alignment probe to reinforce differences in fitting progress. A shared token-wise projector serves as the probe, with both encoders frozen and no additional spatial or temporal aggregation. We evaluate its predictions through *spatial alignment consistency*, defined as the cosine similarity between each predicted latent and its reference at the same location. Our *consistency-guided token reweighting* uses the current consistency scores as a proxy for fitting progress, assigning larger loss weights to high-consistency tokens and smaller weights to the remaining tokens. This deliberately imbalanced learning process is designed to promote two separable distributions of consistency scores. After training, a threshold calibrated for each encoder assigns low-consistency tokens to the spurious group and the remaining tokens to the regular group.

For refinement, the original video encoder serves as a frozen teacher. Following UniRefiner (Qiu et al., 2026), regular teacher features supervise student features at corresponding spatial locations, while spurious teacher features supervise the student's boundary registers.

Align2Refine improves spatial prediction and temporal correspondence across five heterogeneous video encoders under frozen-backbone evaluation. Across these encoders, refinement yields gains of 4.4–21.5 mIoU on ADE20K and 7.7–13.2 $\mathcal{J}\&\mathcal{F}$ on DAVIS at M resolution. The feature visualizations in Figure 1 further illustrate the reduction in local fragmentation after refinement.

Our contributions are threefold:

- We introduce a diagnostic perspective that identifies spurious video tokens through task feedback. Spatial latent alignment provides this feedback without downstream annotations, with both high-level semantic and low-level visual encoder representations serving as diagnostic references.
- We introduce **Align2Refine**, a framework for post-hoc video representation refinement on unlabeled videos. It reinforces fitting differences through deliberately imbalanced alignment learning, promoting score separation for token identification and using the resulting partition to guide selective refinement.
- Align2Refine delivers consistent and substantial gains in spatial prediction and temporal correspondence across different video encoders under frozen-backbone evaluation.

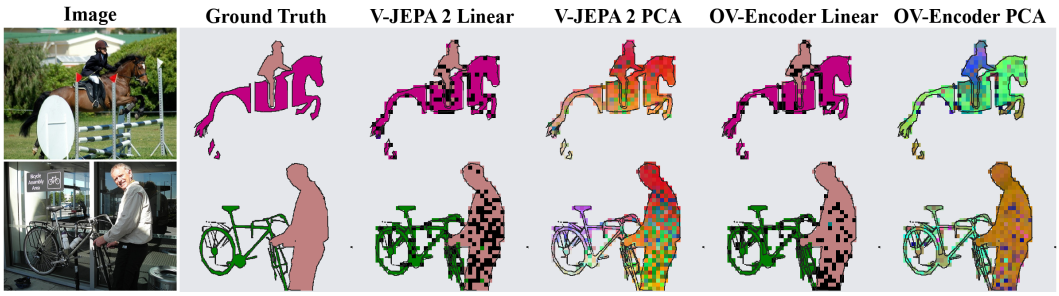

Figure 2: **Prediction errors and local feature fragmentation.** Ground-truth masks, linear segmentation predictions, and PCA visualizations for V-JEPA 2 and OV-Encoder. Predictions and feature visualizations are shown on each encoder's native token grid. For clarity, predictions and PCA visualizations are displayed only within object regions selected by ground-truth masks. PCA is fitted using tokens from the full image.

2 Related Work

Artifacts and post-hoc refinement. ViT artifact studies show that patch tokens can support global computation at the expense of local content (Darcet et al., 2024). Analyses of positional artifacts and token aggregation have broadened this account of spatial distortion (Yang et al., 2024; Shi et al., 2026). Denoising and register-based methods make it possible to repair such features in pre-trained models (Yang et al., 2024; Jiang et al., 2025; Yan et al., 2025). UniRefiner organizes selective refinement around whether each token represents the content at its location (Qiu et al., 2026), making the identification of spurious tokens central to deciding which features to preserve. UniRefiner and SINDER obtain identification cues by analyzing internal token behavior (Qiu et al., 2026; Wang et al., 2024a). For video encoders, diverse spatiotemporal interactions complicate this analysis. We retain UniRefiner's refinement framework but approach diagnosis through task feedback: how well a token supports prediction of its local visual content.

Video representation learning. Masked reconstruction and latent prediction exploit spatiotemporal context to learn video representations without annotations (Tong et al., 2022; Assran et al., 2025). Image-teacher distillation provides semantic supervision (Li et al., 2023), and frameworks such as InternVideo2 combine this knowledge transfer with video modeling and multimodal learning (Wang et al., 2024b). V-JEPA 2.1 further emphasizes dense feature quality by explicitly supervising spatial and temporal grounding (Mur-Labadia et al., 2026). Our work addresses local representation quality after pre-training, using diagnosis to guide selective refinement of existing video encoders on unlabeled videos.

3 Empirical Study

We examine whether local feature fragmentation is reflected in the predictions made from those features. We freeze V-JEPA 2 ViT-g and OV-Encoder ViT-L and train a linear segmentation probe for each encoder. The linear readout assesses how well the existing token representations support local semantic prediction. Figure 2 compares the predictions with ground-truth masks and PCA visualizations on each encoder's native token grid.

In the visualizations, prediction errors spatially overlap with local feature fragmentation. This association suggests that task responses can help locate candidate spurious tokens by revealing where local content is poorly represented.

Using segmentation errors directly for diagnosis would require task-specific annotations and a trained segmentation head. We therefore formulate spatial latent alignment to obtain feedback about local content prediction without segmentation labels. The task predicts a reference latent from the video token at the same location, using alignment performance as the diagnostic signal (Sec. 4).

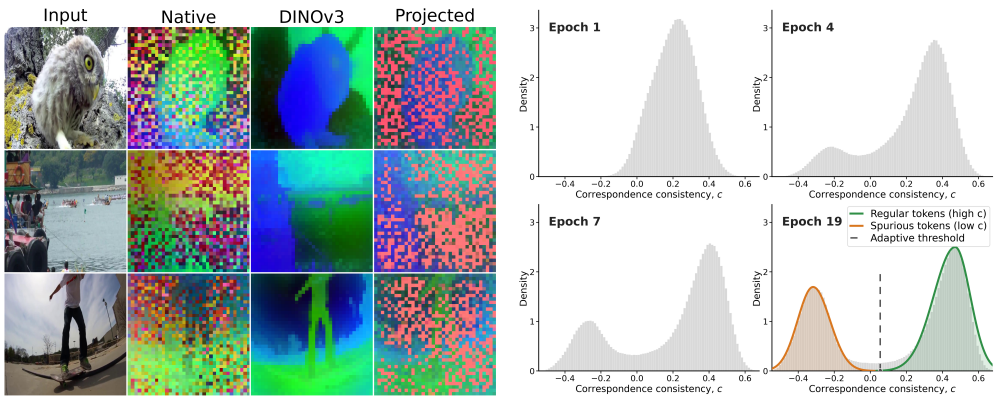

Figure 3: **Progressive consistency learning on V-JEPA 2 ViT-g. (a)** PCA visualizations of native video features, projected features, and DINOv3 reference features. Fragmented patterns occur at similar locations in the native and projected features. **(b)** Consistency scores separate into high- and low-consistency distributions during training; a two-component GMM calibrates the threshold τ after training.

4 METHODOLOGY

Align2Refine uses token-level diagnosis to guide post-hoc refinement of a pre-trained video encoder on unlabeled videos. As shown in Figure 4, the framework consists of two stages: *Align* and *Refine*. The Align stage uses spatial latent alignment with a reference encoder to assess how well video tokens represent their local content, yielding a partition into regular and spurious tokens (Secs. 4.1 and 4.2). The Refine stage uses this partition to adapt the video encoder, preserving regular representations while suppressing spurious responses.

4.1 LEARNING SPATIAL LATENT ALIGNMENT

To diagnose spurious tokens without downstream annotations, we formulate spatial latent alignment as a local prediction task. A probe predicts reference features from video tokens at corresponding locations, using prediction consistency to assess how faithfully each token represents its local content. We train the probe through deliberately imbalanced learning to reinforce differences in fitting progress, making the resulting consistency scores more discriminative for token identification.

Spatial latent alignment. Given an unlabeled video clip, we extract token representations from the video encoder F and reference features from F_{ref}. Representations encoding high-level semantics (DINOv3) or low-level visual details (SD3.5 VAE) can both provide reference targets for this diagnostic task. We use DINOv3 by default and report results with the SD3.5 VAE in the appendix. Let $z_{t,i}$ denote the video token at frame t and spatial location i, and let $z^{\text{ref}}_{t,i}$ denote its corresponding reference latent. With both encoders frozen, we train a shared token-wise MLP g to predict $z^{\text{ref}}_{t,i}$ from $z_{t,i}$. We define *spatial alignment consistency* as

$$c_{t,i} = \cos\big(g(z_{t,i}), z^{\text{ref}}_{t,i}\big)\,. \tag{1}$$

Consistency-guided token reweighting. For each location (t, i), we compute an InfoNCE loss (Oord et al., 2018) $\ell_{t,i}$ on ℓ_2-normalized features, pairing the reference latent with the projected feature at the same location and using projected features at other locations in the same frame as negatives. Using current consistency as a proxy for fitting progress, we set $w^{(e)}_{t,i} = 1$ for the highest-consistency ρ_e fraction of tokens in each frame at epoch e, and 0.1 otherwise. The weighted alignment objective is

$$\mathcal{L}^{(e)}_{\text{probe}} = \sum_{t,i} w^{(e)}_{t,i}\, \ell_{t,i}. \tag{2}$$

We linearly decrease ρ_e during the initial training epochs and then keep it fixed, progressively concentrating the larger weights on fewer tokens. As illustrated in Figure 3(b), this deliberately imbalanced learning process promotes separation between high- and low-consistency token distributions, providing the basis for subsequent token identification.

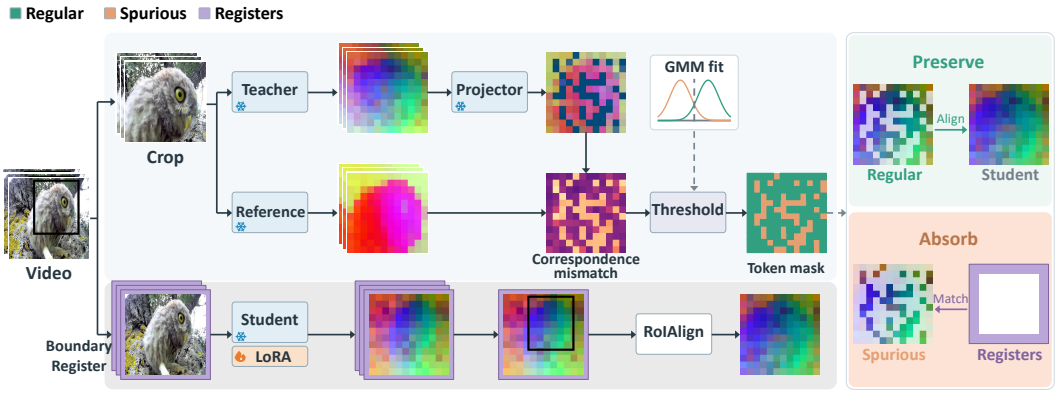

Figure 4: **Overview of Align2Refine.** The frozen alignment probe partitions teacher tokens into regular and spurious subsets. Regular teacher features supervise spatially corresponding student features, while spurious teacher features supervise the student's temporal boundary registers.

