ANCHORMAE: BRIDGING TEMPORAL CORRESPONDENCE AND TASK SEMANTICS THROUGH FRAME-LEVEL SEMANTIC TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Ego–exo video learning must recognize activities across changing viewpoints without losing the temporal detail needed to distinguish action phases. Masked cross-view reconstruction provides dense temporal supervision, while clip-level language alignment alone does not directly specify task supervision for individual deployable visual frames. Prior methods also combine language and temporal prediction; the question examined here is *how to couple* dense visual reconstruction with same-video frame-level task supervision without requiring language at inference in the unpaired ego–exo setting. We introduce **AnchorMAE**, a shared-encoder dual-path framework with an unconditioned visual pathway and a training-only task-conditioned pathway. Dual-stage Semantic Anchor Adapters condition the latter; task-aware clip alignment and detached, same-video frame-level transfer supervise the former while masked self-/cross-view prediction retains pre-semantic visual targets. Only the visual pathway is used at inference. On AE2, AnchorMAE reports 0.8687 Phase F1, 0.8264 frame mAP@10, and 0.9966 Kendall's τ. Evaluations on AE2 and Ego-Exo4D, reported ablations and semantic/temporal diagnostics examine whether the proposed coupling supports both phase-sensitive correspondence and activity discrimination, while revealing remaining rank-dependent retrieval trade-offs.

1 INTRODUCTION

Relating egocentric and exocentric (ego–exo) videos requires recognizing the same activity across substantial viewpoint changes while preserving the temporal detail that distinguishes its constituent action phases. These requirements are complementary but not interchangeable: frames from different activities may share similar appearances, whereas frames from the same activity may depict different temporal stages. A useful cross-view representation must therefore encode both *what activity is being performed* and *where an observation lies within that activity*. This tension is relevant to cross-view frame retrieval, instructional video understanding, and learning from human demonstrations. AE2 (Xue & Grauman, 2023) probes fine-grained temporal correspondence between activity-matched yet temporally unaligned ego–exo videos, while Ego-Exo4D (Grauman et al., 2024) enables broader evaluation of cross-view representation transfer.

Existing work develops complementary solutions. BYOV (Park et al., 2025) uses masked self- and cross-view prediction to learn frame-sensitive visual representations, without explicit task-text supervision of reconstruction. VRS (Luo et al., 2025) learns cross-view alignment via feature translation, while PRISM (Chee et al., 2026) combines language-supervised latent recomposition with frame-level self-prediction for temporal learning. These mechanisms differ from same-video semantic feature transfer alongside detached visual reconstruction. Language-guided masked video learning is also established: MILES (Ge et al., 2022) and TGM (Fan et al., 2024) combine masking and language differently. The remaining question is thus not whether reconstruction or language supervision is individually effective, or whether the two can be combined at all, but *how their supervision should be coupled to preserve fine-grained visual correspondence and expose task semantics without requiring text at inference*.

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.

Simply adding a pooled video–text alignment objective to masked reconstruction does not explicitly determine how semantic information should reach individual visual frame representations. Moreover, making the reconstruction target itself text-conditioned changes the prediction task, potentially confounding the effects of dense visual learning and semantic supervision. These observations motivate a specific design distinction: *keep reconstruction targets visual while transferring training-time task semantics at the frame level*. Crucially, semantic transfer should operate between corresponding representations *within each video*, rather than assuming frame-level synchronization between independently sampled ego and exo recordings. Whether this coupling helps activity discrimination while retaining temporal detail is assessed with within-framework ablations and diagnostics.

