LEARNING FROM NEXT-FRAME PREDICTION: TOWARDS UNIFIED AND SCALABLE VISUAL WORLD MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video provides a natural source of supervision: what happens next serves as a learning signal for visual world modeling. Yet masked video pretraining focuses on recovering missing content from surrounding context rather than predicting the future from the past. Autoregressive methods instead predict forward one token at a time, but have yet to match joint-embedding methods in representation quality. This gap motivates us to revisit the granularity and space of future prediction. In this work, we introduce VANE, a self-supervised visual pretraining framework built on next-frame prediction in embedding space. A causal encoder summarizes past observations for a state predictor that infers the representation of the entire next frame, making temporal evolution the primary learning signal. This prediction is carried out in a fixed embedding space provided by a frozen semantic tokenizer, emphasizing scene content and structure. To retain fine-grained spatial detail alongside semantic content, a lightweight auxiliary generative decoder reconstructs future-frame patches conditioned on the predicted representations. To unify visual modeling across images and videos, we construct a local-to-global crop sequence for still images, allowing the same encoder to predict newly revealed regions from earlier crops under the same objective. Experiments show that VANE improves with model and data scaling, achieving SOTA on most anticipation, recognition, and dense perception benchmarks among unified encoders. Especially on EPIC-KITCHENS-100 action anticipation, it surpasses V-JEPA 2.1 by 7.1 points in mean-class recall@5. These results demonstrate that next-frame prediction in embedding space is effective and scalable for visual world modeling.

1 INTRODUCTION

Humans come to understand the world by watching it unfold, mapping a continuous stream of sensory input into a stable sense of objects, motion, and events. What they acquire is not merely a catalogue of appearances, but an internal model of how the world evolves (Rao & Ballard, 1999; LeCun, 2022): what a scene contains, what is happening within it, and what is likely to happen next. Video captures the world in precisely this form, with temporal order providing a natural learning signal that requires no annotation. Predicting what comes next therefore requires a model to represent both the appearance and the dynamics of the visual world.

Yet most video pretraining methods do not learn by predicting what comes next. Masked video models (Feichtenhofer et al., 2022), such as VideoMAE (Tong et al., 2022), hide a large fraction of spatiotemporal patches and reconstruct their pixels from the visible ones (He et al., 2022). Joint-embedding methods (Assran et al., 2023), such as V-JEPA (Bardes et al., 2024; Assran et al., 2025), likewise mask part of the video, but predict representations of the masked regions instead of their pixels (Fig. 1a). Both remain forms of contextual recovery: the model infers what is missing from surrounding observations rather than predicting the future from the past. Autoregressive pretraining (Chen et al., 2020; El-Nouby et al., 2024), extended to video by Toto (Rajasegaran et al., 2025), instead predicts forward in time by treating video as a sequence of visual tokens and learning through next-token prediction. While viable and scalable, this formulation has not matched joint-embedding methods in representation quality (Rajasegaran et al., 2025). This raises a central question: *at what granularity, and in what space, should the future be predicted?*

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.

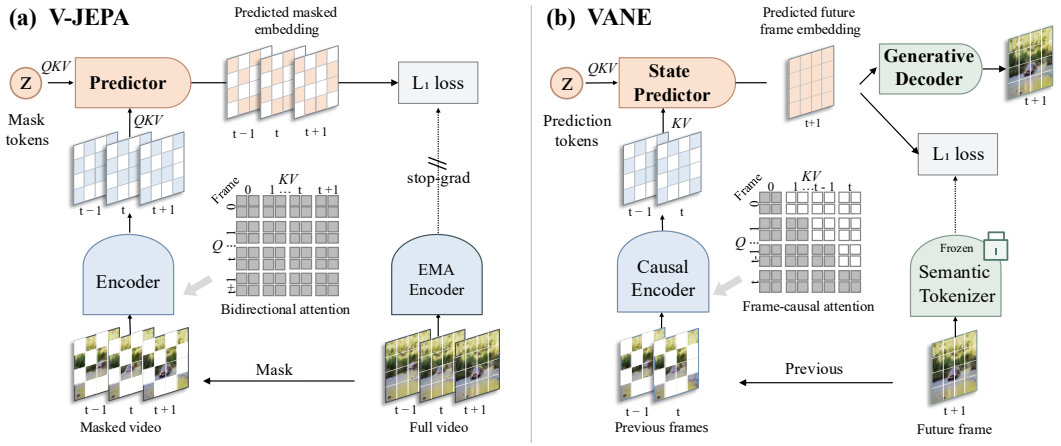


Figure 1: **From masked completion to next-frame prediction in embedding space.** (a) V-JEPA masks part of a clip, encodes the visible patches with bidirectional attention, and predicts the embeddings of the masked regions, supervised by an EMA encoder of the full video. (b) VANE encodes previous frames with frame-causal attention and predicts the embedding of the entire next frame from them, with targets from a frozen semantic tokenizer. An auxiliary generative decoder, used only during pretraining, regularizes the predicted embedding to retain fine-grained spatial detail.

In this work, we answer this question with VANE, a self-supervised visual pretraining framework built on **next-frame prediction in embedding space**. Specifically, VANE shifts prediction from individual tokens to the frame level and performs it in a semantic representation space. At the frame level, the distinction between spatial and temporal structure becomes explicit: structure within a frame is largely spatial, whereas changes across consecutive frames capture how objects, actions, and scenes evolve over time. Frame-level prediction therefore places temporal evolution at the center of the learning objective, rather than devoting much of the objective to completing individual tokens within the same frame. Additionally, prediction is carried out in semantic embedding space, building on the principle of joint-embedding learning (Assran et al., 2023; Bardes et al., 2024). Semantic representations emphasize the higher-level content and structure of a scene, allowing the model to focus on the aspects of the visual world that persist and evolve over time.

We instantiate this objective with four components (Fig. 1b). A causal encoder summarizes the observed history through frame-causal attention, with each frame attending only to itself and preceding frames. From this temporal context, a state predictor infers the representation of the next frame. A frozen semantic tokenizer, implemented as a pretrained image encoder (Oquab et al., 2024), provides a fixed semantic space in which future scene content and structure are predicted. As in generative models operating in fixed latent spaces (Rombach et al., 2022), we keep the tokenizer frozen to provide fixed prediction targets and support stable optimization. Finally, a lightweight generative decoder reconstructs each future-frame patch conditioned on its predicted embedding. This auxiliary reconstruction objective complements semantic prediction by encouraging the model to learn representations that preserve fine-grained spatial detail.

To unify visual modeling across images and videos, we extend the same predictive objective to still images. Unlike videos, still images lack a temporal order, so we introduce **local-to-global observation**: each image is presented as a sequence of local-to-global crops, and the model predicts representations of newly revealed regions in the next crop from preceding crops, just as it predicts the next frame from preceding frames in video. This enables joint image–video training, yielding a unified visual encoder without separate pretraining stages or modality-specific objectives.

Extensive experiments demonstrate that VANE unifies predictive and perceptual understanding within a single pretrained visual model. VANE achieves state-of-the-art performance among unified image–video encoders on most evaluated benchmarks spanning anticipation, global recognition, and dense perception. On EPIC-KITCHENS-100 (Damen et al., 2022), VANE achieves 45.5 action mean-class recall@5, improving over V-JEPA 2.1 (Mur-Labadia et al., 2026) by 7.1 points. Beyond its broad transferability, VANE scales consistently with both model capacity and pretraining data. Together,

these results establish next-frame prediction in embedding space as a unified and scalable pretraining paradigm for visual foundation models.

Our contributions are as follows:

- We formulate visual pretraining as next-frame prediction in embedding space. Instead of completing masked content or reconstructing pixels, the model predicts the semantic embedding of the next frame from past frames, learning directly from how scenes evolve over time.
- We instantiate this objective by combining a causal encoder–predictor, a semantic tokenizer, and auxiliary generative decoding, then extend it to images through local-to-global observation for unified and scalable image–video pretraining.
- We demonstrate that next-frame prediction in embedding space is an effective and scalable visual pretraining paradigm, achieving state-of-the-art performance among unified encoders across most anticipation, global recognition, and dense perception benchmarks.

2 Related Work

Masked visual pretraining. Masked visual pretraining learns representations by recovering content withheld from an observation. Building on masked autoencoding (He et al., 2022), VideoMAE (Tong et al., 2022; Wang et al., 2023a) and MAE-ST (Feichtenhofer et al., 2022) reconstruct masked spatiotemporal patches in pixel space, while the InternVideo series (Wang et al., 2022; 2024; 2025) incorporates masked modeling into broader video pretraining frameworks. The Joint-Embedding Predictive Architecture (JEPA) (LeCun, 2022) formulates prediction in representation space rather than pixel space. I-JEPA (Assran et al., 2023) instantiates this principle for images, while the V-JEPA family (Bardes et al., 2024; Assran et al., 2025; Mur-Labadia et al., 2026) extends it to video, scaling masked feature prediction and improving global and dense representations. Related methods distill pretrained visual features (Wang et al., 2023b; Li et al., 2023b), and Rethinking JEPA (Li et al., 2026) shows that a frozen target encoder can support effective masked latent prediction. InternVideo-Next (Wang et al., 2026) combines semantic alignment, generative reconstruction, and fixed-target feature prediction within a two-stage masked modeling framework. Despite differences in target construction and supervision, these pretraining objectives remain centered on recovering or matching content within the observed clip rather than explicitly predicting its temporal continuation. By predicting future-frame embeddings from past observations, VANE makes temporal evolution central to pretraining while preserving fine-grained spatial structure through auxiliary reconstruction.

Autoregressive and predictive visual learning. Autoregressive visual pretraining depends on how visual observations are represented and ordered. iGPT (Chen et al., 2020) and AIM (El-Nouby et al., 2024) predict pixels or patches, while Toto (Rajasegaran et al., 2025) learns from images and videos through next-token prediction over discrete visual codes. Alternative formulations vary the prediction order, granularity, or target space: ARVideo (Ren et al., 2025) randomizes spatiotemporal ordering, VAR (Tian et al., 2024) predicts progressively finer-resolution token maps, and NEPA (Xu et al., 2025) predicts successive image patch embeddings with stop-gradient targets. A complementary line learns from future visual features through regression (Vondrick et al., 2016) or contrastive objectives (van den Oord et al., 2018; Han et al., 2019). More recently, DINO-Foresight (Karypidis et al., 2025) and DINO-world (Baldassarre et al., 2025) forecast future features for scene understanding, while DINO-WM (Zhou et al., 2025) uses action-conditioned feature prediction for planning. Causal-JEPA (Nam et al., 2026) explores object-level latent masking to learn interaction-aware dynamics. These recent world models learn prediction on top of pretrained visual encoders that remain fixed during dynamics learning. VANE instead uses next-frame embedding prediction to train the causal encoder itself, extending the same objective to unified image–video pretraining through local-to-global image views.

3 Method

This section presents VANE (Fig. 2). We first formulate next-frame prediction in embedding space (Sec. 3.1). Realizing this objective raises three questions, which we address in turn: how to encode the observed frames without access to the future (Sec. 3.2), how to predict the next frame from this encoding (Sec. 3.3), and how to retain the fine-grained spatial detail that semantic targets abstract away (Sec. 3.4). Finally, we extend the objective to still images, which lack temporal order, for unified image–video pretraining (Sec. 3.5).