4.2 Detecting Spurious Tokens

We use the learned consistency scores to partition tokens into regular and spurious groups, with a threshold calibrated for each video encoder. With the alignment projector g frozen, we fit a two-component Gaussian mixture to consistency scores computed on the videos used to train the alignment probe:

$$p(c) = \pi_r \mathcal{N}(c; \mu_r, \sigma_r^2) + \pi_s \mathcal{N}(c; \mu_s, \sigma_s^2), \qquad \mu_s < \mu_r. \tag{3}$$

Here, r and s denote the high- and low-consistency components, respectively. We set $\tau \in (\mu_s, \mu_r)$ at the score where a token is equally likely to belong to either component under the fitted mixture. Using this point as the decision boundary, we identify tokens with $c_{t,i} < \tau$ as spurious and the remaining tokens as regular. This calibration requires no downstream annotations or validation results, and the threshold remains fixed throughout refinement.

To assess the validity of the token partition, we perform linear segmentation probing on ADE20K using frozen V-JEPA 2 ViT-g features. Separate probes are trained on the full token set and on each identified subset. Table 1 shows the same ordering on both metrics: regular tokens yield the strongest probing performance, followed by the full token set and then spurious tokens. The agreement between this partition and downstream prediction performance provides empirical support for spatial alignment consistency as a diagnostic criterion.

Table 1: **Linear segmentation on ADE20K.**

Training tokens	mIoU	mAcc
All	25.4	33.2
Regular	**27.8**	**35.7**
Spurious	21.9	29.8

4.3 Selective Representation Refinement

We extend UniRefiner's preservation-and-absorption framework (Qiu et al., 2026) to video encoders, using the token partition obtained through spatial latent alignment to guide refinement. The student F_s and teacher F_t are initialized from the same pre-trained encoder F. The teacher remains frozen, and only the student's LoRA parameters are optimized.

Video refinement with temporal boundary registers. We sample spatial crops using shared coordinates across all frames of a clip. Each cropped clip is passed through both F_t and F_{ref} to obtain spatially corresponding teacher and reference feature maps. The teacher features serve as refinement targets, while the reference features are used to assess spatial alignment consistency. We compute consistency between teacher features projected by the frozen g and the corresponding reference features, then apply the calibrated threshold τ to identify regular and spurious teacher locations, denoted by Γ_r and Γ_s.

We zero-pad each frame of the full clip before passing it to the student, using the resulting padding tokens as temporal boundary registers. RoIAlign (He et al., 2017) extracts student features spatially aligned with each teacher crop. Only teacher tokens are partitioned: regular teacher features supervise the corresponding student features, while spurious teacher features supervise the registers.

Regular preservation and spurious absorption. For regular preservation, we reweight the alignment and uniformity terms of InfoNCE (Wang & Isola, 2020). The uniformity term maintains discriminability within each frame, but can also separate locations with similar semantic content; we therefore retain it with a reduced weight. For each cropped frame, let s_i denote the ℓ_2-normalized student feature after RoIAlign and t_i the ℓ_2-normalized teacher feature at the corresponding location:

$$\mathcal{L}_{\text{align}} = -\frac{1}{|\Gamma_r|}\sum_{i\in\Gamma_r}\frac{s_i^\top t_i}{\tau_r}, \qquad \mathcal{L}_{\text{uniform}} = \frac{1}{|\Gamma_r|}\sum_{i\in\Gamma_r}\log\sum_j \exp\left(\frac{s_i^\top t_j}{\tau_r}\right), \tag{4}$$

where τ_r is the refinement temperature and j ranges over all teacher tokens in the same cropped frame. The resulting preservation objective is

$$\mathcal{L}_{\text{pres}} = \mathcal{L}_{\text{align}} + \lambda_{\text{u}}\mathcal{L}_{\text{uniform}}. \tag{5}$$

We set $\lambda_{\text{u}} = 0.3$ and report its sensitivity in the appendix.

Spurious absorption trains the boundary registers to capture the identified spurious teacher features. Each register is matched to its most similar teacher token in Γ_s from the same cropped frame. We optimize these pairs with an InfoNCE objective $\mathcal{L}_{\text{abs}}$, using the other registers in the same frame as negatives. The overall refinement objective combines the two losses:

$$\mathcal{L}_{\text{refine}} = \mathcal{L}_{\text{pres}} + \lambda_{\text{abs}}\mathcal{L}_{\text{abs}}. \tag{6}$$

We set $\lambda_{\text{abs}} = 0.1$ and average the loss over crops and frames.

5 Experiments

We evaluate whether Align2Refine improves video representations across progressively more demanding forms of spatial and temporal localization. We first test frame-level spatial quality using semantic segmentation and monocular depth estimation. We then evaluate region-level correspondence across time through video object segmentation, and finally move to point-level temporal correspondence with zero-shot point tracking. Together, these evaluations test whether refinement improves not only per-frame dense structure but also the ability to preserve local correspondence across time. We evaluate five heterogeneous video encoders, with all backbones frozen during downstream evaluation.

5.1 Implementation Details

Refinement Setup. We evaluate V-JEPA 2 (Assran et al., 2025), MoonViT-V2 (Team et al., 2026), OV-Encoder (Tang et al., 2026), Mage-ViT (Yang et al., 2026), and InternVideo-Next (Wang et al., 2026). All models are refined on the complete YouTube-VOS training split (Xu et al., 2018) using only RGB videos; the provided segmentation masks are never used.

Evaluation Protocols. For semantic segmentation and monocular depth, we follow the frozen-backbone dense probing procedure of DINOv3 (Siméoni et al., 2025). We additionally report mAcc for segmentation and AbsRel/δ_1 for depth to provide a more complete evaluation. Semantic segmentation is evaluated on ADE20K (Zhou et al., 2017), Cityscapes (Cordts et al., 2016), and Pascal VOC (Everingham et al., 2015); monocular depth is evaluated on NYUv2 (Silberman et al., 2012). For video object segmentation, we adopt DINOv3's non-parametric label-propagation procedure and patch-size-matched input resolutions (Siméoni et al., 2025). We evaluate on DAVIS, MOSE, and the official YouTube-VOS validation split. Zero-shot point tracking follows the first-frame query protocol of DiffTrack (Nam et al., 2025) on TAP-Vid (Doersch et al., 2022).

5.2 Image-Level Dense Prediction

We first evaluate frame-level spatial representation quality. Semantic segmentation tests category-aware spatial localization, while monocular depth probes scene geometry. Together, they examine whether refinement improves dense spatial structure across both semantic and geometric prediction.

Table 2: **Frozen-backbone linear probing for semantic segmentation and monocular depth.** We report mIoU and mAcc in percent on ADE20K, Cityscapes, and VOC. On NYUv2, we report RMSE, AbsRel as a ratio, and δ_1 in percent. Highlighted rows denote our refined models.

Model	ViT	ADE20K		Cityscapes		VOC		NYUv2		
		mIoU(↑)	mAcc(↑)	mIoU(↑)	mAcc(↑)	mIoU(↑)	mAcc(↑)	RMSE(↓)	AbsRel(↓)	δ_1(↑)
V-JEPA 2	g/16	25.4	33.2	53.4	62.2	65.2	74.5	0.542	0.186	74.1
+Ours	g/16	**36.4**	**46.2**	**62.2**	**72.0**	**77.9**	**84.6**	**0.454**	**0.152**	**81.4**
MoonViT-V2	0.4B/14	18.7	23.1	46.6	54.8	49.0	55.1	0.572	0.199	71.4
+Ours	0.4B/14	**40.2**	**50.9**	**62.5**	**72.5**	**74.8**	**83.5**	**0.456**	**0.144**	**82.2**
OV-Encoder	L/14	40.0	51.2	59.6	70.6	73.3	81.7	0.529	0.159	77.2
+Ours	L/14	**46.5**	**58.6**	**66.3**	**76.5**	**81.0**	**88.7**	**0.449**	**0.135**	**83.4**
Mage-ViT	L/16	42.1	53.2	60.8	72.2	74.9	83.6	0.523	0.159	77.7
+Ours	L/16	**46.9**	**58.6**	**65.1**	**76.9**	**80.9**	**88.7**	**0.459**	**0.135**	**83.0**
InternVideo-Next	L/14	40.9	52.7	61.1	70.5	76.1	84.2	0.463	0.148	82.4
+Ours	L/14	**45.3**	**57.5**	**65.3**	**75.6**	**81.7**	**89.2**	**0.411**	**0.129**	**86.6**

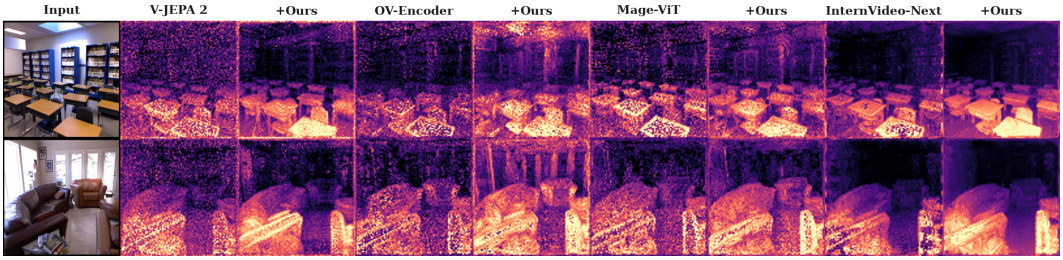

Figure 5: **Depth estimation on NYUv2.** Frozen-backbone depth predictions for Base and Refined representations under the same probing protocol.

Linear probing. Table 2 shows that refinement improves every reported metric for all five video encoders across three segmentation datasets and monocular depth. These gains span both semantic localization and geometric prediction, indicating that the improvement is neither backbone-specific nor tied to a single dense prediction target. For instance, V-JEPA 2 improves by 11.0 mIoU on ADE20K while reducing NYUv2 RMSE from 0.542 to 0.454; OV-Encoder gains 6.5 and 7.7 mIoU on ADE20K and VOC, respectively. The qualitative depth maps in Figure 5 further show cleaner local structure and more coherent object and surface boundaries across different backbones. Appendix C further compares our task-motivated token criterion with UniRefiner on image encoders, where Align2Refine yields stronger dense-prediction results under the corresponding frozen-backbone protocol.