We address this question with **AnchorMAE**, a shared-encoder dual-path framework (Fig. 1). Its *deployable visual pathway* processes unconditioned frame features using a temporal encoder and learns from masked self-view and cross-view reconstruction against detached, pre-semantic visual targets. A separate *training-only semantic pathway* injects a task-level text anchor through pre- and post-encoder Semantic Anchor Adapters (SAA), reusing *the same temporal encoder weights* in a separate forward pass. Task-aware clip-level alignment relates ego and exo visual embeddings and aligns semantic clip embeddings with their task text. More importantly for coupling the two pathways, frame-level semantic transfer supervises each visual token using its detached semantic counterpart from the *same* video. Ego and exo clips are matched by activity, not required to be temporally synchronized. At deployment, the text encoder, semantic adapters, semantic pass, and reconstruction decoders are discarded; only the visual pathway produces the frame and clip representations used downstream.

We evaluate the confirmed visual-only checkpoint on AE2 and Ego-Exo4D using reported objective baselines and module ablations. The reported AE2 scores are 0.8687 Phase F1, 0.8264 frame mAP@10, 0.9247 progression R^2, and 0.9966 Kendall's τ. On Ego-Exo4D, AnchorMAE reports gains in HR@5/10 and recognition over the listed VRS measurements but trails it at HR@1; the comparisons are contextualized by differences in training supervision and backbone choices. Activity/view probes, controlled semantic and temporal distractors, and visual-only retrieval analyses examine whether task discrimination and temporal correspondence are both represented. The qualitative comparison in Fig. 2 illustrates the distinct retrieval errors considered by these diagnostics; the quantitative ablations, rather than selected examples, provide the corresponding experimental evidence.

Contributions. We introduce AnchorMAE to investigate how dense visual reconstruction and task-semantic supervision can be coupled for fine-grained ego–exo correspondence. Our contributions are threefold: (i) a shared-encoder dual-path framework that separates deployable visual representations from training-only task-conditioned representations, enabling visual-only inference; (ii) dual-stage semantic anchoring and detached same-video frame-level transfer, which provide task-semantic guidance while retaining visual reconstruction targets; and (iii) an evaluation on AE2 and Ego-Exo4D, complemented by component ablations and semantic–temporal diagnostics that examine the proposed coupling.

2 Related Work

Temporal Correspondence and Ego–Exo Reconstruction. Fine-grained video representations must distinguish nearby action stages while remaining robust to viewpoint changes. TCC (Dwibedi et al., 2019) learns from temporal cycle consistency, GTA (Hadji et al., 2021) addresses global temporal alignment, and CARL (Chen et al., 2022) uses sequence-level contrastive supervision. In the unpaired ego–exo setting, AE2 (Xue & Grauman, 2023) learns temporal correspondence despite substantial appearance differences between views. Most directly relevant, BYOV (Park et al., 2025) introduces masked self-view and cross-view prediction over unpaired ego–exo videos, providing dense visual supervision for temporal modeling and cross-view correspondence. However, BYOV's masked prediction provides temporal supervision without explicit task-text targets for individual visual frame representations. AnchorMAE builds on this visual reconstruction setting but studies a different interface for supervision: it maintains detached, pre-semantic visual targets and transfers task-conditioned information to a separate deployable visual pathway at the frame level. The aim is to examine whether explicit task supervision can complement, rather than replace, reconstruction-driven temporal learning.

Language-Guided Video Representation Learning. Video–language pretraining offers semantic signals complementary to visual correspondence. EgoVLP (Lin et al., 2022), EgoVLPv2 (Pramanick et al., 2023), and LaViLa (Zhao et al., 2023) investigate egocentric video–text representation learning and cross-modal supervision. Language and masked video modeling have also been combined explicitly. MILES (Ge et al., 2022) reconstructs masked visual patches using features from an evolving snapshot encoder whose targets become language-aligned, whereas TGM (Fan et al., 2024) uses accompanying text to guide video masking and couples masked autoencoding with video–text contrastive learning. Thus, neither masked learning with language nor local semantic supervision is new in itself. AnchorMAE examines a different coupling: a training-only task-conditioned pathway shares temporal-encoder weights with an unconditioned visual pathway, and detached semantic frame states supervise their corresponding visual states *within each video*. Reconstruction continues to predict pre-semantic visual features, and the deployed pathway does not require a task prompt. Gated cross-attention, used in the semantic adapters, also has established precedent in Flamingo (Alayrac et al., 2022); our focus is its role in this dual-path training configuration, not a new attention operator.