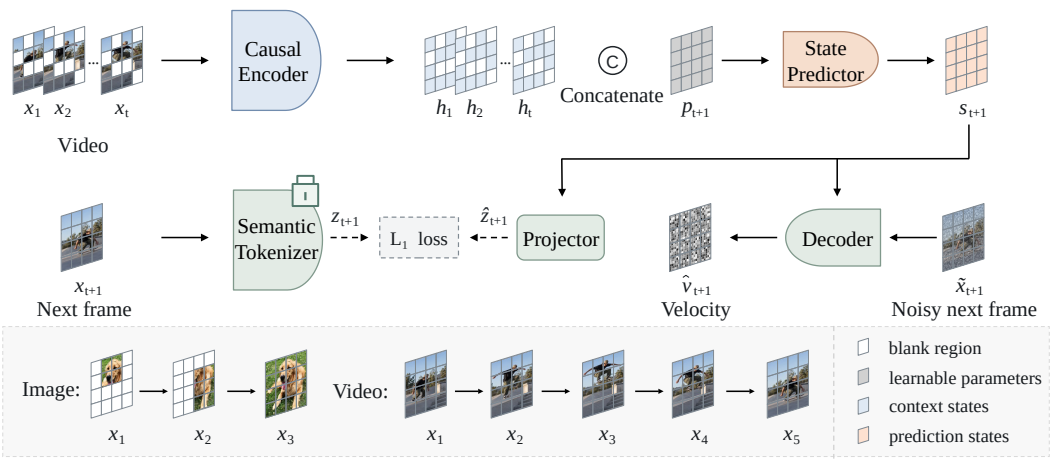


Figure 2: **Overview of VANE.** VANE predicts the embedding of the next frame from the frames observed so far. *Top:* the causal encoder maps the observed frames $x_1, \ldots, x_t$ to context states $h_1, \ldots, h_t$, which are concatenated with prediction tokens p_{t+1}; the state predictor uses the context states only as keys and values and outputs the predicted state s_{t+1}. *Middle:* a projector maps s_{t+1} to the embedding space of the frozen semantic tokenizer, where it is supervised by the embedding of the next frame x_{t+1} with an L_1 loss; the generative decoder, conditioned on s_{t+1}, predicts the flow velocity $\hat{v}_{t+1}$ from the noisy next frame $\tilde{x}_{t+1}$. *Bottom:* images are presented as local-to-global crops and videos as frame sequences, so both share the same objective.

3.1 Next-Frame Prediction in Embedding Space

To make temporal evolution the primary learning signal, VANE learns by predicting what comes next. Given a video clip $(x_1, \ldots, x_N)$ of N frames, the model predicts each frame x_{t+1} from the frames observed so far, $x_{\leq t}$. Every transition in the clip thus provides supervision, and each intermediate frame serves both as a target given its predecessors and as context for its successors.

Rather than predicting pixels, which would require modeling low-level appearance details, we predict the next frame in a semantic embedding space. This space is defined by a frozen semantic tokenizer F, a pretrained image encoder that maps each frame to embeddings representing its content and structure while abstracting over appearance variation. We keep F fixed throughout pretraining so that the target space cannot co-adapt with the model to simplify the prediction task, as in generative models trained in fixed latent spaces (Rombach et al., 2022). The objective is thus

$$\min_{f} \sum_{t=1}^{N-1} d\Big(\underbrace{f\big(x_{\leq t}\big)}_{\text{prediction}}, \underbrace{F\big(x_{t+1}\big)}_{\text{target}}\Big), \tag{1}$$

where f is the predictive model and d is a distance in the embedding space (instantiated in Sec. 3.3).

3.2 Causal Encoder

We realize f with a causal encoder followed by a state predictor. To supervise every transition in a clip within a single forward pass, the encoder must represent each frame using only the frames observed so far; otherwise the target of a transition would leak into its own context. The encoder E_θ therefore processes the context frames $x_1, \ldots, x_{N-1}$ with *frame-causal* attention (Fig. 1b): attention is unidirectional across frames, so that tokens of frame t never attend to later frames, and bidirectional within each frame, since the patches of a frame are observed at the same time and have no natural order. The resulting context state h_t thus depends only on frames up to t:

$$h_t = E_\theta\big(x_{\leq t}\big), \qquad 1 \leq t < N. \tag{2}$$

We apply tube masking (Tong et al., 2022; Bardes et al., 2024) to the context frames, so that E_θ receives only their visible patches, while the prediction targets remain unmasked (Appendix C.3).

3.3 State Predictor and Semantic Objective

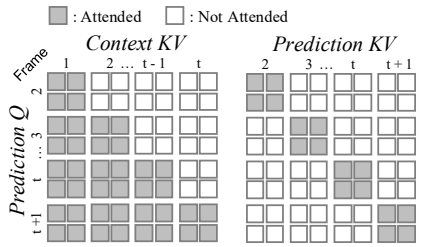


Figure 3: **State predictor attention.** Context states supply only keys and values; prediction tokens attend within their own target frame.

The state predictor P_ϕ then infers the representation of the next frame from the encoded history. It places copies of a learned token p at the positions of frame $t+1$, forming prediction tokens p_{t+1}. As in the encoder, the prediction tokens of all target frames are processed together in a single forward pass, under two attention constraints (Fig. 3). Under *context isolation*, context states supply only keys and values and are not updated inside the predictor, so that the encoder, rather than the predictor, must carry the historical information needed for prediction. Under *target isolation*, prediction tokens for frame $t+1$ attend only to $h_{\le t} = (h_1, \dots, h_t)$ and to one another, so that each frame is predicted jointly over its spatial extent from its own history alone, independently of the other targets:

$$s_{t+1} = P_\phi\big(p_{t+1}; h_{\le t}\big), \qquad 1 \le t < N, \tag{3}$$

where s_{t+1} is the predicted state of frame $t+1$. To supervise the prediction in the semantic embedding space, a learned projection g_ω maps the predicted state to the tokenizer space, $\hat{z}_{t+1} = g_\omega(s_{t+1})$, and we minimize its distance to the frozen target $z_{t+1} = F(x_{t+1})$ with an L_1 loss:

$$\mathcal{L}_{\text{sem}} = \big\|\hat{z}_{t+1,j} - z_{t+1,j}\big\|_1, \tag{4}$$

where $\hat{z}_{t+1,j}$ and $z_{t+1,j}$ denote the prediction and target at patch j of frame $t+1$, averaged over all transitions t and supervised patches j. The encoder, predictor, and projection thus together form the predictive model of Eq. (1): $f(x_{\le t}) = g_\omega\big(P_\phi(p_{t+1}; h_{\le t})\big)$, with $h_{\le t}$ given by Eq. (2).

3.4 Auxiliary Generative Reconstruction

Semantic prediction emphasizes abstract visual content but may underrepresent fine-grained spatial information useful for dense perception. We therefore introduce a lightweight auxiliary generative decoder D_ψ that regularizes each predicted state through patch-level reconstruction. Let $x_{t+1,j}$ and $s_{t+1,j}$ denote the pixels and predicted state of patch j of frame $t+1$. The decoder reconstructs each supervised patch conditioned on $s_{t+1,j}$, and the reconstruction loss is back-propagated through the predictor into the encoder.

Since the next frame is not fully determined by the past, a deterministic pixel regressor would be driven toward the average of plausible outcomes. We therefore model the distribution of each patch given its predicted state, as in per-token diffusion heads (Li et al., 2024), and formulate the decoder as a conditional flow matching model (Lipman et al., 2023). Each supervised patch is corrupted as $\tilde{x}_{t+1,j} = (1-\tau)x_{t+1,j} + \tau\epsilon_{t+1,j}$ with $\epsilon_{t+1,j} \sim \mathcal{N}(0, I)$ and $\tau \sim \mathcal{U}[0,1]$, and the decoder predicts the velocity $\epsilon_{t+1,j} - x_{t+1,j}$ with an L_1 loss:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{\tau,\epsilon}\big\|D_\psi\big(\tilde{x}_{t+1,j}, \tau, s_{t+1,j}\big) - (\epsilon_{t+1,j} - x_{t+1,j})\big\|_1, \tag{5}$$

averaged over the same transitions and supervised patches as $\mathcal{L}_{\text{sem}}$.

3.5 Unified Image–Video Pretraining

To train a unified encoder on both images and videos with the same objective, we arrange each still image into an ordered sequence of views through local-to-global observation. The sequence consists of two local crops x_1 and x_2 of increasing area followed by a larger global crop x_3 ($K = 3$; Appendix C.2). Like successive frames, successive crops reveal new content that must be inferred from what has been observed so far (examples in Appendix Fig. 7). Each crop is treated as one frame by the encoder and predictor, and all crops are aligned to the patch grid of the full image. Let c_t be the binary mask of the region covered by crop t on this grid. Supervision is restricted to the newly revealed region, i.e., the part of crop $t+1$ covered by none of the preceding crops:

$$m_{t+1} = c_{t+1} \odot \prod_{i=1}^{t}\big(1 - c_i\big), \tag{6}$$

where $\odot$ denotes element-wise multiplication, and both losses are averaged over the positions where $m_{t+1} = 1$. Regions covered by earlier crops are excluded because they are already visible in the

Table 1: **Action anticipation on EPIC-KITCHENS-100.** Mean-class recall@5 (%) for verb, noun, and action on the validation set, with a one-second anticipation gap. Bold denotes the best result.

Method	Param.	Verb	Noun	Action
InAViT (Roy et al., 2024)	160M	51.9	52.0	25.8
Video-LLaMA (Zhang et al., 2023)	7B	52.9	52.0	26.0
PlausiVL (Mittal et al., 2024)	8B	55.6	54.2	27.6
V-JEPA 2 ViT-g (Assran et al., 2025)	1B	63.6	57.1	39.7
V-JEPA 2.1 ViT-g (Mur-Labadia et al., 2026)	1B	63.6	56.2	38.4
VANE (ours)	1B	**65.7**	**67.6**	**45.5**

context, so predicting them would amount to reconstructing the input rather than inferring unseen content. This progressively expanding context transfers better than both a fixed order over image cells and masked-image completion (Table 5b).

Training objective. We jointly pretrain on images and videos by combining semantic prediction and generative reconstruction:

$$\mathcal{L} = \mathcal{L}_{\text{sem}} + \lambda_{\text{flow}}\, \mathcal{L}_{\text{flow}}, \tag{7}$$

where λ_{flow} weights the reconstruction loss. Both losses are computed separately on video and image batches and averaged with equal weight, and all components except the frozen tokenizer are optimized jointly (Appendix C).

4 Experiments

We evaluate whether next-frame prediction provides an effective objective for visual world modeling on anticipation, global classification, and dense perception tasks, and then ablate the design choices.

Implementation details. Our main model pairs a 1B-parameter ViT-g encoder with a frozen DINOv2 ViT-L (Oquab et al., 2024) semantic tokenizer. It is pretrained on VISTA, our curated image–video corpus, which combines public image datasets (ImageNet-21K (Deng et al., 2009), SA-1B (Kirillov et al., 2023)) and video datasets (SSv2 (Goyal et al., 2017), K710 (Kay et al., 2017; Li et al., 2023a), HowTo100M (Miech et al., 2019)) with additional retrieved and class-balanced data (Appendix B). Training proceeds in four stages: warm-up and base pretraining on 16-frame clips at 256×256, long-video post-training on 64-frame clips, and a high-resolution cooldown at 384×384. The first two stages take about 26.6k B200 GPU-hours and the last two a further 30.7k (Appendices A and C).