5.3 Video Object Segmentation

Having established improvements in frame-level spatial representations, we next ask whether the refined features support stronger correspondence across time. Video object segmentation evaluates this property at the region level through non-parametric label propagation based on cross-frame patch similarity. No task-specific decoder is trained.

Table 3 shows consistent gains for every backbone, dataset, and resolution setting. Since refinement uses only the YouTube-VOS training split, improvements on the held-out YouTube-VOS validation set demonstrate generalization to unseen videos within the refinement domain, while the gains on DAVIS and MOSE establish cross-dataset transfer. At the higher M resolution, InternVideo-Next improves by 7.7 points on DAVIS, 9.0 on YouTube-VOS, and 6.6 on MOSE. Figure 6 reflects the same trend qualitatively: the base representation often yields fragmented regions during propagation, whereas refinement maintains cleaner object support and more coherent boundaries across time.

5.4 Zero-Shot Point Tracking

Finally, we move from region-level propagation to the stricter setting of point-level temporal correspondence. Zero-shot point tracking tests whether individual spatial locations remain matchable

Table 3: **Video object segmentation.** We report mean $\mathcal{J}\&\mathcal{F}$. For patch sizes 14/16, S and M use short-side resolutions of 420/480 and 840/960 pixels, respectively.

Model	ViT	DAVIS		YouTube-VOS		MOSE	
		S	M	S	M	S	M
V-JEPA 2	g/16	53.4	57.3	52.7	56.3	30.5	33.0
+Ours	g/16	**64.5**	**70.5**	**64.5**	**68.8**	**36.8**	**40.3**
MoonViT-V2	0.4B/14	50.1	63.2	50.5	60.8	25.9	33.6
+Ours	0.4B/14	**63.3**	**73.7**	**63.8**	**72.6**	**35.6**	**41.6**
OV-Encoder	L/14	58.4	63.2	56.8	56.3	32.6	35.3
+Ours	L/14	**65.0**	**71.6**	**65.9**	**66.8**	**37.7**	**40.9**
Mage-ViT	L/16	58.1	61.4	55.4	54.9	31.8	34.4
+Ours	L/16	**65.0**	**70.1**	**64.4**	**65.4**	**36.8**	**39.5**
InternVideo-Next	L/14	60.7	66.0	58.9	64.3	34.0	39.1
+Ours	L/14	**66.6**	**73.7**	**68.5**	**73.3**	**40.1**	**45.7**

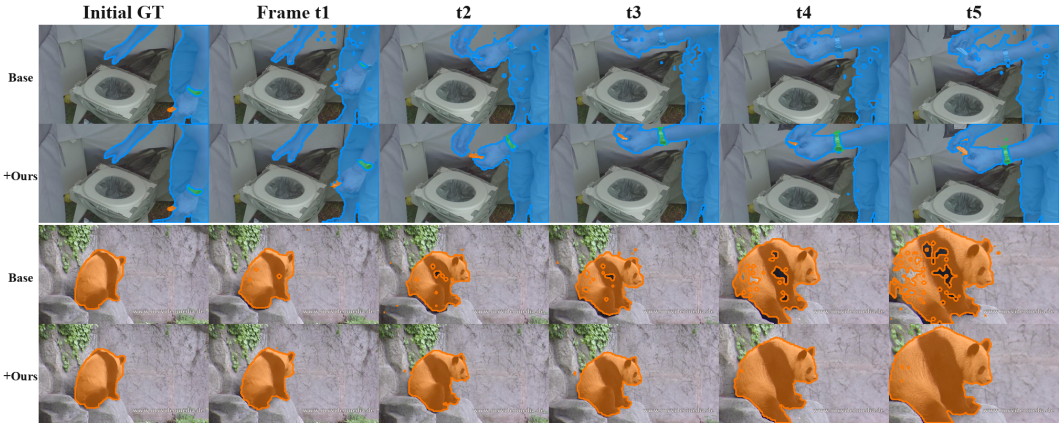

Figure 6: **Video object segmentation with InternVideo-Next ViT-L.** Base and Refined predictions use the same first-frame mask and M-resolution label-propagation protocol.

across frames without task-specific adaptation. We follow DiffTrack's first-frame query protocol on TAP-Vid Kinetics and DAVIS. Localization accuracy is measured at pixel thresholds of 1, 2, 4, 8, and 16, reported as $< \delta^0, \ldots, < \delta^4$ and their average $< \delta^x_{\text{avg}}$. No TAP-Vid annotations are used for training or refinement. Evaluation uses the prescribed first-frame queries and ground-truth tracks to compute localization accuracy.

We report literature baselines from DiffTrack Table 1 and additionally evaluate DINOv3 under the same protocol. Refinement improves V-JEPA 2 and InternVideo-Next in Table 4; Appendix F reports improvements across all five video encoders on both datasets. InternVideo-Next improves $< \delta^x_{\text{avg}}$ from 38.2 to 44.3 on Kinetics and from 42.4 to 47.5 on DAVIS. Its DAVIS score exceeds the best average reported in DiffTrack Table 1, 46.9 for CogVideoX-5B. Figure 7 further shows that refinement reduces cross-frame mismatches and yields more temporally consistent trajectories.

5.5 Ablation Study

Component ablations. For the controlled variants, *w/o* token reweighting assigns the same weight to all tokens during probe training; *w/* random spurious-token mask replaces the detected mask with a random mask of matched sparsity; *w/o* regular preservation removes regular-token preservation supervision; and *w/o* spurious absorption removes the boundary-register absorption objective.

Table 5: **Ablations on V-JEPA 2 (ViT-g).** Results on ADE20K (mIoU) and DAVIS-S ($\mathcal{J}\&\mathcal{F}$).

	ADE20K	DAVIS-S
Original	25.4	53.4
w/o token reweighting (4.1)	34.6	61.2
w/ random spurious-token mask (4.2)	33.4	58.6
w/o regular preservation (4.3)	33.5	59.4
w/o spurious absorption (4.3)	35.6	63.1
Full method	**36.4**	**64.5**

Table 4: **Zero-shot point tracking on TAP-Vid Kinetics and DAVIS.** Baseline grouping follows DiffTrack Table 1; highlighted rows denote ours, with best/second-best values bold/underlined.

Method	TAP-Vid Kinetics						TAP-Vid DAVIS					
	$< \delta^0$	$< \delta^1$	$< \delta^2$	$< \delta^3$	$< \delta^4$	$< \delta^x_{\text{avg}}$	$< \delta^0$	$< \delta^1$	$< \delta^2$	$< \delta^3$	$< \delta^4$	$< \delta^x_{\text{avg}}$
DIFT (SD2.1) (Tang et al., 2023)	3.7	14.9	45.4	70.9	79.6	42.9	3.6	13.3	40.1	65.8	75.7	39.7
DINOv2-Reg (ViT-B/14) (Darcet et al., 2024)	2.8	10.8	32.8	60.6	78.7	37.1	3.0	11.7	36.5	67.5	84.1	40.6
DINOv2 (ViT-B/14) (Oquab et al., 2023)	3.0	11.4	34.6	63.0	80.3	38.4	3.1	12.1	38.7	70.7	85.8	42.1
DINOv3 (ViT-7B/16) (Siméoni et al., 2025)	3.1	12.1	37.4	67.9	84.8	41.1	3.5	13.6	42.4	74.2	88.5	44.4
SMTC (ViT-S/16) (Qian et al., 2023)	4.1	15.5	34.2	54.4	72.1	36.1	2.6	12.1	29.4	52.5	73.0	33.9
Spa-then-Temp (ResNet-50) (Li & Liu, 2023)	5.3	19.4	41.6	58.9	69.7	39.0	3.2	13.8	33.1	53.4	67.5	34.2
SVD (Blattmann et al., 2023)	4.3	16.0	37.9	56.3	69.8	36.6	3.6	14.6	34.1	55.7	71.4	35.9
ZeroCo (CroCo) (An et al., 2025)	**14.5**	22.9	35.9	60.4	79.7	42.6	4.6	8.8	19.5	44.9	65.6	28.7
DiffTrack (HunyuanVideo (Kong et al., 2024))	5.9	22.0	49.1	70.4	80.3	45.5	4.4	18.2	44.8	70.1	82.8	44.1
DiffTrack (CogVideoX-2B (Yang et al., 2025))	6.2	23.3	51.2	71.2	79.9	46.3	4.8	19.4	49.2	73.6	84.3	46.3
DiffTrack (CogVideoX-5B)	6.8	**25.9**	**55.4**	**74.9**	82.7	**49.2**	**5.2**	**20.5**	**50.7**	73.9	84.3	46.9
V-JEPA 2 (ViT-g/16)	3.1	11.2	33.7	61.2	76.2	37.1	3.5	12.9	38.0	64.2	77.6	39.2
+Ours	3.9	14.3	41.3	67.7	79.6	41.4	4.4	15.9	47.8	75.0	84.7	45.6
InternVideo-Next (ViT-L/14)	2.9	11.0	33.5	62.7	81.1	38.2	3.4	12.8	39.6	69.9	86.3	42.4
+Ours	4.1	15.3	44.3	72.5	**85.4**	44.3	4.5	16.7	48.9	**78.3**	**89.3**	**47.5**

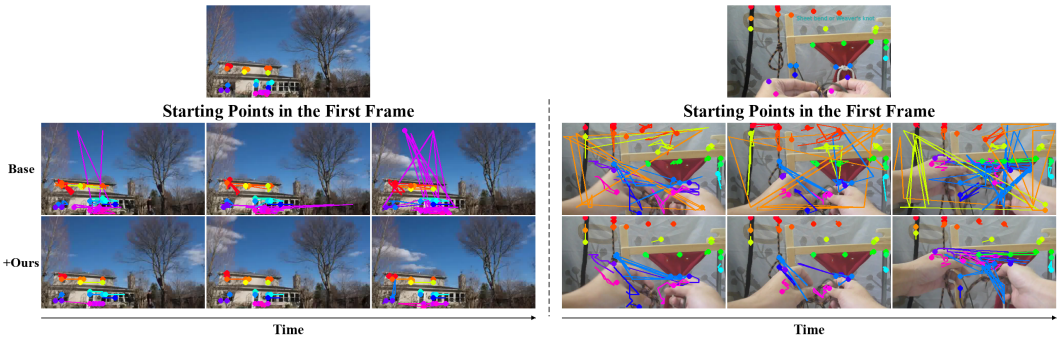

Figure 7: **Zero-shot point tracking with InternVideo-Next ViT-L.** Base and Refined predictions use identical first-frame queries and the same feature-matching procedure.

Replacing the detected spurious-token mask with a random mask of matched sparsity produces the largest drop, showing that consistency-based spurious-token detection provides more useful refinement supervision than random selection. Removing token reweighting, regular preservation, or spurious absorption also degrades performance, supporting the contribution of each component.