Semantic Cross-View Alignment. A complementary line of research uses semantic supervision to address the ego–exo viewpoint gap. SUM-L (Wang et al., 2023) constructs semantic pseudo-pairs and aligns unpaired multiview video representations; EgoInstructor (Xu et al., 2024) leverages relevant exocentric instructional videos and textual supervision for egocentric learning. VRS (Luo et al., 2025) uses a diffusion-based feature translator trained with synchronized multiview data to support contrastive alignment on larger unpaired collections. The recent PRISM preprint (Chee et al., 2026) separates view-invariant and view-variant latents, uses language-supervised predictive recomposition, and includes frame-level self-prediction to learn temporal dynamics. Thus, prior semantic cross-view work is not devoid of frame-level temporal learning. Its supervisory interface—latent decomposition and recomposition with an additional self-predictive objective—differs from AnchorMAE's frame-wise transfer from a training-only task-conditioned pathway into an unconditioned visual pathway, alongside masked self- and cross-view reconstruction. The transfer is performed between corresponding tokens *of the same video*; it does not assume that class-matched ego and exo clips are temporally synchronized.

Taken together, prior work establishes masked visual temporal prediction, language-mediated cross-view learning, and, in PRISM, explicit frame-level self-prediction within a semantic framework. The specific question addressed here is *how to couple them so that task semantics supervise individual deployable visual representations without altering the visual reconstruction target or requiring text at inference*. AnchorMAE investigates this question through shared-encoder dual-path learning, dual-stage semantic anchoring, task-aware clip-level alignment, and detached frame-level transfer. We examine the proposed coupling through reported objective and mechanism comparisons and semantic–temporal diagnostics rather than attributing gains to architectural combination alone.

3 METHODOLOGY

Masked ego–exo prediction supplies dense visual supervision, while language-guided methods supply task semantics and sometimes frame-level temporal supervision. AnchorMAE addresses a narrower *supervision- interface problem*: transferring task semantics to individual deployable visual frames without changing pre-semantic reconstruction targets or requiring text at inference as shown in Fig. 1. Its unconditioned visual and training-only semantic pathways share temporal- encoder weights; semantic transfer matches frame indices *within* each video, not synchronized ego–exo frames.

Let $v \in \{e, x\}$ denote the egocentric or exocentric view, $\bar{v}$ its opposite, T the sampled sequence length, and d the working feature dimension. A frozen visual backbone, spatial aggregation, and a learned projection produce language-independent column-vector frame features $\mathbf{x}^v_{i,t} \in \mathbb{R}^d$. We stack their transposes as $\mathbf{X}^v_i = [(\mathbf{x}^v_{i,1}); \ldots; (\mathbf{x}^v_{i,T})]^\top \in \mathbb{R}^{T\times d}$, where semicolons denote row-wise stacking. Each training item pairs ego and exo clips with the same activity label y_i, not necessarily the same temporal phase. A frozen text encoder and learned projection map a fixed task description q_i to the normalized anchor $\mathbf{s}_i \in \mathbb{R}^d$. The anchor is shared between both views in the item. We omit i when unambiguous.