Evaluation protocol. All evaluations use frozen features: the pretrained encoder is kept fixed, and a lightweight probe is trained on its features for each task. Probes follow the protocols of V-JEPA 2.1 (Mur-Labadia et al., 2026) unless otherwise stated. We compare primarily with V-JEPA 2 and V-JEPA 2.1 at the same 1B scale; other encoders are either taken from the literature or re-evaluated under the same protocol, as marked in each table.

4.1 Anticipation

We first examine whether next-frame pretraining supports anticipation, on action anticipation (EK100) and short-term object-interaction anticipation (Ego4D).

Action anticipation. On EPIC-KITCHENS-100 (EK100) (Damen et al., 2022), we follow the V-JEPA 2.1 (Mur-Labadia et al., 2026) protocol with a one-second anticipation gap and report mean-class recall@5 on the validation set. The attentive probe uses only frozen encoder features; the state predictor is not used. VANE reaches 45.5 on actions, exceeding the 1B V-JEPA 2.1 and V-JEPA 2 by 7.1 and 5.8 points, respectively (Table 1), with the largest gain on nouns (67.6 vs. 56.2). It also outperforms the 8B PlausiVL by 17.9 points. These gains suggest that next-frame prediction in embedding space encourages representations that capture how visual events evolve and what is likely to happen next.

Short-term object-interaction anticipation. Ego4D STA v2 (Grauman et al., 2022) additionally requires localizing the next-active object and predicting its category, the interaction verb, and the time to contact. Following V-JEPA 2.1, we train an attentive probe with frame-guided temporal pooling and a detection head on frozen features and report Top-5 AP and mAP. VANE reaches 7.81 mAP All,

Table 2: **Short-term object-interaction anticipation on Ego4D STA v2.** The model localizes the next-active object and predicts its noun, the interaction verb, and the time to contact. Each column reports Top-5 AP or mAP counting a prediction as correct only if all listed components are correct: box (b), noun (N), verb (V), and time to contact (δ). Bold denotes the best results.

Model	AP b	AP $b{+}N$	AP $b{+}V$	AP $b{+}\delta$	AP $b{+}N{+}V$	AP $b{+}\delta{+}N$	AP $b{+}\delta{+}V$	AP All	mAP N	mAP $N{+}V$	mAP $N{+}\delta$	mAP All
Results reported in the literature												
StillFast (Ragusa et al., 2023)	—	—	—	—	—	—	—	—	20.2	10.3	7.17	3.96
STAformer (Mur-Labadia et al., 2024)	38.3	28.3	16.6	12.3	12.8	8.89	5.47	4.06	29.4	15.3	9.94	5.67
Evaluated using the same protocol												
DINOv2 ViT-g (Oquab et al., 2024)	47.8	38.1	23.6	14.1	19.1	10.9	7.29	5.73	30.6	16.8	9.37	5.44
DINOv3 ViT-H+ (Siméoni et al., 2025)	48.9	40.2	23.4	13.6	19.4	10.8	6.71	5.45	32.4	17.2	9.53	5.20
DINOv3 ViT-7B (Siméoni et al., 2025)	50.2	40.9	23.8	14.2	19.8	11.5	7.33	5.82	33.8	17.4	10.3	5.68
V-JEPA 2 ViT-g (Assran et al., 2025)	45.7	34.4	22.2	16.7	16.9	12.1	8.56	6.26	27.2	14.9	10.4	6.02
V-JEPA 2.1 ViT-g (Mur-Labadia et al., 2026)	47.3	36.5	23.6	18.1	18.2	13.5	9.55	7.21	28.7	15.5	11.4	6.75
VANE ViT-g	**57.9**	**46.4**	**24.6**	**21.8**	**19.9**	**17.1**	**10.2**	**8.18**	**37.6**	**17.7**	**14.7**	**7.81**

Table 3: **Global classification.** Top-1 accuracy (%) of attentive probes on frozen features, covering appearance understanding (IN1K, K400) and motion understanding (SSv2, Diving-48). Avg. is reported for models evaluated on all four tasks. Bold denotes the best result in image–video encoders.

Method	**Param.**	**Appearance Understanding**		**Motion Understanding**		**Avg.**
		IN1K	K400	SSv2	Diving-48	
Results reported in the literature						
Toto (Rajasegaran et al., 2025)	1B	75.3	74.4	–	–	–
VideoMAEv2 (Wang et al., 2023a)	1B	71.4	82.8	56.1	–	–
InternVideo2-1B (Wang et al., 2024)	1B	–	87.9	67.3	–	–
VideoPrism (Zhao et al., 2024)	1B	–	87.6	68.5	71.3	–
DINOv3 (Siméoni et al., 2025)	7B	88.4	87.8	70.1	–	–
Image encoders evaluated under the same protocol						
DINOv2 (Oquab et al., 2024)	1.1B	86.1	83.6	50.7	82.5	75.7
DINOv3 ViT-H+ (Siméoni et al., 2025)	0.8B	88.2	86.6	63.1	85.3	80.8
PE_{core} G (Bolya et al., 2025)	1.9B	87.6	88.5	55.4	76.9	77.1
SigLIP 2 (Tschannen et al., 2025)	1.2B	88.0	87.3	49.9	75.3	75.1
Image–video encoders evaluated under the same protocol						
V-JEPA (Bardes et al., 2024)	630M	75.9	82.0	71.4	87.9	79.3
V-JEPA 2 (Assran et al., 2025)	1B	84.6	86.5	75.5	**90.1**	84.2
V-JEPA 2.1 (Mur-Labadia et al., 2026)	1B	83.7	86.2	75.3	89.0	83.6
VANE (ours)	1B	**87.5**	**88.6**	**75.9**	89.7	**85.4**

compared with 6.75 for V-JEPA 2.1 and 6.02 for V-JEPA 2, and improves box AP by 10.6 points over V-JEPA 2.1 (Table 2). It outperforms all encoders evaluated under the same protocol, including DINOv3 ViT-7B, on all twelve metrics. Together with EK100, these results suggest that VANE anticipates not only what happens next but also where and when it happens. Qualitative examples are provided in Appendix D.1.

4.2 Global Classification

To assess recognition beyond anticipation, we evaluate image classification on IN1K (Deng et al., 2009) and action recognition on K400, SSv2, and Diving-48 (Kay et al., 2017; Goyal et al., 2017; Li et al., 2018). Following V-JEPA 2.1, we train four-layer attentive probes on frozen features with 256×256 inputs and report top-1 accuracy. Under the same protocol, VANE outperforms V-JEPA 2.1 on all four benchmarks, by 3.8 points on IN1K, 2.4 on K400, 0.6 on SSv2, and 0.7 on Diving-48, raising the average from 83.6 to 85.4 (Table 3); V-JEPA 2 remains slightly better on Diving-48 (90.1 vs. 89.7). VANE also exceeds the published results of Toto, which autoregressively predicts future visual tokens, by 12.2 points on IN1K and 14.2 on K400. Against image encoders, VANE nearly matches DINOv3 ViT-H+ on IN1K (87.5 vs. 88.2) while exceeding it by 12.8 points on SSv2 (75.9 vs. 63.1), suggesting that unified image–video pretraining retains appearance understanding while adding motion understanding. Notably, it also surpasses DINOv2 ViT-g, a larger variant of its own frozen tokenizer, by 25.2 points on SSv2 (75.9 vs. 50.7), indicating that next-frame prediction learns temporal structure absent from its targets.

Table 4: **Dense perception.** Semantic segmentation (mIoU, %) and monocular depth estimation (RMSE, lower is better) with a linear probe on frozen final-layer patch features. Param./patch denotes encoder size/patch size. Bold denotes the best result among image–video encoders.

Method	Param./patch	Segmentation			Depth	
		ADE20K	Citysc.	VOC	NYUv2↓	KITTI↓
Image encoders evaluated under the same protocol						
PE_{core} (Bolya et al., 2025)	2B/14	38.9	61.1	69.2	0.590	4.119
$\text{PE}_{\text{spatial}}$ (Bolya et al., 2025)	2B/14	49.3	73.2	82.7	0.362	3.082
AM-RADIOv2.5 (Ranzinger et al., 2024)	1B/14	53.0	78.4	85.4	0.340	2.918
SigLIP 2 (Tschannen et al., 2025)	1B/16	42.7	64.8	72.7	0.494	3.273
DINOv2 (Oquab et al., 2024)	1B/14	49.5	75.6	83.1	0.372	2.624
DINOv3 ViT-H+ (Siméoni et al., 2025)	0.8B/16	54.8	79.5	85.8	0.352	2.635
Image–video encoders evaluated under the same protocol						
InternVideo2-1B (Wang et al., 2024)	1B/14	28.4	35.6	65.2	0.471	4.739
V-JEPA 2 (Assran et al., 2025)	1B/16	24.4	45.9	63.9	0.642	4.650
V-JEPA 2.1 (Mur-Labadia et al., 2026)	1B/16	47.8	71.8	**84.7**	0.350	2.601
VANE (ours)	1B/16	**51.0**	**75.4**	84.5	**0.329**	**2.431**

4.3 Dense Perception

To assess spatial detail beyond global recognition, we evaluate semantic segmentation on ADE20K, Cityscapes, and VOC (Zhou et al., 2017; Cordts et al., 2016; Everingham et al., 2010) and monocular depth estimation on NYUv2 and KITTI (Silberman et al., 2012; Geiger et al., 2012). Following DINOv3 (Siméoni et al., 2025) and V-JEPA 2.1 (Mur-Labadia et al., 2026), we train a dense linear projection on frozen final-layer patch features for each task and report mIoU for segmentation and RMSE for depth. Compared with V-JEPA 2.1 (Table 4), VANE improves ADE20K and Cityscapes by 3.2 and 3.6 mIoU, respectively, while remaining comparable on VOC (84.5 vs. 84.7). It also reduces depth RMSE from 0.350 to 0.329 on NYUv2 and from 2.601 to 2.431 on KITTI, the lowest errors among all encoders in Table 4, including image encoders. Since these benchmarks consist of still images, we attribute the gains mainly to next-crop prediction and the generative decoder, both of which improve ADE20K in our ablations (Table 5b,c). Figure 4 further shows that VANE produces cleaner and more coherent segments than V-JEPA 2.1.

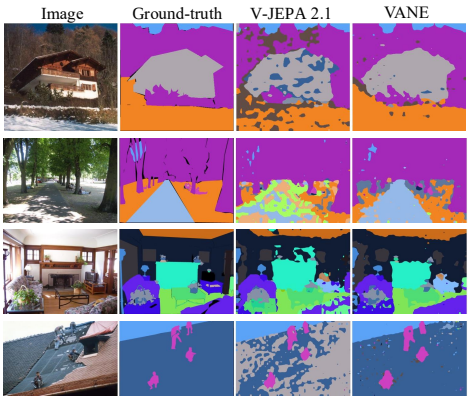


Figure 4: **Qualitative comparison on ADE20K.** V-JEPA 2.1 and VANE with the same dense linear probe on frozen features.

4.4 Ablation Studies

We ablate the key design choices of VANE. Unless otherwise specified, all ablations use a ViT-L encoder trained for 50K iterations. Additional ablations on the predictor attention and the number of image crops, as well as a matched comparison with V-JEPA 2.1, are provided in Appendix E.