Choice of reference encoder. We compare two reference encoders with different visual information: DINOv3 provides high-level semantic features, and the SD3.5 VAE retains low-level visual details. As shown in Table 6, the two reference encoders produce comparable performance. This shows that spatial latent alignment remains effective with both high-level semantic and low-level visual features. Appendix B provides additional results on more video encoders and dense prediction tasks.

Table 6: **Reference ablation.** ADE20K results on V-JEPA 2.

Reference	mIoU	mAcc
None (Original)	25.4	33.2
SD3.5 VAE	36.2	45.9
DINOv3	36.4	46.2

6 Conclusion

We presented Align2Refine, a framework for diagnosing and refining pre-trained video representations on unlabeled videos. We train the alignment probe through a deliberately imbalanced learning process that reinforces differences in fitting progress, making its consistency scores more discriminative for token identification. The resulting partition guides refinement with supervision from the original encoder, consistently improving spatial dense prediction and temporal correspondence across five heterogeneous video encoders. Local prediction performance provides a common basis for token diagnosis across heterogeneous video encoders. Results with both semantic and low-level visual references further support the use of spatial latent alignment to diagnose spurious tokens and guide representation refinement.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and presentation of the manuscript, including grammar, wording, readability, and conciseness. All AI-assisted edits were reviewed and revised by the authors. The authors take full responsibility for the accuracy, claims, and final content of the submission.

REFERENCES

Honggyu An, Jin Hyeon Kim, Seonghoon Park, Jaewoo Jung, Jisang Han, Sunghwan Hong, and Seungryong Kim. Cross-view completion models are zero-shot correspondence estimators. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1103–1115. IEEE, 2025.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

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)*, pp. 9630–9640. IEEE, 2021.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International conference on learning representations*, volume 2024, pp. 2632–2652, 2024.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. Mose: A new dataset for video object segmentation in complex scenes. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20167–20177. IEEE, 2023.

Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*, 35:13610–13626, 2022.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Allan Jabri, Andrew Owens, and Alexei Efros. Space-time correspondence as a contrastive random walk. *Advances in neural information processing systems*, 33:19545–19560, 2020.

Nicholas Jiang, Amil Dravid, Alexei Efros, and Yossi Gandelsman. Vision transformers don't need trained registers. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 56557–56595. Curran Associates, Inc., 2025. doi: 10.52202/085713-1891. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/51bae6441380c4c4cab87f430be89ac2-Paper-Conference.pdf.

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.

Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19948–19960, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Li_Unmasked_Teacher_Towards_Training-Efficient_Video_Foundation_Models_ICCV_2023_paper.html.

Rui Li and Dong Liu. Spatial-then-temporal self-supervised learning for video correspondence. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2279–2288. IEEE, 2023.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

Jisu Nam, Soowon Son, Dahyun Chung, Jiyoung Kim, Siyoon Jin, Junhwa Hur, and Seungryong Kim. Emergent temporal correspondences from video diffusion transformers. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 123683–123727. Curran Associates, Inc., 2025. doi: 10.52202/085713-4122. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/b2fea79b1137d917e8b7cce9434ab5fa-Paper-Conference.pdf.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

Rui Qian, Shuangrui Ding, Xian Liu, and Dahua Lin. Semantics meets temporal correspondence: Self-supervised object-centric learning in videos. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16629–16641. IEEE, 2023.

Congpei Qiu, Zhaoyu Hu, Wei Ke, Zhuotao Tian, Yanhao Wu, and Tong Zhang. Unirefiner: Teaching pre-trained vits to self-dispose dross via contrastive register. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10061–10070, June 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Cheng Shi, Yizhou Yu, and Sibei Yang. Vision transformers need more than registers. *arXiv preprint arXiv:2602.22394*, 2026. URL https://arxiv.org/abs/2602.22394.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pp. 746–760. Springer, 2012.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Feilong Tang, Xiang An, Yunyao Yan, Yin Xie, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Chunyuan Li, Shikun Feng, et al. Onevision-encoder: Codec-aligned sparsity as a foundational principle for multimodal intelligence. *arXiv preprint arXiv:2602.08683*, 2026.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in neural information processing systems*, 36: 1363–1389, 2023.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

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*, 2022. URL https://arxiv.org/abs/2203.12602.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Chenting Wang, Yuhan Zhu, Yicheng Xu, Jiange Yang, Ziang Yan, Yali Wang, Yi Wang, and Limin Wang. Internvideo-next: Towards world-understanding video models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16966–16976, June 2026.

Haoqi Wang, Tong Zhang, and Mathieu Salzmann. SINDER: Repairing the singular defects of DINOv2. *arXiv preprint arXiv:2407.16826*, 2024a. URL https://arxiv.org/abs/2407.16826.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European conference on computer vision*, pp. 396–416. Springer, 2024b.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European conference on computer vision*, pp. 432–448. Springer, 2018.

Jiarui Xu and Xiaolong Wang. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10055–10065. IEEE, 2021.

Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Zipeng Yan, Yinjie Chen, Chong Zhou, Bo Dai, and Andrew Luo. Vision transformers with self-distilled registers. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 137517–137548. Curran Associates, Inc., 2025. doi: 10.52202/085713-4590. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/c8e5ccdcd98d1152361076180d85d384-Paper-Conference.pdf.

Jiawei Yang, Katie Z Luo, Jiefeng Li, Congyue Deng, Leonidas Guibas, Dilip Krishnan, Kilian Q Weinberger, Yonglong Tian, and Yue Wang. Denoising vision transformers. *arXiv preprint arXiv:2401.02957*, 2024. URL https://arxiv.org/abs/2401.02957.

Senqiao Yang, Kaichen Zhang, Zhaoyang Jia, Jinghao Guo, Yifei Shen, Xinjie Zhang, Xiaoyi Zhang, Haoqing Wang, Xiao Li, Peng Zhang, et al. Mage-vl: An efficient codec-native streaming multimodal foundation model. *arXiv preprint arXiv:2607.24904*, 2026.

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. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5122–5130. IEEE, 2017.

A Implementation and Evaluation Details

This appendix details the refinement configuration and downstream evaluation protocols. Unless otherwise stated, all five video encoders use the same refinement recipe, and each Base/Refined pair uses the same downstream protocol. We also report image-encoder comparisons, UPerNet evaluation, uniformity-weight sensitivity, complete tracking results, and additional qualitative examples.

A.1 Refinement Data and Clip Construction

Unlabeled refinement data. We refine all video encoders on the training split of YouTube-VOS (Xu et al., 2018). Only RGB frames are used: object masks, category annotations, and any other task-specific supervision are never accessed during spatial correspondence probing or representation refinement. The official YouTube-VOS validation split is excluded from refinement and is used only for downstream video object segmentation evaluation. Consequently, the refinement stage does not use annotations from any of the downstream benchmarks considered in the paper.

Clip sampling. Training examples consist of temporally contiguous clips of 16 RGB frames. During probe learning, a clip is sampled from each video and forwarded through both the frozen video encoder and the frozen reference encoder. For videos shorter than the required temporal extent, the last available frame is repeated to complete the clip. During refinement, all spatial transformations applied to a clip are temporally consistent: the same crop coordinates are shared by every frame in the clip. This avoids introducing artificial frame-to-frame jitter into the spurious-token mask or the refinement objective.

Spatial token resolution. For the five video encoders studied in the main paper, we use a common 36×36 spatial token lattice during refinement. During refinement, models with patch size 16 use 576×576 inputs, while models with patch size 14 use 504×504 inputs. Reference features are organized on a 36×36 spatial grid to match the video-token lattice before computing correspondence consistency. The concrete input resolution and feature extraction depend on the reference instantiation and are specified in Sec. A.2.

Reference features from the clip are matched to the native temporal units of each video encoder. When a temporal unit spans multiple frames, the corresponding reference features are averaged and re-normalized.

A.2 Spatial Correspondence Probe

We use two frozen reference encoders that provide different visual information: DINOv3 ViT-L/16 (Siméoni et al., 2025) provides high-level semantic features, while the SD3.5 VAE encoder (Esser et al., 2024) retains low-level visual details. Unless otherwise stated, the main experiments use DINOv3 as the reference encoder; Sec. 5.5 compares the two choices. The reference encoder and target video encoder remain frozen during probe learning.

For each video encoder, we train a separate three-layer token-wise MLP g whose weights are shared across token locations:

$$d_v \to 2048 \to 2048 \to d_{\mathrm{ref}}. \tag{7}$$

GELU nonlinearities are inserted between consecutive linear layers. The projector contains no convolution, attention, positional input, or neighborhood aggregation. Its output dimension d_{ref} matches the selected reference: $d_{\mathrm{ref}} = 1024$ for DINOv3 and $d_{\mathrm{ref}} = 512$ for the SD3.5 VAE. Both projected video tokens and reference tokens are ℓ_2-normalized when computing the contrastive correspondence objective.

We optimize the probe for 20 epochs using AdamW with learning rate 10^{-3} and zero weight decay. The contrastive temperature is $\tau_p = 0.07$, and training uses bfloat16 precision. At each epoch, the highest-consistency fraction receives weight 1, while the remaining tokens receive $\beta = 0.1$. For all five backbones, we set $\rho_{\mathrm{start}} = 0.9$ and $\rho_{\mathrm{end}} = 0.6$. The full-weight ratio is linearly annealed from 0.9 to 0.6 during the first $E_w = 10$ probe epochs and remains fixed thereafter.

DINOv3 reference. We use the frozen DINOv3 ViT-L/16 (Siméoni et al., 2025). Frames in each clip are resized to 576×576, producing a 36×36 patch-token grid with 1024-dimensional features.

SD3.5 VAE reference. We use the frozen SD3.5 VAE encoder (Esser et al., 2024). Frames in each clip are resized to 288×288, and we extract the 512-dimensional final encoder activation before the posterior projection, yielding a 36×36 spatial feature grid.

A.3 GMM Calibration and Representation Refinement

Backbone-specific consistency calibration. After probe learning, g is frozen and we collect correspondence consistency scores over the probe-training data. For each backbone, we fit the mixture in Eq. 3 and select the equal-posterior intersection between its component means as τ. The resulting threshold is fixed throughout refinement and requires no downstream annotations or validation results.

Temporally consistent crop-space relocalization. For every training clip, we sample three spatial crop proposals. The crop coordinates are shared across all 16 frames. Crop areas are sampled with scale in $[0.25, 0.5]$ relative to the full frame and aspect ratio in $[2/3, 3/2]$. Each crop is resized to the input resolution corresponding to the common 36×36 token grid. We then recompute the teacher and reference representations on the crop and apply the frozen projector and backbone-specific GMM threshold. Thus, the spurious-token mask is relocalized in crop space rather than being obtained by simply cropping a full-frame mask.