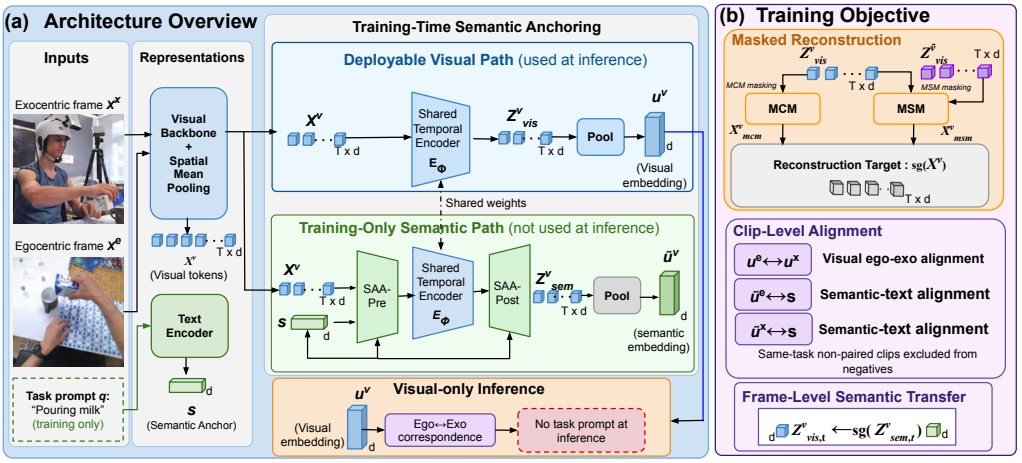

Figure 1: **AnchorMAE dual-path architecture.** A shared temporal encoder processes unconditioned visual features and training-only, SAA-conditioned features in separate passes. Visual masked self-/cross-view reconstruction predicts detached pre-semantic targets. Clip alignment and detached same-video frame-level transfer connect the two pathways; inference retains only the visual path.

3.1 Training-Time Semantic Anchoring

Separate pathways with shared weights. Let $\mathcal{E}_\phi$, $\mathcal{A}_{\text{pre}}$, and $\mathcal{A}_{\text{post}}$ be the temporal encoder, pre, and post semantic anchor adapters, respectively. Rather than injecting text into the only representation available at inference, we construct a second training-time pathway that uses the same encoder parameters:

$$\mathbf{Z}^v_{\text{vis}} = \mathcal{E}_\phi(\mathbf{X}^v), \tag{1}$$

$$\mathbf{Z}^v_{\text{sem}} = \mathcal{A}_{\text{post}}(\mathcal{E}_\phi(\mathcal{A}_{\text{pre}}(\mathbf{X}^v, \mathbf{s})), \mathbf{s}). \tag{2}$$

Both outputs $\mathbf{Z}^v_{\text{vis}}, \mathbf{Z}^v_{\text{sem}} \in \mathbb{R}^{T\times d}$. Only the training-time semantic path depends on the prompt; the visual path supplies test-time features.

Semantic Anchor Adapter. Each Semantic Anchor Adapter (SAA) maps an anchor $\mathbf{s} \in \mathbb{R}^d$ to M key/value slots $\mathbf{M}^K, \mathbf{M}^V \in \mathbb{R}^{M\times d}$ using $\mathbf{W}^K_m, \mathbf{W}^V_m \in \mathbb{R}^{d\times d}$ equation 10. For input $\mathbf{R} \in \mathbb{R}^{T\times d}$, one adapter computes:

$$\begin{aligned} \mathbf{C} &= \text{MHA}\big(\mathbf{R}, \mathbf{M}^K(\mathbf{s}), \mathbf{M}^V(\mathbf{s})\big), \\ \mathbf{Y} &= \text{LN}(\mathbf{R} + \alpha\mathbf{C}), \\ \mathcal{A}(\mathbf{R}, \mathbf{s}) &= \text{LN}(\mathbf{Y} + \text{FFN}(\mathbf{Y})). \end{aligned} \tag{3}$$

Here $\mathbf{C}, \mathbf{Y}, \mathcal{A}(\mathbf{R}, \mathbf{s}) \in \mathbb{R}^{T\times d}$. The two adapters have distinct parameters and gates $\alpha_{\text{pre}}, \alpha_{\text{post}}$ (each initialized to 0.1), denoted generically by α above; within each stage, parameters are shared across views. SAA-Pre and SAA-Post condition features before and after temporal encoding; our contribution concerns their *supervisory role*, not a new attention operator.