Prediction space and granularity. Replacing pixel targets with frozen DINOv2 ViT-L embeddings improves SSv2 by 16.2 points and ADE20K by 21.0 mIoU (Table 5a). Among alternative targets, DINOv2 outperforms SigLIP 2 and VA-VAE (Yao et al., 2025) by 6.8 and 4.3 points on SSv2, and the online EMA target by 4.8 points, suggesting that effective targets should capture the content and local structure of a scene while abstracting over appearance variation. We adopt DINOv2 as the default tokenizer because it performs best on IN1K and SSv2, whereas DINOv3 is only slightly better on ADE20K (45.2 vs. 45.1). To isolate the effect of prediction granularity, we also compare next-frame prediction with standard next-token prediction, which predicts each visual token from the preceding tokens (Appendix E.4). With the target space fixed, next-frame prediction improves IN1K, SSv2, and ADE20K by 3.7, 19.1, and 22.7 points, respectively, showing that the gains stem from the prediction objective rather than from the target encoder itself.

Table 5: **Ablation studies.** Top-1 accuracy on IN1K/SSv2 and mIoU on ADE20K with a ViT-L encoder trained for 50K iterations. (a) Prediction space and granularity; next-frame and next-token variants share the frozen DINOv2 ViT-L target space. (b) Image objectives, with the video objective, tokenizer, and generative decoder fixed. (c) Auxiliary generative decoder.

(a) Prediction space and granularity

Target	Prediction	IN1K	SSv2	ADE20K
Reference				
Online EMA	Next frame	81.0	67.5	35.3
Pixels	Next frame	75.8	56.1	24.1
Semantic tokenizers				
VA-VAE	Next frame	81.8	68.0	37.1
SigLIP 2	Next frame	77.5	65.5	29.6
DINOv2 ViT-L	Next frame	**84.5**	**72.3**	45.1
DINOv2 ViT-L	Next token	80.8	53.2	22.4
DINOv3 ViT-L	Next frame	84.0	71.8	**45.2**

(b) Image objectives

Image objective	IN1K	SSv2	ADE20K
Masked image	81.0	69.3	38.3
Next cell	83.7	71.6	43.1
Next crop (default)	84.5	72.3	45.1

(c) Generative decoder

Setting	IN1K	SSv2	ADE20K
w/o Decoder	84.3	72.0	42.8
w/ Decoder	**84.5**	**72.3**	**45.1**

Unified image prediction. Table 5b compares three image objectives under the same video objective: masked-image completion, next-cell prediction, and local-to-global next-crop prediction. For next-cell prediction, we uniformly partition an image into a 2×2 grid, treat the four cells as an ordered sequence, and predict each cell from the preceding ones; next-crop prediction instead orders views from local to global and supervises only newly revealed regions. Next-crop prediction performs best on all three benchmarks, improving ADE20K by 2.0 and 6.8 mIoU over next-cell prediction and masked-image completion, respectively. Both ordered objectives outperform masked completion, indicating that images benefit from the same predictive structure as videos, and the further gain of next-crop prediction suggests that progressively expanding context is preferable to a fixed grid.

Generative supervision for spatial detail. We test whether the auxiliary generative objective helps preserve spatial information beyond semantic prediction. Adding the generative decoder raises ADE20K from 42.8 to 45.1 mIoU (Table 5c), with smaller gains of 0.2 points on IN1K and 0.3 on SSv2. The larger gain on dense perception suggests reconstructing visual content in pixel space complements semantic prediction by preserving fine-grained spatial structure.

Scaling and the training recipe. Figure 5 summarizes the cumulative effects of model scaling, data expansion, and post-training. Scaling the encoder from 300M to 1B parameters improves both SSv2 and ADE20K, while data expansion brings further gains. Long-video post-training yields the clearest improvement on SSv2 (73.8 to 75.6), suggesting that longer temporal context strengthens next-frame prediction by exposing the model to richer visual evolution. The final 384×384 cooldown further improves SSv2 and ADE20K. Overall, the complete recipe improves SSv2 and ADE20K by 3.6 and 5.9 points, respectively, highlighting complementary benefits from model and data scaling, longer temporal context, and higher spatial resolution.

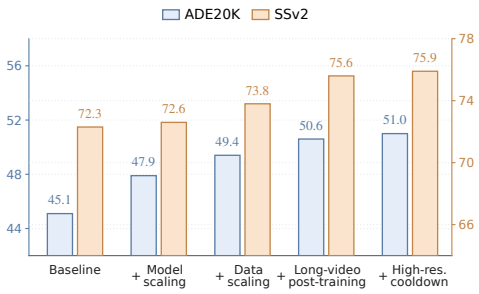


Figure 5: **Cumulative training recipe.**

5 Conclusion

We presented VANE, a self-supervised visual pretraining framework that predicts the next frame in embedding space rather than recovering missing content within an observed context, making temporal evolution the primary learning signal. Extending this objective to images through local-to-global next-crop prediction enables unified image–video pretraining. Across anticipation, global recognition, and dense perception, VANE achieves state-of-the-art performance among unified encoders on most benchmarks and improves consistently with model and data scale, supporting next-frame prediction in embedding space as a scalable principle for visual world modeling that captures both what the visual world contains and how it evolves.

AI USE STATEMENT

Generative AI tools were used solely to assist with information search, language editing, and polishing of the manuscript. All AI-assisted content was reviewed and revised by the authors, who take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We provide the details required to reproduce our method and experiments throughout the main paper and appendix. The main paper specifies the learning objectives, model components, training procedure, and downstream evaluation protocols. The appendix provides additional details on dataset construction and sampling, model architecture, video context masking, stage-wise training configurations and hyperparameters, and evaluation settings. For standard downstream benchmarks, we follow the established protocols cited in the corresponding experimental sections.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023.

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

Federico Baldassarre, Marc Szafraniec, Basile Terver, Vasil Khalidov, Francisco Massa, Yann LeCun, Patrick Labatut, Maximilian Seitzer, and Piotr Bojanowski. Back to the features: DINO as a foundation for video world models. *arXiv preprint arXiv:2507.19468*, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *TMLR*, 2024.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, et al. Perception encoder: The best visual embeddings are not at the output of the network. In *NeurIPS*, 2025.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, 2020.

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 *CVPR*, 2016.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *IJCV*, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Angel Bautista, Vaishaal Shankar, Alexander Toshev, Joshua M. Susskind, and Armand Joulin. Scalable pre-training of large autoregressive image models. In *ICML*, 2024.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV*, 2010.

Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. In *NeurIPS*, 2022.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *ICCV Workshops*, 2019.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

Efstathios Karypidis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. DINO-Foresight: Looking into the future with DINO. In *NeurIPS*, 2025.

Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023.

Yann LeCun. A path towards autonomous machine intelligence. OpenReview, 2022.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. UniFormerV2: Unlocking the potential of image ViTs for video understanding. In *ICCV*, 2023a.

Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *ICCV*, 2023b.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *NeurIPS*, 2024.

Xianhang Li, Chen Huang, Chun-Liang Li, Eran Malach, Josh Susskind, Vimal Thilak, and Etai Littwin. Rethinking JEPA: Compute-efficient video self-supervised learning with frozen teachers. In *ICLR*, 2026.

Yingwei Li, Yi Li, and Nuno Vasconcelos. RESOUND: Towards action recognition without representation bias. In *ECCV*, 2018.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019.

Himangi Mittal, Nakul Agarwal, Shao-Yuan Lo, and Kwonjoon Lee. Can't make an omelette without breaking some eggs: Plausible action anticipation using large video-language models. In *CVPR*, 2024.

Lorenzo Mur-Labadia, Ruben Martinez-Cantin, Josechu Guerrero, Giovanni Maria Farinella, and Antonino Furnari. AFF-ttention! affordances and attention models for short-term object interaction anticipation. In *ECCV*, 2024.

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.

Heejeong Nam, Quentin Le Lidec, Lucas Maes, Yann LeCun, and Randall Balestriero. Causal-JEPA: Learning world models through object-level latent masking. In *ICML*, 2026.

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. *TMLR*, 2024.

Francesco Ragusa, Giovanni Maria Farinella, and Antonino Furnari. StillFast: An end-to-end approach for short-term object interaction anticipation. In *CVPR Workshops*, 2023.

Jathushan Rajasegaran, Ilija Radosavovic, Rahul Ravishankar, Yossi Gandelsman, Christoph Feichtenhofer, and Jitendra Malik. An empirical study of autoregressive pre-training from videos. In *ICCV*, 2025.

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. AM-RADIO: Agglomerative vision foundation model – reduce all domains into one. In *CVPR*, 2024.

Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 1999.

Sucheng Ren, Hongru Zhu, Chen Wei, Yijiang Li, Alan Yuille, and Cihang Xie. ARVideo: Autoregressive pretraining for self-supervised video representation learning. *TMLR*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Debaditya Roy, Ramanathan Rajendiran, and Basura Fernando. Interaction region visual transformer for egocentric action anticipation. In *WACV*, 2024.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 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.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *NeurIPS*, 2024.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022.

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.

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

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *CVPR*, 2016.

Chenting Wang, Yuhan Zhu, Yicheng Xu, Jiange Yang, Lang Lin, Ziang Yan, Yali Wang, Yi Wang, and Limin Wang. InternVideo-Next: Towards world-understanding video models. In *CVPR*, 2026.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE V2: Scaling video masked autoencoders with dual masking. In *CVPR*, 2023a.

Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *CVPR*, 2023b.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. InternVideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, et al. InternVideo2: Scaling foundation models for multimodal video understanding. In *ECCV*, 2024.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. InternVideo2.5: Empowering video MLLMs with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025.

Sihan Xu, Ziqiao Ma, Wenhao Chai, Xuweiyi Chen, Weiyang Jin, Joyce Chai, Saining Xie, and Stella X. Yu. Next-embedding prediction makes strong vision learners. *arXiv preprint arXiv:2512.16922*, 2025.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *CVPR*, 2025.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *EMNLP (System Demonstrations)*, 2023.

Long Zhao, Nitesh B. Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J. Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, et al. VideoPrism: A foundational visual encoder for video understanding. In *ICML*, 2024.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, 2017.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *ICML*, 2025.

APPENDIX

A MODEL ARCHITECTURE

VANE consists of a shared causal visual encoder, a context-isolated state predictor, a frozen semantic tokenizer, and a lightweight generative decoder. The encoder is the only component retained for downstream evaluation; the state predictor, semantic tokenizer, and generative decoder are used only during pretraining.

Table 6 summarizes the model configurations. Unless otherwise specified, the predictor uses the same depth and hidden dimension across encoder scales.

Causal encoder. The encoder processes all observations in a sequence in a single forward pass. Under frame-causal attention, tokens in observation t may attend to tokens from observations $1, \ldots, t$, but never to future observations. Tokens within the same observation are processed jointly. For videos, the input is patchified into $2 \times 16 \times 16$ temporal-spatial tubelets ($T \times H \times W$), such that each temporal encoder position spans two consecutive sampled raw frames. For images, each local-to-global crop constitutes one observation. Images and videos use separate patch embeddings and a learned modality embedding, while all other components are shared; image sequences use no context masking. Fig. 6 (left) illustrates the attention mask.

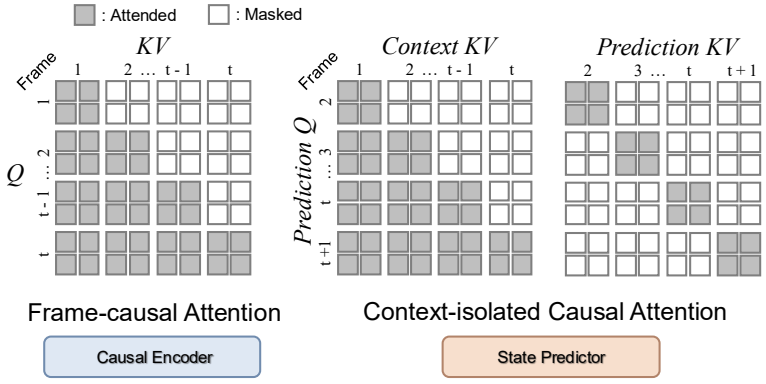


Figure 6: **Attention masks.** *Left:* frame-causal attention in the causal encoder, bidirectional within each frame and unidirectional across frames. *Right:* context-isolated causal attention in the state predictor. Prediction tokens for frame $t+1$ read the context states of frames up to t, which supply only keys and values, and attend to each other only within the same target frame.