Temporal boundary registers. The teacher is a frozen copy of the original encoder, whereas the student is initialized from the same pre-trained weights and adapted only through LoRA. For the student forward pass, each frame is padded by one patch width on all four sides, creating a one-token-wide ring of temporal boundary registers at every temporal unit. The padding values are initialized to zero. Only the interior feature map is exposed to downstream tasks; the boundary tokens remain active during feature computation but are discarded at the output.

Optimization. Refinement is performed for **four epochs**. We use AdamW with learning rate 10^{-4} and weight decay 0.1. The refinement temperature is $\tau_r = 0.2$. For the regular-token preservation objective in Eq. 5, we use $\lambda_{\mathrm{u}} = 0.3$. For the spurious-token absorption objective in Sec. 4.3, we use $\lambda_{\mathrm{abs}} = 0.1$. All refinement experiments use bfloat16 precision.

We apply LoRA (Hu et al., 2022) to the linear transformations inside the transformer blocks of the video encoder. We use rank $r = 8$, scaling parameter $\alpha = 16$, dropout 0.1, and no LoRA bias. The pre-trained backbone parameters, the reference encoder, the token-wise projector g, and the teacher encoder remain frozen. Only the LoRA parameters of the student are optimized. Table 7 summarizes the default probe and refinement configuration used across backbones.

A.4 Backbone-Specific Feature Extraction

Table 8 summarizes the spatial configuration used during refinement. The input resolution is chosen such that all five encoders expose the same 36×36 spatial token lattice. For all downstream experiments, we use dense representations from the final transformer layer unless explicitly stated otherwise.

A.5 Image-Level Dense Prediction

We follow the frozen-backbone dense probing protocol of DINOv3 (Siméoni et al., 2025) for both semantic segmentation and monocular depth estimation. The backbone is kept completely frozen. We extract final-layer dense patch representations after the backbone's output normalization and apply a trainable batch-normalization layer followed by a linear prediction head. Base and refined encoders use exactly the same data processing, probe architecture, optimization grid, and evaluation code.

For semantic segmentation, we evaluate on ADE20K (Zhou et al., 2017), Cityscapes (Cordts et al., 2016), and Pascal VOC (Everingham et al., 2015). For each benchmark, we train the linear probe with AdamW and sweep learning rate over

$$\left\{10^{-4},\ 3 \times 10^{-4},\ 10^{-3}\right\} \tag{8}$$

Table 7: **Default Align2Refine training configuration.** These settings are shared across the five video encoders unless otherwise noted.

Component	Setting
Refinement data	YouTube-VOS training split, RGB only
Clip length	16 frames
Spatial token grid	36×36
Reference encoder (main setting)	DINOv3 ViT-L/16, frozen
Reference dimension (main setting)	1024
Projector	3-layer token-wise MLP, hidden dim. 2048
Projector activation	GELU
Probe epochs	20
Probe optimizer	AdamW, lr 10^{-3}, wd 0
Probe temperature τ_p	0.07
Full-weight token ratio	$\rho_{\text{start}} = 0.9,\ \rho_{\text{end}} = 0.6$
Full-weight annealing horizon	10 epochs
Low-weight factor β	0.1
GMM	2 components, one fit per backbone
Refinement crops	3 temporally consistent crops / clip
Crop scale	$[0.25, 0.5]$
Crop aspect ratio	$[2/3, 3/2]$
Temporal boundary register width	1 token
Refinement epochs	**4**
Refinement optimizer	AdamW, lr 10^{-4}, wd 0.1
Refinement temperature τ_r	0.2
λ_{u}	0.3
λ_{abs}	0.1
LoRA	$r = 8$, $\alpha = 16$, dropout $= 0.1$
Precision	bfloat16

Table 8: **Backbone configurations used for refinement.** Input resolution is selected to produce a common 36×36 spatial token grid.

Model	Architecture	Patch	Input	Token grid
V-JEPA 2	ViT-g	16	576^2	36^2
MoonViT-V2	0.4B	14	504^2	36^2
OV-Encoder	ViT-L	14	504^2	36^2
Mage-ViT	ViT-L	16	576^2	36^2
InternVideo-Next	ViT-L	14	504^2	36^2

and weight decay over

$$\left\{10^{-4},\ 10^{-3}\right\}. \tag{9}$$

We report mean Intersection-over-Union (mIoU) and mean class accuracy (mAcc).

Monocular depth estimation is evaluated on NYUv2 (Silberman et al., 2012) using the same frozen-feature probing principle and the same optimizer search space. In addition to RMSE, we report AbsRel and δ_1. Base and Refined representations use the same learning-rate and weight-decay search spaces.

A.6 Video Object Segmentation

We follow the non-parametric label-propagation procedure of DINOv3 (Siméoni et al., 2025). The backbone remains frozen and no segmentation decoder is trained. The ground-truth mask of the first frame is converted to patch-level instance labels. For every subsequent frame, correspondence is obtained from cosine similarity between the current patch features and features from the first frame together with a limited temporal context of previous frames. For each query patch, label propagation aggregates the labels of the most similar features within a spatial neighborhood. Predicted labels from previously processed frames are retained as soft labels for subsequent propagation.

Table 9: **Extended comparison of reference encoders.** We compare DINOv3 and the SD3.5 VAE across three video encoders under the same refinement and frozen-backbone evaluation protocols.

Model	Reference	ADE20K		Cityscapes		VOC		NYUv2		
		mIoU(↑)	mAcc(↑)	mIoU(↑)	mAcc(↑)	mIoU(↑)	mAcc(↑)	RMSE(↓)	AbsRel(↓)	δ_1(↑)
V-JEPA 2	Baseline	25.4	33.2	53.4	62.2	65.2	74.5	0.542	0.186	74.1
	SD3.5 VAE	36.2	45.9	61.9	71.8	77.8	84.3	0.461	0.153	81.1
	DINOv3	36.4	46.2	62.2	72.0	77.9	84.6	0.454	0.152	81.4
Mage-ViT	Baseline	42.1	53.2	60.8	72.2	74.9	83.6	0.523	0.159	77.7
	SD3.5 VAE	46.3	57.9	64.8	76.6	80.4	88.1	0.464	0.138	82.5
	DINOv3	46.9	58.6	65.1	76.9	80.9	88.7	0.459	0.135	83.0
InternVideo-Next	Baseline	40.9	52.7	61.1	70.5	76.1	84.2	0.463	0.148	82.4
	SD3.5 VAE	44.8	57.1	64.9	75.1	81.3	88.7	0.411	0.128	86.8
	DINOv3	45.3	57.5	65.3	75.6	81.7	89.2	0.411	0.129	86.6

Similarity computation and propagation are carried out directly in the dense feature space. The resulting patch-level mask probabilities are bilinearly resized to the native video resolution before computing the standard $\mathcal{J}\&\mathcal{F}$ metric. Following DINOv3, the label-propagation hyperparameters are selected on the DAVIS training set and are then kept fixed across DAVIS, YouTube-VOS, and MOSE.

We evaluate two resolutions in the main paper. For patch-size-14 encoders, the short side is 420 pixels in the **S** setting and 840 pixels in the **M** setting. For patch-size-16 encoders, the corresponding resolutions are 480 and 960 pixels. Images preserve their aspect ratio up to the nearest resolution compatible with the patch lattice. Only objects annotated in the first frame are propagated and evaluated.

We report results on DAVIS (Pont-Tuset et al., 2017), the held-out YouTube-VOS validation split (Xu et al., 2018), and MOSE (Ding et al., 2023). No VOS labels are used to adapt either the base or refined backbone.

A.7 Zero-Shot Point Tracking

Zero-shot point tracking follows the evaluation protocol of DiffTrack (Nam et al., 2025) on TAP-Vid Kinetics and DAVIS (Doersch et al., 2022). Following DiffTrack, we initialize query points in the first video frame and match them to subsequent frames. All point localization metrics are computed at an evaluation resolution of 256×256.

Following DiffTrack, a predicted point is counted as correct when its distance from a visible ground-truth point is below one of five pixel thresholds,

$$\{1, 2, 4, 8, 16\}, \tag{10}$$

reported as $< \delta^0, \ldots, < \delta^4$. For models evaluated in this work, $< \delta^x_{\text{avg}}$ is the arithmetic mean across the five thresholds. Literature baseline scores are reproduced as reported in DiffTrack. Two reported averages differ from the arithmetic means of their displayed threshold scores: SVD on Kinetics reports 36.6 rather than approximately 36.9, and DINO on DAVIS reports 37.3 rather than approximately 36.9. We retain the published values for these baselines.

For all video encoders evaluated in this work, we use the dense output of the final transformer layer. The same feature extraction and matching pipeline is used for the base and refined encoder. No TAP-Vid annotations are used for training or refinement. Evaluation uses the prescribed first-frame queries and ground-truth tracks to compute localization accuracy. Unlike video diffusion models evaluated through DiffTrack, our discriminative video encoders do not require diffusion inversion, denoising-timestep selection, or diffusion-layer selection; the final dense representation is used directly.

B Extended Analysis of Reference Encoders

Table 9 extends the main-paper reference comparison to additional video encoders across all image-level dense-prediction benchmarks. DINOv3 and the SD3.5 VAE use the same refinement and frozen-backbone evaluation protocols.

Table 10: **Comparison with UniRefiner on image encoders.** UniRefiner results are taken from the original paper, while **+Ours** denotes Align2Refine under our refinement setup. Results use frozen-backbone dense prediction on all benchmarks.

Model	ViT	ADE20K		Cityscapes		VOC		NYUv2		
		mIoU(↑)	mAcc(↑)	mIoU(↑)	mAcc(↑)	mIoU(↑)	mAcc(↑)	RMSE(↓)	AbsRel(↓)	δ_1(↑)
CLIP	G/14	39.8	52.4	58.6	68.2	74.6	83.8	0.485	0.147	81.3
+UniRefiner	G/14	42.1	54.9	61.7	71.5	76.7	86.0	0.421	0.125	85.9
+Ours	G/14	**45.2**	**57.0**	**64.1**	**73.1**	**80.6**	**88.2**	**0.399**	**0.117**	**87.5**
SigLIP 2	So/16	45.6	58.6	63.4	74.6	77.8	85.3	0.469	0.143	82.0
+UniRefiner	So/16	49.8	62.9	67.9	78.5	82.1	89.9	0.387	0.112	88.6
+Ours	So/16	**50.9**	**63.4**	**68.7**	**79.1**	**83.9**	**90.4**	**0.373**	**0.108**	**89.5**

Across all three video encoders, DINOv3 and the SD3.5 VAE produce similar improvements across semantic segmentation and monocular depth estimation. These results further show that spatial latent alignment remains effective with reference encoders that provide substantially different visual information.