3.2 Masked Visual Reconstruction

Preserving a visual reconstruction task. We retain masked *visual* prediction rather than predicting semantic features. Following Park et al. (2025), we apply separate masked-input operators $\text{Mask}_{\mathcal{M}^v_{\text{MSM}}}$ and $\text{Mask}_{\mathcal{M}^v_{\text{MCM}}}$ to $\mathbf{X}^v$ *before* temporal encoding. Their masked-position sets $\mathcal{M}^v_{\text{MSM}}, \mathcal{M}^v_{\text{MCM}} \subseteq \{1, \ldots, T\}$ have sampling ratios 0.40 and 0.80, respectively. For $b \in \{\text{MSM}, \text{MCM}\}$, we define

$$\mathbf{X}^v_b = \text{Mask}_{\mathcal{M}^v_b}(\mathbf{X}^v), \qquad \mathbf{Z}^{v,b}_{\text{vis}} = \mathcal{E}_\phi(\mathbf{X}^v_b). \tag{4}$$

Here the sequence shape is $T_b \times d$: $T_b = T$ with mask-token replacement, or $T_b \leq T$ when masked positions are removed. Unlike equation 1, these states are computed from masked inputs; decoders produce predictions at the original T positions.

MSM targets past-only same-view prediction, which requires causal encoder–decoder information flow; MCM also uses activity-matched opposite-view context:

$$\widehat{\mathbf{X}}^v_{\mathrm{MSM}} = \mathcal{D}^{\mathrm{self}}(\mathbf{Z}^{v,\mathrm{MSM}}_{\mathrm{vis}}), \tag{5}$$

$$\widehat{\mathbf{X}}^v_{\mathrm{MCM}} = \mathcal{D}^{\mathrm{cross}}\left(\mathbf{Z}^{v,\mathrm{MCM}}_{\mathrm{vis}}, \mathbf{Z}^{\bar{v}}_{\mathrm{vis}}\right). \tag{6}$$

Both predictions lie in $\mathbb{R}^{T\times d}$, but losses use only masked positions. The decoder symbols do not imply shared or separate parameters; opposite-view clips share activity, not frame correspondence. Masked-position losses are given in the appendix.

Detached Visual Target. Each decoder predicts its view's detached pre-semantic target $\mathrm{sg}(\mathbf{X}^v)$, where sg stops target-side gradients. The per-item losses average squared vector errors over masked positions, sum over views, and do not divide by d, (refer. equation 11 and equation 12). This isolates target type from task-conditioned supervision; the reported T2 comparison changes the detached target, not the mask task.

3.3 Task-Aware Alignment and Semantic Transfer

Mean pooling and learned, normalized projections yield unmasked clip vectors $\mathbf{u}^v_i, \widetilde{\mathbf{u}}^v_i \in \mathbb{R}^d$ from the visual and semantic paths, respectively, in the anchor's $\mathbf{s}_i \in \mathbb{R}^d$ contrastive space. Clip alignment is complemented by frame-wise supervision.

Task-Aware Clip-Level Alignment. Activity-matched ego–exo clips need not be synchronized: their designated pair is a class-matched pseudo-positive; other same-task clips are excluded from negatives. For a batch of size B with task labels y_i, let $\mathcal{N}_i = \{j \neq i : y_j \neq y_i\}$ be the valid negative indices. For normalized embedding sets $\{\mathbf{a}_i\}$ and $\{\mathbf{b}_i\}$, define the temperature-scaled similarity $h_{ij} = \mathbf{a}_i^\top \mathbf{b}_j/\tau_c$, where $\tau_c > 0$ is the contrastive temperature. The directional loss is

$$\ell^{\mathrm{TA}}_{a\to b} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(h_{ii})}{\exp(h_{ii}) + \sum_{j\in\mathcal{N}_i}\exp(h_{ij})}. \tag{7}$$