Context-isolated state predictor. The state predictor receives encoded historical states as read-only context. Since the encoder and predictor use different hidden dimensions, encoder states are first projected from the encoder width to the 384-dimensional predictor space. Historical context states then provide keys and values but are never updated inside the predictor. Prediction tokens are obtained by duplicating a learned token and assigning the corresponding target patch positions.

Prediction tokens for observation $t+1$ attend to the context states of observations $1, \ldots, t$ and to one another, while predictions belonging to different future observations remain isolated (Fig. 6, right). The resulting predictor states are projected back to the encoder width before being passed to the semantic and generative objectives. This design separates representation learning from future-state readout: the predictor can only use information already encoded by the causal backbone.

Semantic tokenizer. We use a frozen DINOv2 ViT-L as the semantic tokenizer for the main model. We take its final patch tokens after removing special tokens and normalize each patch feature with a LayerNorm without affine parameters; no additional ℓ_2 normalization is applied.

To align tokenizer targets with the encoder patch grid, we resize the tokenizer input so that its 14×14 patchification produces the same $H_p \times W_p$ spatial grid, without interpolating tokenizer features. For example, a 256×256 encoder input produces a 16×16 grid, while the corresponding tokenizer input is resized to 224×224.

Table 6: **Model architectures.** The main model uses a ViT-g causal encoder, a 12-layer context-isolated state predictor, a frozen DINOv2 ViT-L semantic tokenizer, and a six-layer patch-isolated DiffMLP decoder.

	ViT-L	**ViT-g**
Causal Encoder		
Block	ViT Block	
Depth	24	40
Width	1024	1408
Num. Heads	16	22
MLP Ratio	4.0	4.36
Parameters	300M	1B
Patch Size	16×16	
Video Tubelet Size	2 frames	
Normalization	LayerNorm	
Position Encoding	3D RoPE	
Attention Mask	Frame-causal	
State Predictor		
Block	Context-isolated attention block	
Depth	12	
Width	384	
Num. Heads	12	
Input Projection	Encoder width $\rightarrow 384$	
Output Projection	$384 \rightarrow$ Encoder width	
Normalization	LayerNorm	
Position Encoding	3D RoPE	
Prediction-token Init.	Learned token duplicated at target positions	
Dropout / DropPath	0.0 / 0.0	
Semantic Tokenizer		
Architecture	DINOv2 ViT-L	
Patch Size	14×14	
Optimization	Frozen throughout pretraining	
Feature	Final patch-token representation	
Per-frame Feature Dim.	1024	
Target Packing	Video: two-frame concatenation; Image: duplication	
Target Dimension	2048	
Semantic Projection	$1024 \rightarrow 2048$	$1408 \rightarrow 2048$
Patch-Isolated Generative Decoder		
Block	DiffMLP	
Depth	6	
Hidden Dimension	1536	
Image Pixel Packing	Duplicate along temporal dimension	
Video Pixel Packing	Two-frame tubelet	
Noisy Pixel Dimension	1536	
Condition Injection	Concatenation	
Input Projection	$2560 \rightarrow 1536$	$2944 \rightarrow 1536$
Time Embedding	$256 \rightarrow 1536 \rightarrow 1536$	
Activation	SiLU	
Conditioning	Predicted state $s_{t,j}$ and flow time τ	
Noisy Input	$\tilde{x}_{t,j}(\tau) \in \mathbb{R}^{1536}$	
Output	1536-d flow velocity $\hat{v}_{t,j}$	
Patch Interaction	None (patch-isolated)	

For images, the tokenizer encodes the full image once, and crop targets are indexed from the aligned full-image grid at their corresponding patch positions. For videos, it encodes individual frames independently. Each normalized patch feature is 1024-dimensional; features at the same spatial position in a two-frame video tubelet are concatenated into a 2048-dimensional target, while image features are duplicated to the same dimension. Both modalities therefore share the same semantic projection head, which maps the predictor output to the 2048-dimensional tokenizer target space.

Patch-isolated generative decoder. The auxiliary generative decoder is implemented as a six-layer DiffMLP with hidden dimension 1536. To share the same decoder across modalities, each image is duplicated along the temporal dimension before pixel patchification, forming a two-frame input with identical image content. Consequently, both image patches and video tubelets are represented by $2 \times 3 \times 16 \times 16 = 1536$ pixel values.

For the main ViT-g model, the 1408-dimensional predicted state is concatenated with the 1536-dimensional noisy pixel token and projected from 2944 to the 1536-dimensional decoder space. Flow time is encoded using a 256-dimensional sinusoidal embedding followed by a $256 \to 1536 \to 1536$ SiLU MLP. Each DiffMLP block contains LayerNorm, an MLP, SiLU activations, and adaptive shift, scale, and gate modulation. The final layer predicts a 1536-dimensional flow velocity. Target patches are decoded independently, without access to neighboring target patches or encoder context other than through the predicted state.

B Pretraining Data

We jointly pretrain VANE on large-scale image and video corpora. Our corpus contains approximately 145M images and 24M video samples, matching V-JEPA 2.1 (Mur-Labadia et al., 2026) in overall image–video sample count. The data are constructed independently, following the large-scale curation principles of DINOv2 (Oquab et al., 2024) and V-JEPA 2.1 (Mur-Labadia et al., 2026), and combine public datasets with additional retrieved and class-balanced data. For all public datasets, only their training splits are used for pretraining and retrieval-based curation; validation and test splits are excluded throughout data construction.

We refer to the resulting unified visual corpus as VISTA (**V**isual **I**mage–video **S**calable **T**raining **A**rchive), with VISTA-I and VISTA-V denoting its image and video subsets, respectively. Each subset combines public seed datasets with additional samples constructed through distribution-matched retrieval and class-balanced curation. The public datasets therefore serve both as high-quality training data and, where applicable, as reference distributions for expanding the corpus.

B.1 Multi-Domain Data Curation

Distribution-matched retrieval. Following DINOv2 (Oquab et al., 2024) and V-JEPA 2.1 (Mur-Labadia et al., 2026), we use the training splits of multiple curated public datasets to define complementary visual distributions and retrieve visually matched samples from a substantially larger independently collected candidate pool.

For images, the retrieval references span broad recognition, fine-grained recognition, textures, scenes, and object-centric domains, including IN21K, IN1K, Caltech-101, CUB-200-2011, DTD, FGVC-Aircraft, Flowers-102, Food-101, Oxford-IIIT Pet, Stanford Cars, SUN397, PASCAL VOC 2007, and WebVision. Additional public image sources, including SA-1B, Google Landmarks v2, and MET, are incorporated directly into VISTA-I. Together, these sources provide broad coverage of objects, scenes, textures, fine-grained categories, and in-the-wild visual content.

For videos, we adopt the same retrieval reference datasets as V-JEPA 2.1 (Mur-Labadia et al., 2026), using the training splits of SSv2, K710, COIN, and EPIC-KITCHENS to cover motion-centric, action-centric, instructional, and egocentric domains, respectively. HowTo100M is additionally incorporated directly as large-scale instructional video data. The public video datasets and their retrieved counterparts are combined to form the distribution-matched component of VISTA-V.

Importantly, the public datasets determine the reference distributions for retrieval, while the retrieved samples themselves are independently collected and are not drawn from the corresponding public datasets. The public seed datasets are retained as part of VISTA-I/V rather than being replaced by their retrieved counterparts.

Class-balanced curation. Distribution-matched retrieval preserves the visual characteristics of the reference datasets but can inherit their long-tailed semantic distributions. We therefore further augment both VISTA-I and VISTA-V with class-balanced data, following the data balancing principle of DINOv3 (Siméoni et al., 2025). This complementary component broadens semantic coverage and reduces the dominance of frequent visual concepts in the retrieved data.

Table 7: **Pretraining data mixture.** VISTA-I and VISTA-V denote the image and video subsets of our unified VISTA corpus. Each combines public seed datasets, distribution-matched retrieval, and class-balanced data as described in Sec. B.1.

Regime	Modality	Training data
Homogeneous (10%)	Image	IN1K
	Video	SSv2
Heterogeneous (90%)	Image	VISTA-I
	Video	VISTA-V

Overall, each VISTA subset combines three complementary components: public seed datasets, distribution-matched retrieved data, and class-balanced data. Applying the same curation principle to both modalities yields a unified image–video corpus for large-scale visual pretraining.

B.2 Homogeneous and Heterogeneous Batch Sampling

We combine homogeneous and heterogeneous sampling during pretraining, as summarized in Table 7. For 10% of optimization steps, we construct homogeneous batches, sampling images exclusively from IN1K and videos from SSv2. These batches provide a stable learning signal from canonical image and video recognition datasets.

For the remaining 90% of steps, image and video batches are sampled from VISTA-I and VISTA-V, respectively. For video sampling, we organize VISTA-V into four source groups with sampling probabilities $0.20/0.20/0.25/0.35$, corresponding to motion-centric data, action-centric data, instructional and retrieved video data, and class-balanced video data, respectively. These probabilities are realized through proportional allocation of data-parallel ranks across the four source groups, with each rank sampling from its assigned group.

Image and video batches are processed in separate forward and backward passes, and their gradients are accumulated before each parameter update. This allows the two modalities to share the same encoder and predictive objective while retaining modality-specific data sampling.

C Multi-Stage Training Strategy

Our training recipe contains four consecutive stages (Table 8): (1) learning-rate warm-up, (2) large-scale base pretraining, (3) long-video post-training, and (4) high-resolution cooldown. The first two stages learn the main representation using 16-frame clips at 256×256 resolution, while the last two stages progressively increase temporal and spatial context.

C.1 Input Sampling and Augmentation

Video sampling. Videos are sampled at a target rate of 4 FPS. Base pretraining uses 16-frame clips and long-video post-training increases the observation window to 64 frames. The encoder uses a temporal tubelet size of two, so each temporal patch corresponds to two consecutive sampled frames. For videos longer than the required observation window, a clip window is randomly selected during training.

Image and video augmentation. At 256×256 resolution, both image and video inputs use random resized cropping with an area range of $[0.3, 1.0]$ and aspect-ratio range of $[0.75, 1.35]$, followed by horizontal flipping with probability 0.5. We do not use AutoAugment, random erasing, or additional color jitter in this training configuration. The same spatial augmentation principle is retained when increasing the input resolution during high-resolution cooldown.