C Comparison with UniRefiner on Image Encoders

UniRefiner (Qiu et al., 2026) is the closest post-hoc refinement method and uses a different criterion to identify spurious tokens. We compare it with Align2Refine on CLIP (Radford et al., 2021) and SigLIP 2 (Tschannen et al., 2025) to examine how the token identification criterion affects dense representation refinement.

For these image encoders, Align2Refine uses the same spatial correspondence probing and selective refinement objectives as in the video experiments. The spatial token lattice is 32×32: CLIP uses a 448×448 input with patch size 14, and SigLIP 2 uses a 512×512 input with patch size 16. The UniRefiner values are taken from the original paper under the corresponding frozen-backbone dense-prediction protocol.

Table 10 shows consistent improvements across semantic segmentation and monocular depth. For CLIP, Align2Refine reaches 45.2 mIoU on ADE20K compared with 42.1 for UniRefiner, while reducing NYUv2 RMSE from 0.421 to 0.399. On SigLIP 2, Align2Refine improves ADE20K mIoU from 49.8 to 50.9 and reduces NYUv2 RMSE from 0.387 to 0.373. These results show that our correspondence-based token criterion also provides effective refinement supervision for image encoders.

D Evaluation with a Stronger Dense Decoder

The main dense-prediction experiments intentionally use lightweight linear probing to expose the quality of the frozen representation with minimal task-specific capacity. We additionally evaluate whether refinement gains extend to UPerNet with multi-level frozen features.

We train UPerNet (Xiao et al., 2018) on ADE20K while keeping the backbone entirely frozen. For a transformer with L blocks, we extract features at approximately one-quarter, one-half, three-quarters, and the full network depth, i.e., $\{\lfloor L/4 \rfloor, \lfloor L/2 \rfloor, \lfloor 3L/4 \rfloor, L\}$. These four feature maps are provided to the standard UPerNet multi-level decoder.

We follow the standard ADE20K UPerNet training recipe: 512×512 training crops, 160k optimization iterations, and AdamW with learning rate 6×10^{-5}, $\beta = (0.9, 0.999)$, and weight decay 0.01. The learning rate uses a 1500-iteration linear warm-up followed by polynomial decay. The backbone remains frozen throughout training, and only the UPerNet decoder is optimized. For every backbone, Base and Refined use exactly the same feature layers, decoder architecture, optimization schedule, data processing, and inference protocol.

As shown in Table 11, refinement improves V-JEPA 2 from 43.6 to 45.9 mIoU and InternVideo-Next from 50.2 to 51.4 mIoU. The corresponding mAcc also increases for both models. Refinement also

Table 11: **Frozen-backbone semantic segmentation with UPerNet on ADE20K.** Only the UPerNet decoder is optimized. Highlighted rows denote refined representations.

Model	ViT	ADE20K	
		mIoU(↑)	mAcc(↑)
V-JEPA 2	g/16	43.6	54.4
+Ours	g/16	**45.9**	**56.5**
InternVideo-Next	L/14	50.2	62.7
+Ours	L/14	**51.4**	**63.6**

Table 12: **Sensitivity to the uniformity weight λ_{u} on V-JEPA 2 ViT-g.** We report frozen-backbone ADE20K mIoU while keeping all other refinement and evaluation settings fixed. The default setting is $\lambda_{\mathrm{u}} = 0.3$.

λ_{u}	0	0.1	**0.3**	0.5	0.7	1.0
ADE20K mIoU	35.9	36.2	**36.4**	36.0	35.5	34.8

improves both tested backbones under multi-level UPerNet evaluation, extending the observed gains beyond final-layer linear probing.

E Sensitivity to the Uniformity Weight

The regular-token preservation objective in Eq. 5 prioritizes location-preserving teacher alignment through $\mathcal{L}_{\mathrm{align}}$, while $\mathcal{L}_{\mathrm{uniform}}$ provides a complementary spatial-discriminability constraint. The main experiments use $\lambda_{\mathrm{u}} = 0.3$. We study the sensitivity to this weighting while keeping all other components of Align2Refine unchanged.

We perform the analysis using V-JEPA 2 ViT-g and evaluate the resulting frozen representation on ADE20K. Setting $\lambda_{\mathrm{u}} = 0$ removes the uniformity term from the regular-token preservation objective, reducing $\mathcal{L}_{\mathrm{pres}}$ to the location-preserving align term. The spurious-token absorption objective remains unchanged with $\lambda_{\mathrm{abs}} = 0.1$.

As shown in Table 12, among the tested values, $\lambda_{\mathrm{u}} = 0.3$ yields the highest ADE20K mIoU. Removing the uniformity term reduces mIoU from 36.4 to 35.9, while $\lambda_{\mathrm{u}} = 0.1$ and 0.5 yield 36.2 and 36.0, respectively. Increasing the weight further to 0.7 and 1.0 results in lower mIoU of 35.5 and 34.8. These results favor a moderate uniformity weight: adding the uniformity term improves over alignment-only preservation, while increasing its weight beyond 0.3 lowers mIoU in this setting.

F Complete Zero-Shot Point Tracking Results

The main paper reports a compact set of representative tracking baselines. Here we provide the complete comparison corresponding to Table 1 of DiffTrack (Nam et al., 2025), together with DINOv3 and all five video encoders studied in this work.

The grouping in Table 13 is based on the origin and training regime of the representation rather than on tracking performance.

Image-based vision foundation models. The first group contains representations learned primarily from single images. DINO, DINOv2, and DINOv2-Reg are image self-supervised vision transformers, while DIFT extracts correspondence features from text-to-image diffusion models. We additionally include DINOv3 ViT-B/16 and ViT-7B/16. This group tests how much temporal correspondence emerges from strong dense image representations without video-specific pre-training.

Video and correspondence-oriented models. The second group contains video- and correspondence-oriented baselines reported by DiffTrack. These methods obtain correspondence cues from videos or paired/cross-view visual observations rather than relying purely on single-image representations. The group therefore provides a complementary comparison to image foundation

Table 13: **Complete zero-shot point tracking comparison on TAP-Vid Kinetics and DAVIS.** Reported results for DINO, DINOv2-Reg, DINOv2, DIFT, SMTC, CRW, Spa-then-Temp, VFS, SVD, ZeroCo, and the DiffTrack variants are taken from DiffTrack (Nam et al., 2025). DINOv3 and the five video encoders are evaluated by us under the same zero-shot point-tracking protocol. Highlighted rows denote Align2Refine.

Method	TAP-Vid Kinetics						TAP-Vid DAVIS					
	$< \delta^0$	$< \delta^1$	$< \delta^2$	$< \delta^3$	$< \delta^4$	$< \delta^x_{\text{avg}}$	$< \delta^0$	$< \delta^1$	$< \delta^2$	$< \delta^3$	$< \delta^4$	$< \delta^x_{\text{avg}}$
Image-based vision foundation models												
DINO (ViT-B/16) (Caron et al., 2021)	2.8	10.9	33.9	58.2	74.8	36.1	2.9	10.7	34.3	59.8	77.0	37.3
DINOv2-Reg (ViT-B/14) (Darcet et al., 2024)	2.8	10.8	32.8	60.6	78.7	37.1	3.0	11.7	36.5	67.5	84.1	40.6
DINOv2 (ViT-B/14) (Oquab et al., 2023)	3.0	11.4	34.6	63.0	80.3	38.4	3.1	12.1	38.7	70.7	85.8	42.1
DIFT (SD1.5) (Tang et al., 2023)	3.7	14.6	44.6	69.0	77.5	41.9	3.5	13.0	39.3	63.1	72.2	38.2
DIFT (SD2.1) (Tang et al., 2023)	3.7	14.9	45.4	70.9	79.6	42.9	3.6	13.3	40.1	65.8	75.7	39.7
DINOv3 (ViT-B/16) (Siméoni et al., 2025)	3.0	11.7	36.1	65.8	82.5	39.8	3.4	13.2	40.9	70.9	85.0	42.7
DINOv3 (ViT-7B/16) (Siméoni et al., 2025)	3.1	12.1	37.4	67.9	84.8	41.1	3.5	13.6	42.4	74.2	88.5	44.4
Video and correspondence-oriented models												
SMTC (ViT-S/16) (Qian et al., 2023)	4.1	15.5	34.2	54.4	72.1	36.1	2.6	12.1	29.4	52.5	73.0	33.9
CRW (ResNet-18) (Jabri et al., 2020)	5.2	19.4	42.7	62.9	74.3	40.9	3.1	13.9	34.7	57.1	70.5	35.9
Spa-then-Temp (ResNet-50) (Li & Liu, 2023)	5.3	19.4	41.6	58.9	69.7	39.0	3.2	13.8	33.1	53.4	67.5	34.2
VFS (ResNet-50) (Xu & Wang, 2021)	5.4	20.1	44.6	65.4	76.6	42.4	3.5	15.2	37.2	60.8	75.2	38.4
SVD (Blattmann et al., 2023)	4.3	16.0	37.9	56.3	69.8	36.6	3.6	14.6	34.1	55.7	71.4	35.9
ZeroCo (CroCo) (An et al., 2025)	14.5	22.9	35.9	60.4	79.7	42.6	4.6	8.8	19.5	44.9	65.6	28.7
Video diffusion transformers with DiffTrack												
DiffTrack (HunyuanVideo (Kong et al., 2024))	5.9	22.0	49.1	70.4	80.3	45.5	4.4	18.2	44.8	70.1	82.8	44.1
DiffTrack (CogVideoX-2B (Yang et al., 2025))	6.2	23.3	51.2	71.2	79.9	46.3	4.8	19.4	49.2	73.6	84.3	46.3
DiffTrack (CogVideoX-5B (Yang et al., 2025))	6.8	25.9	55.4	74.9	82.7	49.2	5.2	20.5	50.7	73.9	84.3	46.9
Discriminative video encoders evaluated in this work												
V-JEPA 2 (ViT-g/16) (Assran et al., 2025)	3.1	11.2	33.7	61.2	76.2	37.1	3.5	12.9	38.0	64.2	77.6	39.2
+Ours	3.9	14.3	41.3	67.7	79.6	41.4	4.4	15.9	47.8	75.0	84.7	45.6
MoonViT-V2 (0.4B/14) (Team et al., 2026)	2.8	10.1	29.8	54.0	73.1	34.0	2.7	9.9	31.5	58.9	78.8	36.4
+Ours	4.0	15.0	43.5	69.5	81.5	42.7	4.2	15.8	47.6	74.7	85.2	45.5
OV-Encoder (ViT-L/14) (Tang et al., 2026)	3.5	13.2	38.7	64.1	77.2	39.4	3.4	13.6	40.8	67.0	80.1	41.0
+Ours	3.9	14.5	42.2	67.7	79.7	41.6	3.7	14.5	43.7	69.6	81.7	42.6
Mage-ViT (ViT-L/16) (Yang et al., 2026)	3.3	12.1	36.1	63.8	79.3	38.9	3.7	14.5	43.9	73.8	87.3	44.6
+Ours	3.8	14.0	41.2	69.0	82.5	42.1	4.2	15.7	46.8	76.3	87.8	46.2
InternVideo-Next (ViT-L/14) (Wang et al., 2026)	2.9	11.0	33.5	62.7	81.1	38.2	3.4	12.8	39.6	69.9	86.3	42.4
+Ours	4.1	15.3	44.3	72.5	85.4	44.3	4.5	16.7	48.9	78.3	89.3	47.5

models, covering representations whose training signal is more directly related to temporal or geometric correspondence.