We use the symmetric loss $\mathcal{L}^{\mathrm{TA}}_{a\leftrightarrow b} = (\ell^{\mathrm{TA}}_{a\to b} + \ell^{\mathrm{TA}}_{b\to a})/2$ and apply it to visual ego–exo embeddings ($\mathbf{u}^e \leftrightarrow \mathbf{u}^x$), semantic ego–text embeddings ($\widetilde{\mathbf{u}}^e \leftrightarrow \mathbf{s}$), and semantic exo–text embeddings ($\widetilde{\mathbf{u}}^x \leftrightarrow \mathbf{s}$). Repeated same-task prompts are also excluded. Effective batches require cross-task negatives; reported label/mask checks support their presence, but per-run T3 negative counts and batch manifests are unavailable.

Frame-Level Semantic Transfer. To supervise individual deployable frames, we align visual tokens with detached semantic tokens at the same index *within each video*. Here $\mathbf{z}^v_{\mathrm{vis},t} = ((\mathbf{Z}^v_{\mathrm{vis}})_{t,:})^\top \in \mathbb{R}^d$ and $\mathbf{z}^v_{\mathrm{sem},t} = ((\mathbf{Z}^v_{\mathrm{sem}})_{t,:})^\top \in \mathbb{R}^d$ are column vectors. For one paired training item, define

$$\mathcal{L}_{\mathrm{ST}} = \frac{1}{2T}\sum_{v\in\{e,x\}}\sum_{t=1}^{T}\left[1 - \cos\left(\mathbf{z}^v_{\mathrm{vis},t}, \mathrm{sg}(\mathbf{z}^v_{\mathrm{sem},t})\right)\right]. \tag{8}$$

Only this loss detaches the semantic target; semantic clip alignment still trains that path and the shared encoder. Equation 8 is per paired item, with batch aggregation implicit in the joint objective. No cross-video frame match is assumed; task discrimination and temporal ordering are evaluated separately.

Joint objective. The joint objective combines masked visual prediction, task-aware clip alignment, and frame-wise transfer. Let $\lambda_{ex}, \lambda_t, \lambda_{\mathrm{ST}} \geq 0$ weight visual cross-view alignment, semantic–text alignment, and frame-level transfer, respectively. We combine their batch aggregates, with per-item reconstruction terms detailed in the appendix:

$$\mathcal{L}_{\mathrm{AnchorMAE}} = \mathcal{L}_{\mathrm{MSM}} + \mathcal{L}_{\mathrm{MCM}} + \lambda_{ex}\mathcal{L}^{\mathrm{TA}}_{e\leftrightarrow x} + \lambda_t\left(\mathcal{L}^{\mathrm{TA}}_{\tilde{e}\leftrightarrow s} + \mathcal{L}^{\mathrm{TA}}_{\tilde{x}\leftrightarrow s}\right) + \lambda_{\mathrm{ST}}\mathcal{L}_{\mathrm{ST}}. \tag{9}$$

T1 tests transfer, T2 compares detached target types, and T3 tests negative handling. Strictly one-factor T3 interpretation remains conditional because the supplement lacks per-run batch manifests.

Table 1: **AE2 benchmark comparison.** Four-activity macro-averages for Phase F1, frame mAP@10, progression R^2, and Kendallś τ. Bold indicates the largest listed value; published methods can differ in backbone and training data.