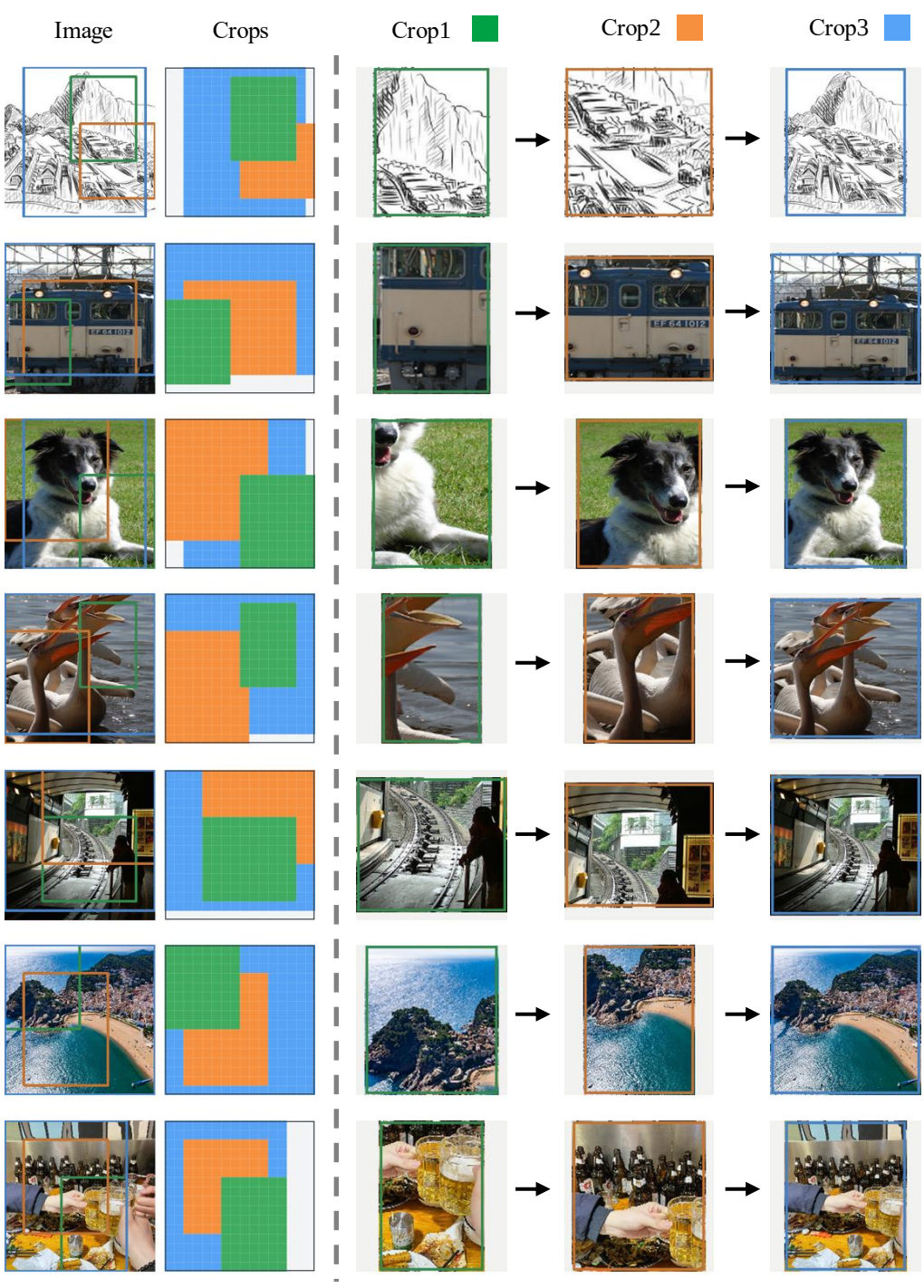


Figure 7: **Visualization of local-to-global next-crop prediction.** Each row shows the spatial layout of three sampled crops and the corresponding prediction sequence $\mathrm{Crop}_1 \rightarrow \mathrm{Crop}_2 \rightarrow \mathrm{Crop}_3$. The first two observations are local crops and the last provides broader scene context. At each step, supervision is restricted to spatial regions newly revealed by the next crop. Crops are resized to a common display size for visualization only; their original spatial coordinates are used during training.

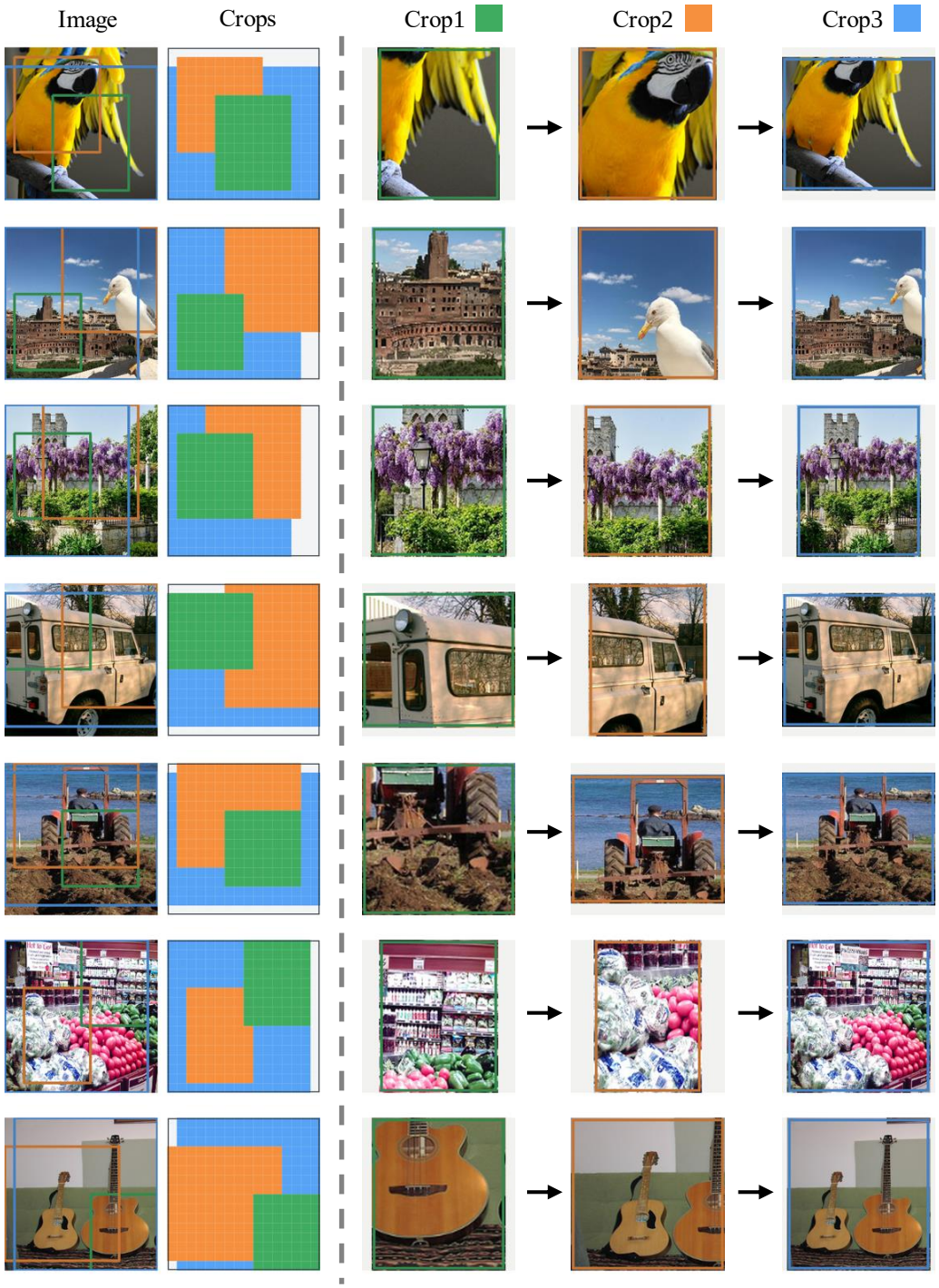


Figure 8: **Additional examples of local-to-global next-crop prediction.** We show additional sampled crop sequences following the same local-to-global sampling strategy as in Fig. 7. Supervision is applied only to regions newly revealed by each subsequent crop. Crops are resized to a common display size for visualization only; their original spatial coordinates are used during training.

C.2 Local-to-Global Image Sampling

We first apply the image augmentation described in Appendix C.1 to obtain a base image. Local-to-global views are then sampled from its patch grid without resizing the individual crops. All view coordinates and semantic targets are defined relative to this shared base image.

For image pretraining, we construct an ordered sequence of $K = 3$ local-to-global crops. Two local crops are sampled with area ratios in $[0.2, 0.6]$, followed by a global crop sampled with an area ratio in $[0.8, 1.0]$. All crops use aspect ratios in $[0.75, 1.35]$ and are sampled on the image patch grid.

The two local crops are ordered by area. Successive crops cover increasingly broad and complementary scene regions, and may partially overlap or extend to adjacent regions. To encourage complementary observations, local-crop sampling favors pairs whose intersection occupies at most 0.4 of the larger local crop. The final global crop provides substantially broader scene context but is not required to coincide exactly with the full uncropped image. Semantic and generative supervision are applied only to patch positions newly revealed relative to preceding observations.

Figure 7 and Figure 8 visualize representative local-to-global crop sequences and their spatial relationships in the original image coordinate system.

C.3 Video Context Masking

To increase the difficulty of next-frame prediction, we apply random multi-block tube masking to the historical video context while keeping future targets fully supervised, following V-JEPA 2.1 (Mur-Labadia et al., 2026).

We follow the two multi-block mask configurations used in V-JEPA 2.1. One samples eight spatial blocks with scale 0.15 while retaining at least 5% of tokens; the other samples two larger blocks with spatial scale 0.70. For both configurations, the temporal scale is 1.0 and the block aspect ratio is sampled from $[0.75, 1.5]$. Block positions are sampled randomly and may overlap, so the resulting fraction of masked context tokens varies across samples rather than being defined by a single fixed mask ratio. The spatial masking pattern is extended temporally to form tube masks across the observed clip.

C.4 Warm-Up

Training starts from randomly initialized VANE parameters while the semantic tokenizer remains frozen. During the first 12K updates, the learning rate is linearly increased from 1×10^{-4} to 6×10^{-4}. We use 16-frame video clips and 256×256 inputs, with global batch sizes of 1536 videos and 4608 images. The warm-up stabilizes optimization before large-scale joint image–video pretraining.

C.5 Base Pretraining

After warm-up, we continue training for 180K updates using the same 16-frame, 256×256 input configuration and a constant learning rate of 6×10^{-4}. This stage constitutes the main representation-learning phase and uses the complete data mixture described in Appendix B.

Image and video batches are processed independently in separate forward/backward passes. Their gradients are accumulated before one optimizer step, allowing the shared encoder and predictor to learn from both modalities without requiring image and video samples to be packed into the same sequence.

C.6 Long-Video Post-Training

We next increase the temporal context from 16 to 64 frames while retaining the 256×256 spatial resolution. This stage runs for 60K additional updates with a global video batch size of 768, providing a longer observation window for predictive pretraining.

C.7 High-Resolution Cooldown

The final stage increases the spatial input resolution from 256×256 to 384×384 for 12K updates while retaining the 64-frame observation window introduced during long-video post-training. It starts from the long-video checkpoint and linearly decays the learning rate during this stage.

The warm-up and base-pretraining stages together require approximately 26.6K B200 GPU-hours, while the two post-training stages require approximately 30.7K B200 GPU-hours in total.

Table 8: **Multi-stage training strategy.** The base model is first trained with 16-frame clips at 256×256, followed by long-video post-training with 64-frame clips and a final high-resolution stage.