Video diffusion transformers with DiffTrack. The third group contains HunyuanVideo and CogVideoX representations accessed through DiffTrack. Unlike the discriminative encoders studied in this work, these models are generative video diffusion transformers and DiffTrack identifies correspondence from selected internal diffusion representations. We keep this group separate because its representation extraction mechanism and pre-training objective are substantially different.

Discriminative video encoders studied in this work. The final group contains V-JEPA 2, MoonViT-V2, OV-Encoder, Mage-ViT, and InternVideo-Next. These are the direct refinement targets of Align2Refine. For each encoder, the **Base** and **+Ours** rows use the same model architecture, the same final-layer feature extraction, and the same point-tracking evaluation pipeline. These paired comparisons evaluate the effect of Align2Refine under the same downstream tracking pipeline.

Across all five video encoders, Align2Refine improves the Base representation under the same zero-shot point-tracking protocol on both TAP-Vid Kinetics and DAVIS. The improvements are observed across every reported localization threshold for all five encoders, rather than being concentrated at a particular tolerance level.

G Additional PCA Visualizations

The quantitative evaluations demonstrate consistent improvements after refinement across both spatial and temporal downstream tasks. To complement these results, we provide additional visualizations of how the dense representations change after refinement. We use exactly the same Base-fitted PCA visualization protocol as in Figure 1 and compare the final-layer patch representations of the Base and Refined encoders at three timestamps $t_a < t_b < t_c$ from each video clip.

For each encoder–video pair, we fit a single PCA basis and channel scaling using only Base tokens from all displayed timestamps. The same PCA basis and scaling are then applied unchanged to the corresponding Refined tokens, with the first three principal components mapped to RGB. Thus, Base and Refined representations are visualized in the same Base-defined PCA coordinate system, without allowing Refined features to influence the projection.

For the Refined representation, boundary-register tokens are excluded before PCA, such that the visualization uses only the interior patch tokens exposed to downstream evaluation. Figures 8 and 9 provide four additional video examples covering all five video encoders considered in the main experiments, using the same visualization protocol as Figure 1.

H Additional Qualitative Results

We provide additional qualitative results for the two temporal correspondence evaluations considered in the paper: video object segmentation and zero-shot point tracking. These examples complement the representation-level visualizations in Sec. G by showing how the refined features affect region-level propagation and point-level correspondence under the same downstream protocols used for the Base representations.

H.1 Additional Video Object Segmentation Results

We provide additional qualitative comparisons of video object segmentation using InternVideo-Next and V-JEPA 2. For every sequence, Base and **+Ours** receive identical first-frame ground-truth masks and use exactly the same non-parametric label-propagation procedure; only the frozen representation is changed. We visualize the propagated masks over multiple subsequent frames to compare the temporal evolution of the predicted object regions.

Figures 10 and 11 provide additional examples under this controlled setting. Across the displayed sequences, **+Ours** generally produces more spatially concentrated and less fragmented propagated masks than Base.

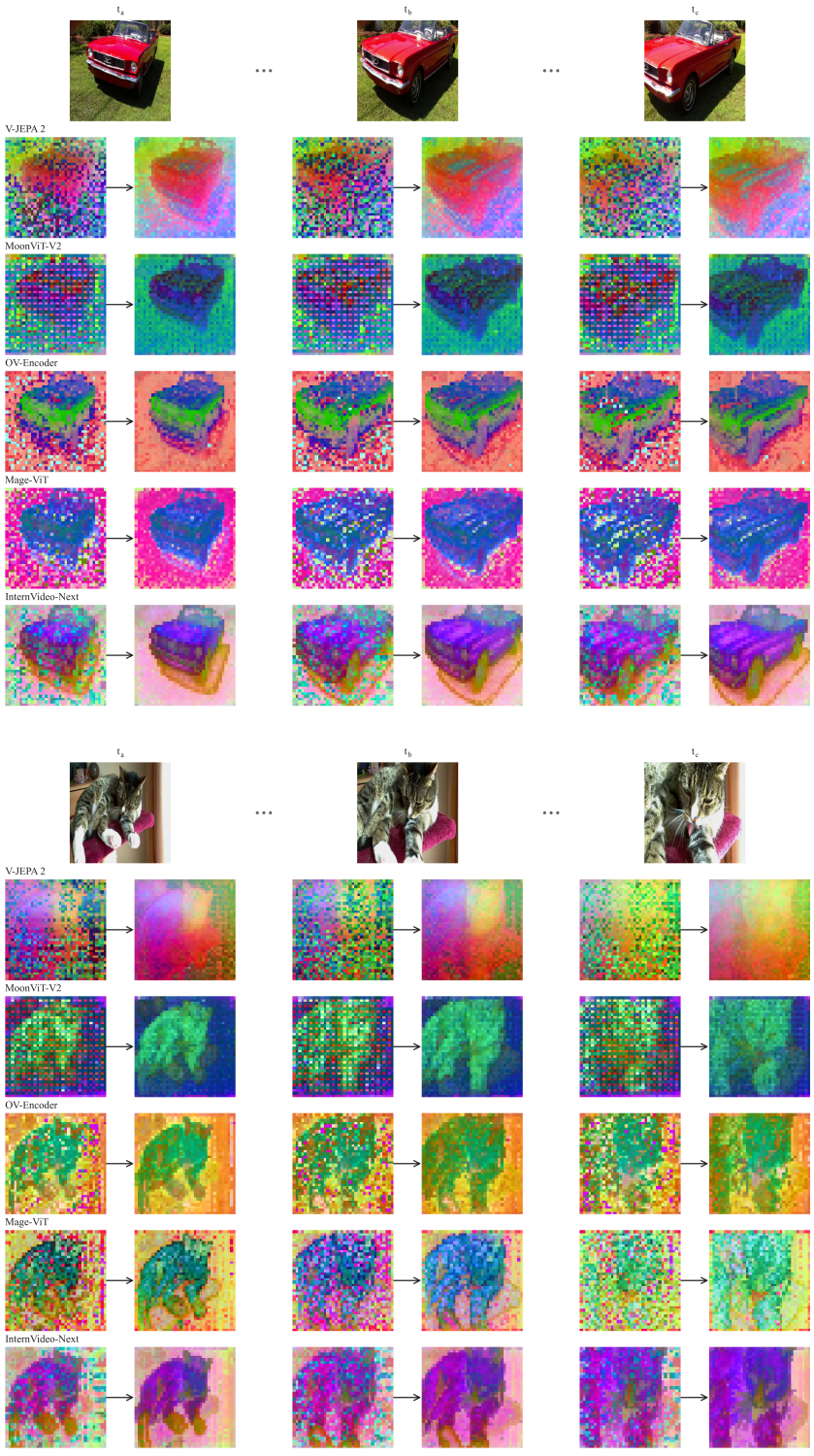

Figure 8: **Additional PCA visualizations of Base and Refined representations.** We show two video examples at three timestamps $t_a < t_b < t_c$. For every encoder and timestamp, Base is shown on the left of the arrow and Refined on the right. Following Figure 1, for each encoder–video pair, the PCA basis and channel scaling are fitted using only Base tokens from all displayed timestamps and reused unchanged for Refined tokens. Results are shown for V-JEPA 2, MoonViT-V2, OV-Encoder, Mage-ViT, and InternVideo-Next.

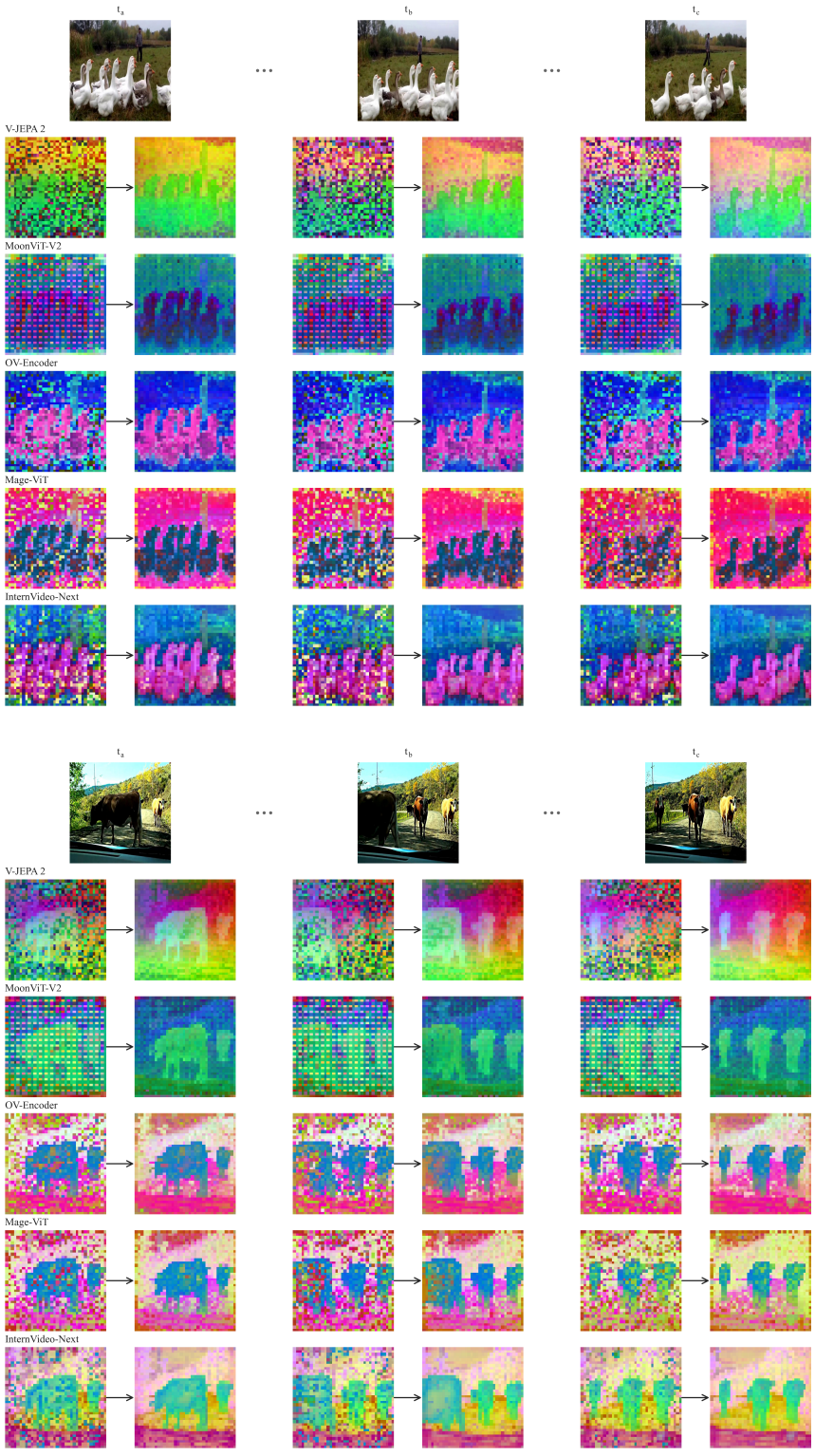

Figure 9: **Additional PCA visualizations (continued).** Two further video examples are shown using the same Base-fitted PCA protocol as Figure 8. For every encoder and timestamp, Base is shown on the left of the arrow and Refined on the right. The same five video encoders are shown.