Method	Venue	Phase F1↑	mAP@10↑	Prog. R^2↑	Kendall τ↑
Temporal / reconstruction-oriented					
TCC (Dwibedi et al., 2019)	CVPR'19	0.6717	0.7057	0.3216	0.3737
GTA (Hadji et al., 2021)	CVPR'21	0.6963	0.7242	0.3217	0.4638
AE2 (Xue & Grauman, 2023)	NeurIPS'23	0.7596	0.7578	0.4796	0.5621
BYOV (Park et al., 2025)	CVPR'25	0.8234	0.7957	0.7472	0.7455
Semantic / alignment-oriented or transfer methods					
LaViLa (Zhao et al., 2023)	CVPR'23	0.5947	0.6048	-0.2724	0.0693
SUM-L (Wang et al., 2023)	ICCV'23	0.6006	0.6017	0.0152	0.0982
EgoInstructor (Xu et al., 2024)	CVPR'24	0.5495	0.7436	0.7923	0.7777
VRS (Luo et al., 2025)	CVPR'25	0.5525	0.6023	-0.3611	0.0323
AnchorMAE (ours)	–	**0.8687**	**0.8264**	**0.9247**	**0.9966**

Visual-only inference. At inference, only the unconditioned visual pathway and task-specific prediction heads remain. Text, SAA modules, the semantic pass, and training decoders/losses are removed; no latency gain is claimed without profiling.

4 Experiments and Results

We evaluate two complementary properties of ego–exo representations: *fine-grained temporal correspondence* on AE2 (Xue & Grauman, 2023), and *activity-sensitive cross-view transfer* on Ego-Exo4D (Grauman et al., 2024). This pairing addresses the motivating question in Sec. 1: whether training-time task semantics can complement reconstruction-driven temporal supervision in a representation intended for text-free deployment. The reported AnchorMAE metrics are obtained from the author-confirmed unconditioned visual pathway; task descriptions and phase labels are never provided as inference inputs.

AE2: Dense Temporal Correspondence. AE2 comprises four activities (Break Eggs, Pour Milk, Pour Liquid, and Tennis Forehand) and measures phase classification, opposite-view same-phase frame retrieval, phase progression, and temporal ordering. Table 1 reports four-activity averages from the confirmed visual-only checkpoint. AnchorMAE achieves 0.8687 Phase F1, 0.8264 frame mAP@10, 0.9247 progression R^2, and 0.9966 Kendall's τ. These values contextualize the combination of dense reconstruction and training-time semantic supervision relative to selected baselines; the within-framework ablations examine the proposed coupling more directly. Importantly, AE2's activity-specific retrieval and phase metrics alone do *not* measure cross-task semantic discrimination; this is examined separately through the controlled diagnostics in Table 5.

The comparison with BYOV, which explicitly optimizes dense masked visual prediction, contextualizes the temporal side of our research question. The comparisons with language-mediated methods contextualize the interaction between task semantics and phase-sensitive features, but cannot independently establish which proposed module produces an improvement. Reported objective and target comparisons in Tables 3 and 4 address that question more directly. We report one AE2 Kendall's τ, using the consistent nearest-neighbor temporal-order evaluator described in the experimental protocol; we do not introduce a separate "within-task τ" metric.

Ego-Exo4D: Cross-View Retrieval and Recognition. Ego-Exo4D complements the phase-centric AE2 benchmark with activity-scale cross-view retrieval and exo-to-ego recognition. We report the evaluated bidirectional-average HR@K and recognition scores in Table 2. AnchorMAE obtains HR@5/10 of 0.5894/0.6875 and recognition Top-1/Top-5 of 0.4946/0.7558. VRS has a higher listed HR@1 (0.2231 versus 0.1504); thus the observed results do not imply uniformly stronger retrieval at every rank. Taken together, the two benchmarks probe complementary aspects of the motivating reconstruction–semantic gap, but cross-study differences in training supervision and initialization preclude attributing between-method differences solely to architecture.

Table 2: **Ego-Exo4D cross-view retrieval and recognition.** HR@K is the bidirectional retrieval average; Top-1/Top-5 are recognition accuracies. Bold marks the largest listed entry, and a dash indicates an unreported value.