	Warm-up	**Base Pretrain**	**Long-Video Post-train**	**High-Res. Cooldown**
Input Configuration				
Video Frames	16	16	64	64
Video Sampling Rate		4 FPS		
Video Tubelet Size		2 frames		
Resolution	256^2	256^2	256^2	384^2
Image Views K	3	3	3	3
Video Context Mask		Random multi-block tube masking		
Optimization				
Updates	12K	180K	60K	12K
Global Video Batch	1536	1536	768	768
Global Image Batch	4608	4608	4608	4608
Learning Rate	$1 \times 10^{-4} \rightarrow 6 \times 10^{-4}$	6×10^{-4}	6×10^{-4}	$6 \times 10^{-4} \rightarrow 6 \times 10^{-5}$
Optimizer		AdamW		
AdamW β		$(0.9, 0.95)$		
AdamW ϵ		10^{-8}		
Weight Decay		0.04		
Gradient Clipping		1.0		
Precision		BF16		
Objectives				
Semantic Loss Weight	1.0	1.0	1.0	1.0
λ_{flow}	0.25	0.25	0.25	0.25
Semantic Tokenizer		Frozen DINOv2 ViT-L		

D Additional Results

D.1 Qualitative Short-Term Anticipation

Figure 9 shows qualitative results on Ego4D STA v2, obtained with the same frozen-feature probe as in Table 2. Each prediction consists of a bounding box for the next-active object and a label combining the interaction verb, the object noun, and the time to contact (TTC). In every example shown, one of the top-scoring predictions matches the ground-truth verb–noun pair and localizes the correct object with a comparable time to contact. The remaining candidates are typically plausible alternatives in the same scene, such as holding rather than marking the wood, or adjusting the garment rather than taking the scissors.

D.2 Qualitative Dense Perception

We provide additional qualitative results for semantic segmentation and monocular depth estimation. All predictions are produced using the same frozen-feature evaluation protocol as in the quantitative experiments in the main paper. These examples complement aggregate metrics by showing how well pretrained features preserve object boundaries, fine-grained scene structure, and geometric layout.

Semantic segmentation. Figures 10 and 11 show representative examples from ADE20K and Cityscapes. We visualize the input image, ground-truth annotation, and predictions obtained from identical dense readouts trained on frozen visual features.

Monocular depth estimation. Figures 12 and 13 provide corresponding qualitative results on NYUv2 and KITTI. NYUv2 emphasizes indoor geometric structure, whereas KITTI provides long-range outdoor scenes and substantially different depth statistics.

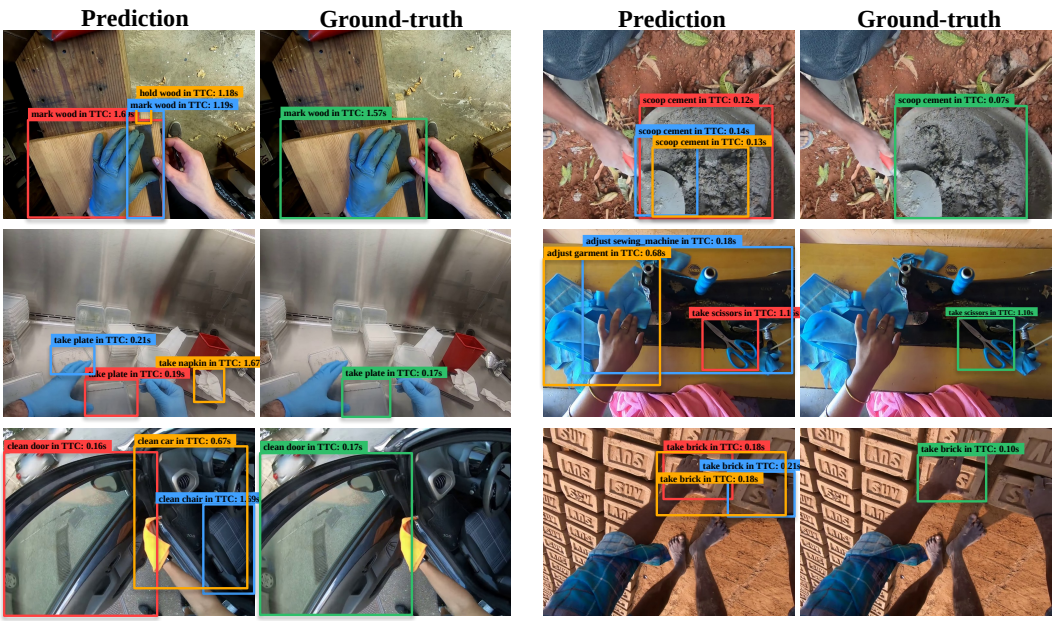


Figure 9: **Qualitative short-term anticipation results on Ego4D STA v2.** Each example pairs the top-scoring predictions of VANE (left) with the ground truth (right). Each box marks a next-active object and is labeled with the interaction verb, the object noun, and the time to contact (TTC). Predictions are produced by a probe trained on frozen VANE features.

Table 9: **Masked versus next-frame prediction under matched pretraining.** All models use ViT-L encoders trained on VISTA for 50K iterations. The frozen-target V-JEPA 2.1 variant uses the same DINOv2 ViT-L semantic tokenizer as VANE. We report top-1 accuracy (%) on IN1K and SSv2 and mIoU (%) on ADE20K.

Method	Objective	Target	IN1K	SSv2	ADE20K
V-JEPA 2.1	Masked	EMA	81.8	72.0	38.4
V-JEPA 2.1	Masked	Frozen DINOv2	82.7	70.6	41.5
VANE	Next frame	Frozen DINOv2	**84.5**	**72.3**	**45.1**

E Additional Ablation Studies

The main paper studies prediction targets, image prediction strategies, generative supervision, scaling, and the training schedule. Here we first provide a controlled comparison with V-JEPA 2.1 under matched pretraining conditions, followed by further analyses of image crops, state-predictor attention, and prediction granularity. Unless otherwise specified, all ablations use a ViT-L encoder trained for 50K iterations.

E.1 Masked versus Next-Frame Prediction under Matched Pretraining

We perform a controlled comparison with V-JEPA 2.1 (Mur-Labadia et al., 2026) to examine whether the gains of VANE persist after accounting for differences in pretraining data and prediction targets. All models use ViT-L encoders trained on VISTA (Appendix B) for 50K iterations under the same training budget. V-JEPA 2.1 retains its bidirectional masked feature-prediction formulation, whereas VANE uses next-frame prediction. We evaluate frozen encoder features using the same downstream protocols as in the main paper, and report top-1 accuracy on IN1K and SSv2 and mIoU on ADE20K.

Matching the pretraining data. We first compare VANE with the standard V-JEPA 2.1 formulation while matching the pretraining corpus, encoder scale, and iteration budget. As shown in Table 9, VANE improves IN1K accuracy by 2.7 points and ADE20K by 6.7 mIoU, while achieving slightly

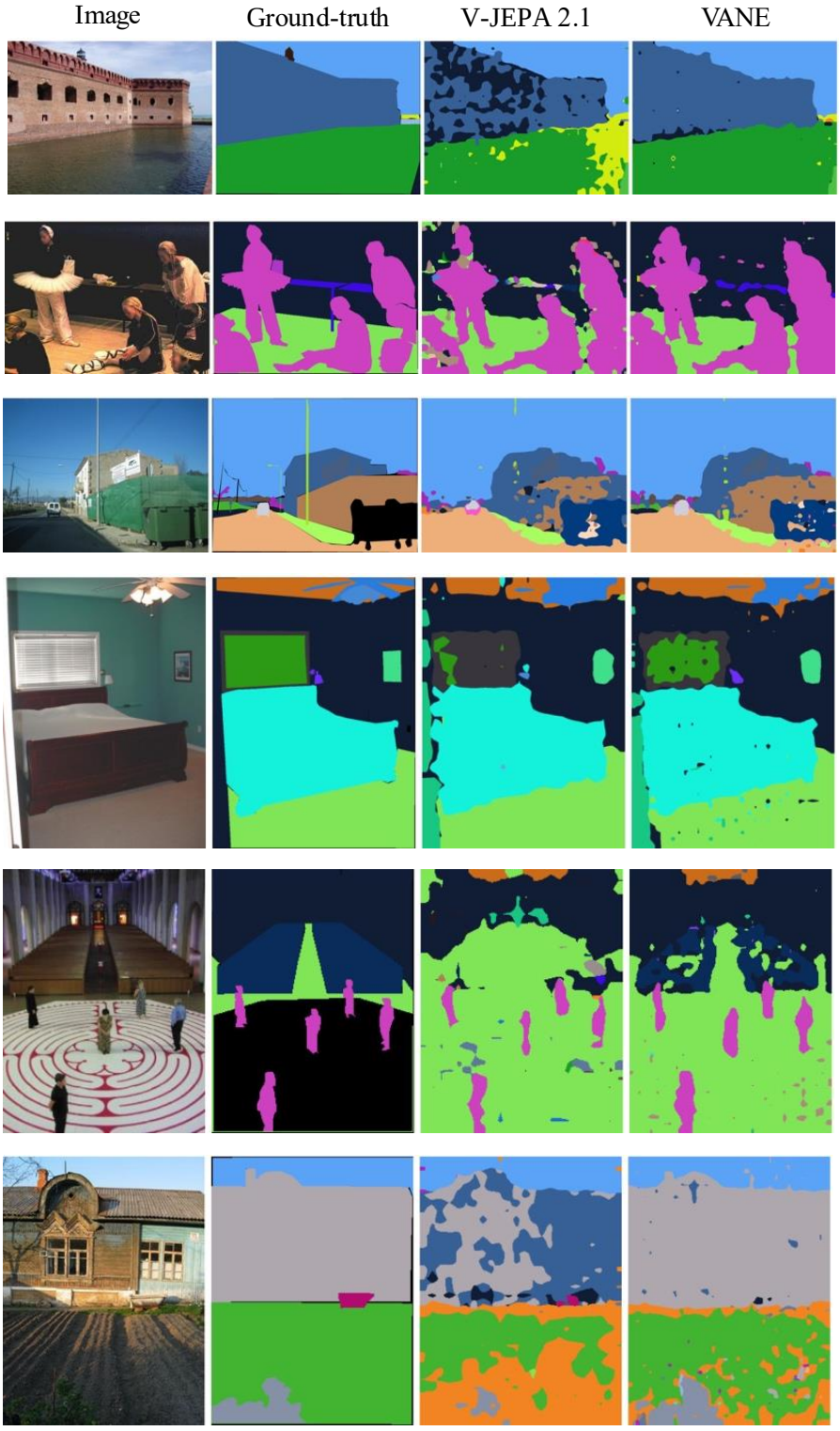


Figure 10: **Qualitative semantic segmentation results on ADE20K.** Each example shows the input image, ground truth, and predictions from dense probes trained on frozen features. All compared encoders use the same downstream protocol.