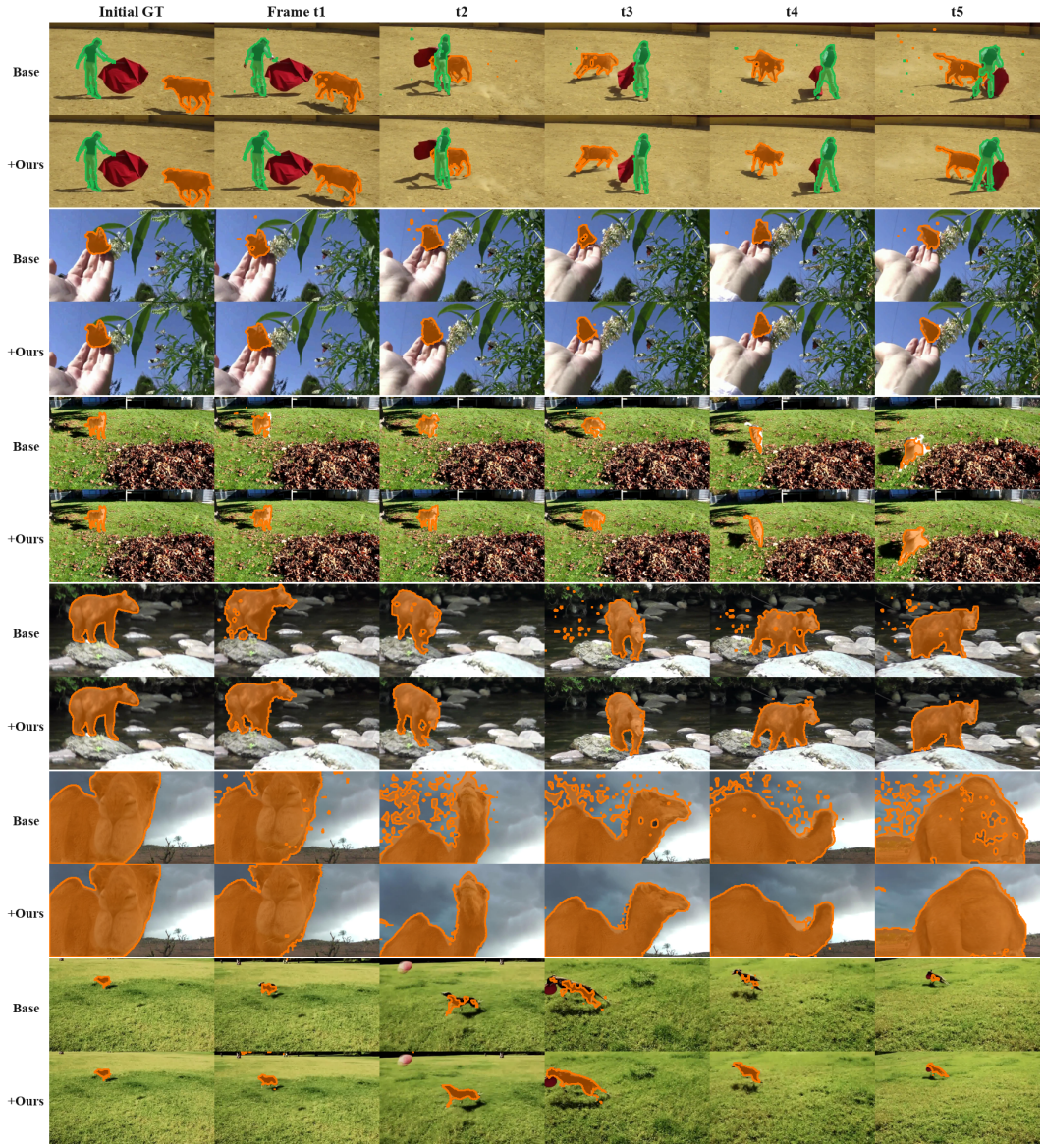

Figure 10: **Additional video object segmentation results with InternVideo-Next.** Each example shows the ground-truth object mask in the initial frame followed by propagated predictions at five subsequent timestamps. Base and **+Ours** use identical first-frame supervision and the same non-parametric label-propagation procedure. The Base results are shown in the upper row of each example and **+Ours** in the lower row. No VOS-specific training is used, and both backbones remain frozen throughout evaluation.



Figure 11: **Additional video object segmentation results with V-JEPA 2 ViT-g.** We show further examples using the same visualization and label-propagation protocol as Figure 10. Base and **+Ours** receive identical first-frame ground-truth masks, with the Base results shown in the upper row and **+Ours** in the lower row of each example. No VOS-specific training is used, and both backbones remain frozen throughout evaluation.

H.2 Additional Zero-Shot Point Tracking Results

We provide additional qualitative comparisons of zero-shot point tracking using InternVideo-Next and V-JEPA 2. For every video, Base and Refined representations receive identical first-frame query points and use exactly the same feature-matching procedure; only the frozen representation is changed. The colored trajectories visualize the resulting point correspondences from the same set of initial query locations over time.

Figures 12 and 13 show additional examples across diverse video content and motion patterns. Consistent with the quantitative results in Sec. F, the Refined representations exhibit more localized point correspondences across the displayed examples, while the Base representations more frequently produce spatially dispersed matches.



Figure 12: **Additional zero-shot point tracking results with InternVideo-Next.** Each example shows the query points initialized in the first frame, followed by tracking results at subsequent timestamps. Base and **+Ours** use identical first-frame queries and the same matching procedure. The Base results are shown in the upper row of each example and **+Ours** in the lower row; colors identify the corresponding query trajectories across time. No tracking-specific training is used, and both backbones remain frozen throughout evaluation.

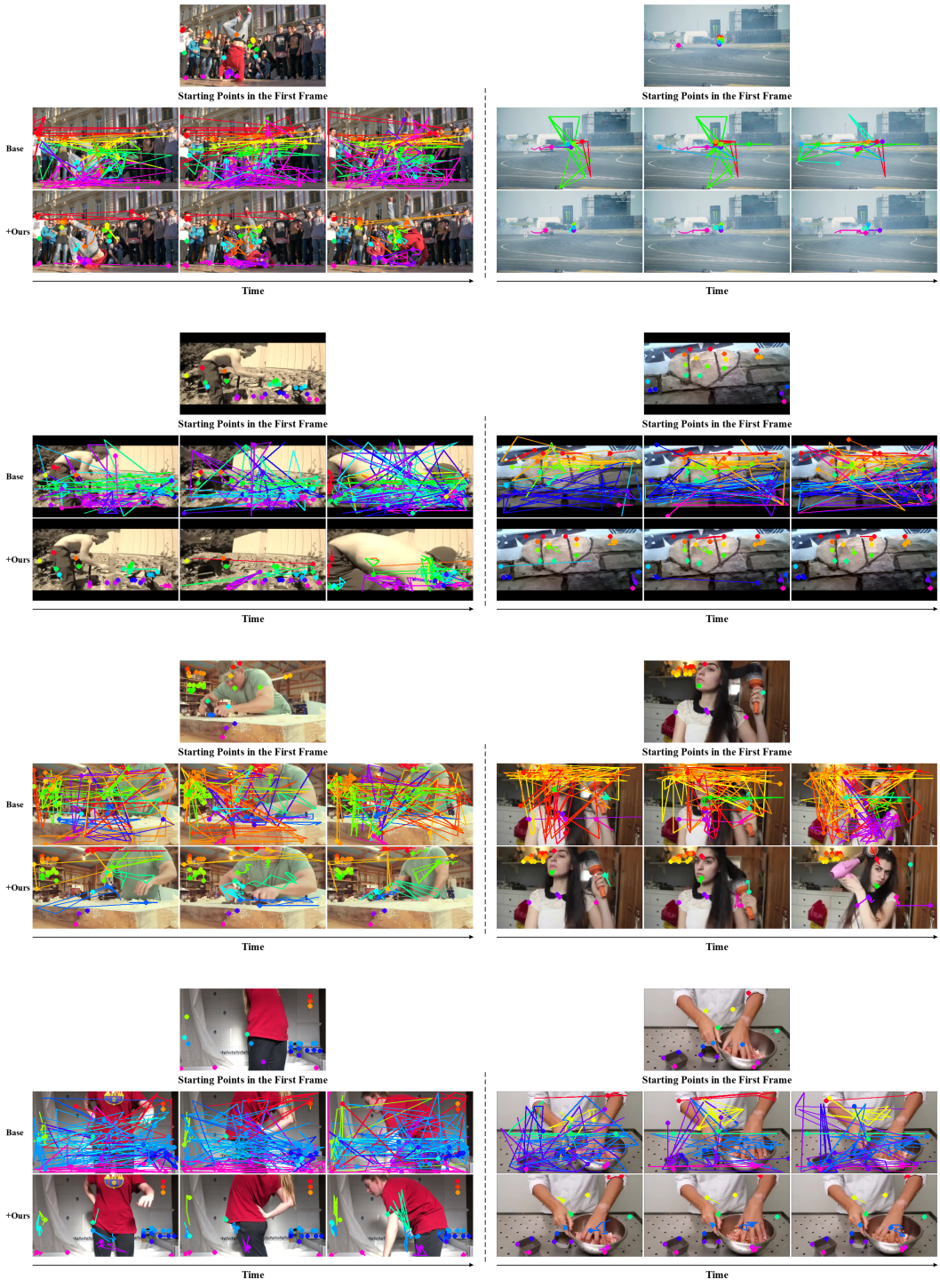

Figure 13: **Additional zero-shot point tracking results with V-JEPA 2 ViT-g.** We show further examples using the same visualization protocol as Figure 12. Base and **+Ours** receive identical first-frame query points, with the Base results shown in the upper row and **+Ours** in the lower row of each example. No tracking-specific training is used, and both backbones remain frozen throughout evaluation.