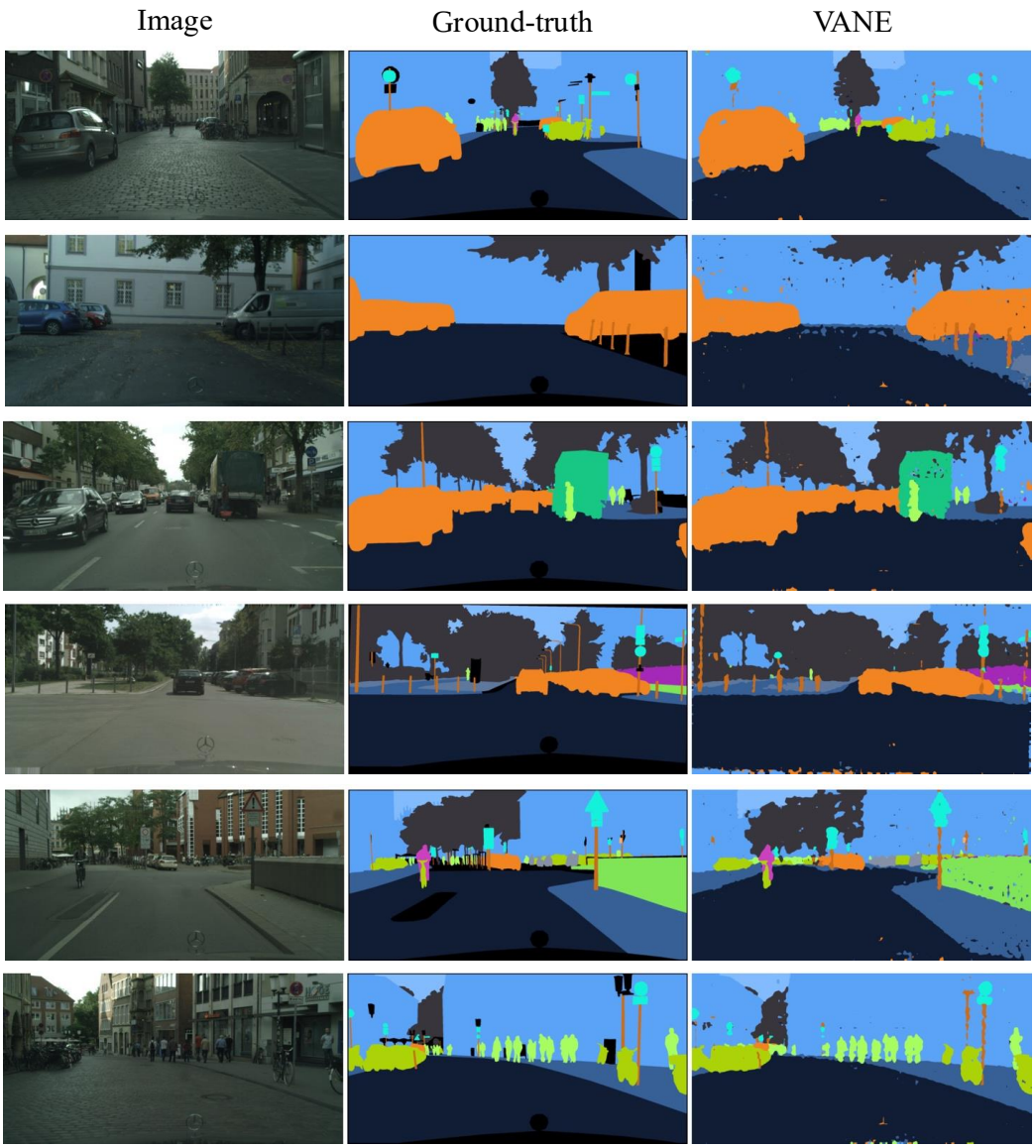


Figure 11: **Qualitative semantic segmentation results on Cityscapes.** Each example shows the input image, ground truth, and the prediction from a dense linear probe trained on frozen VANE features.

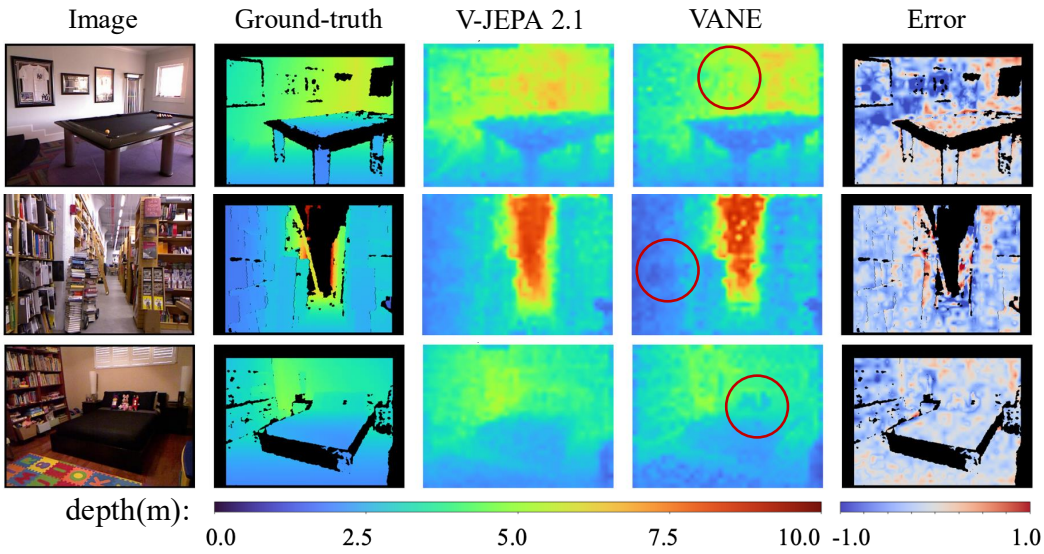


Figure 12: **Qualitative monocular depth estimation on NYUv2.** The error column shows $\Delta e = \left|\hat{d}_{VANE} - d_{\text{gt}}\right| - \left|\hat{d}_{\text{V-JEPA 2.1}} - d_{\text{gt}}\right|$; blue favors VANE, while red favors V-JEPA 2.1.

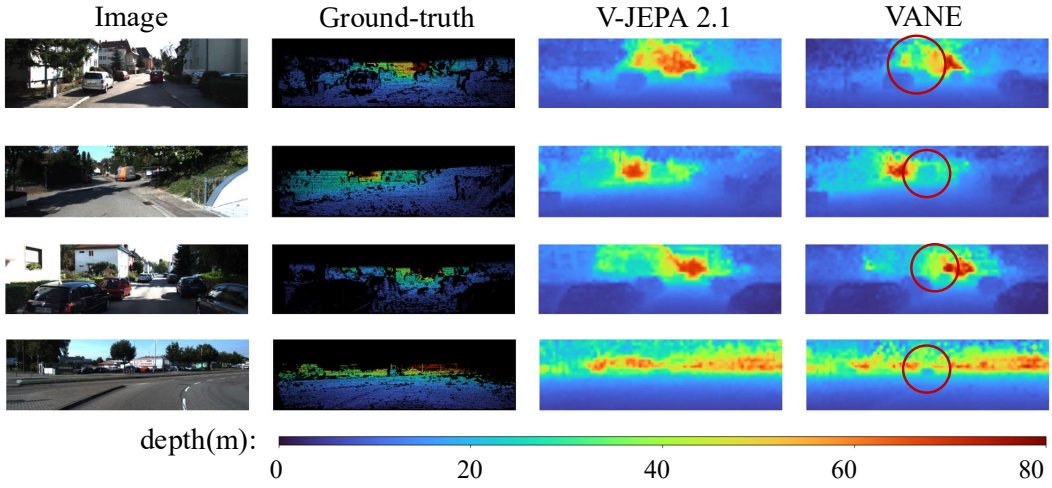


Figure 13: **Qualitative monocular depth estimation results on KITTI.** Predicted depth maps are obtained from dense readouts trained on frozen visual features using the same protocol as the quantitative evaluation.

Table 10: **Ablation of the number of image crops.** We vary the number of local-to-global views K while keeping the remaining training configuration fixed.

Number of Crops K	IN1K	SSv2	ADE20K
2	84.3	72.1	44.4
3 (default)	**84.5**	**72.3**	**45.1**
4	84.4	71.9	44.7

Table 11: **Ablation of state-predictor attention.** Context-isolated prediction treats historical encoder states as read-only context rather than jointly updating them with prediction tokens inside the predictor.

Predictor Attention	IN1K	SSv2	ADE20K
Causal self-attention	81.2	66.8	35.5
Context-isolated cross-attention	84.5	72.3	45.1

higher SSv2 accuracy (72.3% vs. 72.0%). Thus, the performance differences persist when both methods are trained on the same data at the same model scale and training budget.

Matching the semantic prediction space. We further replace the EMA target encoder of V-JEPA 2.1 with the same frozen DINOv2 ViT-L (Oquab et al., 2024) tokenizer used by VANE. This variant retains V-JEPA 2.1's bidirectional masked feature-prediction objective while predicting targets in the same semantic embedding space as VANE.

The frozen DINOv2 target improves the masked baseline on IN1K and ADE20K, from 81.8 to 82.7 and from 38.4 to 41.5, respectively, while SSv2 decreases from 72.0 to 70.6. With both the pretraining corpus and semantic target matched, VANE retains gains of 1.8 points on IN1K, 1.7 points on SSv2, and 3.6 mIoU on ADE20K. Thus, neither the pretraining corpus nor the frozen DINOv2 target alone explains the performance gap between masked and next-frame pretraining.

E.2 Number of Image Crops

Our default local-to-global image objective uses $K = 3$ local-to-global crops. We additionally vary the number of crops to examine how the amount of intermediate visual context affects representation learning. For all settings, crops are ordered from local to global, and supervision is restricted to regions newly revealed by the next crop. The video objective, semantic tokenizer, model architecture, and training schedule are kept fixed.

As shown in Table 10, performance is relatively stable across the three settings, with $K = 3$ providing the best overall results. Using only two views provides less intermediate context during local-to-global prediction, while increasing to four views brings no further improvement. We therefore use $K = 3$ as the default throughout the main experiments.

E.3 State-Predictor Attention

We next study how the state predictor should interact with historical encoder states. Our default predictor uses *context-isolated cross-attention*, where encoded historical states are treated as read-only context and provide keys and values to the predictor. Prediction tokens are updated by attending to this historical context to produce the future visual state, while the encoded context itself is not jointly updated inside the predictor.

As a control, we replace context-isolated attention with causal self-attention that jointly updates historical states and prediction tokens. Both variants enforce the same causal boundary: the prediction for observation t depends only on observations preceding t, and prediction-token groups remain isolated across target observations. The predictor depth, hidden dimension, target positions, and training objective are otherwise unchanged.

Table 11 shows that context-isolated prediction consistently improves downstream representation quality. Compared with causal self-attention, it improves IN1K accuracy by 3.3 points and SSv2 accuracy by 5.5 points, with a larger gain of 9.6 mIoU on ADE20K.

These results favor a separation between history encoding and future-state prediction. Causal self-attention allows the predictor to further transform historical states before producing the target representation. Context isolation instead requires the encoder outputs to carry the historical information needed for prediction, leaving the predictor to read these states and construct the future representation. This division of roles improves frozen-feature transfer in our experiments, particularly for dense perception.

E.4 Prediction Granularity

We provide implementation details for the next-token baseline used in the prediction-granularity comparison in Table 5. We compare our next-frame objective with standard next-token prediction (NTP), while keeping the frozen DINOv2 ViT-L semantic target, encoder scale, training schedule, batch size, input resolution, and clip length fixed.

In standard NTP, visual patches are ordered first temporally and then spatially. When predicting the k-th token of frame t, the model observes all preceding frames and the first $k-1$ tokens of the current frame, but not later tokens. The encoder therefore follows token-causal attention, analogous to a conventional autoregressive Transformer.

Next-frame prediction instead treats the entire frame as the prediction unit. All patches within an observed frame are jointly available, while no patch from the target frame is observed. Accordingly, the encoder uses frame-causal attention, allowing tokens to interact within each observed frame and attend to preceding frames, but not future frames. The target frame is then predicted jointly from the preceding observations.

Both variants use the same semantic and generative objectives and differ in how the visual sequence is factorized for prediction: NTP predicts successive visual tokens from prior tokens, whereas next-frame prediction predicts the next visual observation from preceding observations.