Method	HR@1↑	HR@5↑	HR@10↑	Top-1↑	Top-5↑
Random Align	0.0623	0.2171	0.3353	0.1219	0.3123
EgoVLP (Lin et al., 2022)	0.0897	0.2140	0.3007	0.2033	0.4632
LaViLa (Zhao et al., 2023)	0.1109	0.2347	0.3244	0.2643	0.5501
EgoInstructor (Xu et al., 2024)	–	0.3886	–	0.2415	0.5140
SUM-L (Wang et al., 2023)	0.1676	0.3996	0.5327	0.2483	0.5208
VRS (Luo et al., 2025)	**0.2231**	0.5268	0.6674	0.3447	0.6485
AnchorMAE (ours)	0.1504	**0.5894**	**0.6875**	**0.4946**	**0.7558**

Table 3: **Reconstruction–semantic complementarity on AE2.** Four-activity averages for three objective baselines and AnchorMAE. The full versus Recon+Align comparison varies multiple mechanisms.

Variant	F1↑	mAP@10↑	R^2↑	τ↑
Recon-only	0.7355	0.7923	0.9010	0.7976
Align-only	0.8105	0.7979	0.8868	0.7662
Recon+Align	0.8110	0.8006	0.9023	0.8244
AnchorMAE (ours)	**0.8687**	**0.8264**	**0.9247**	**0.9966**

These results are consistent with the hypothesis that training-time semantic supervision can organize transferable visual features without requiring task text at deployment. However, the mechanism is examined within AnchorMAE by the reported ablations rather than inferred from the published-method comparison. In particular, retrieval HR@K is rank-dependent and cannot be summarized by recognition accuracy alone; both are reported to preserve this distinction.

5 Ablation and Analysis

The main design question is whether dense visual reconstruction and task-semantic supervision provide complementary information *when coupled at the frame level*, rather than merely optimized together. We examine objective complementarity, reported mechanism variants, and representation diagnostics. The supplied point estimates use the author-confirmed visual-only pathway and the stated evaluation protocol. Between-method comparisons are contextual, not architectural controls.

Reconstruction–Semantic Complementarity. Table 3 separates masked visual reconstruction (Recon-only) from clip-level alignment (Align-only), combines the two without semantic adapters or frame-wise transfer (Recon+Align), and compares them with the full two-path design. The purpose is to test whether adding objectives alone supplies the frame-level semantics motivating AnchorMAE, or whether the proposed training-time semantic pathway offers an additional effect. Align-only has higher reported phase F1 than Recon-only, whereas Recon-only retains higher progression R^2 and headline τ. Recon+Align exceeds either single-objective variant on headline τ but does not equal the full configuration. The full row reports increases over Recon+Align across all four metrics; nevertheless, this comparison adds multiple mechanisms simultaneously. T1–T3 report narrower comparisons of the intended mechanisms. These aggregate values do not establish consistency across individual activities.

Critical Training Mechanisms. The distinctions in Table 4 correspond to the proposed solutions to specific weaknesses of a naive reconstruction-plus-alignment design. T1 tests whether *direct frame-level transfer* adds information beyond having a semantic pathway and clip losses. T2 tests whether keeping the *reconstruction target visual* matters, rather than making the model predict text-conditioned features. T3 examines how activity-matched but nonsynchronized ego–exo pairs are treated in the contrastive objective. A one-factor interpretation requires keeping the semantic pathway active in T1, detaching both target types in T2, and fixing minibatches and all non-contrastive

Table 4: **Critical training mechanisms on AE2.** T1–T3 compare reported variants for frame-level transfer, reconstruction targets, and contrastive negative handling. Bold marks the largest listed value within each block and metric.

Variant	F1↑	mAP@10↑	R^2↑	τ↑
T1. Frame-level semantic transfer				
Without transfer	0.8096	0.7826	0.9072	0.9211
With transfer	**0.8687**	**0.8264**	**0.9247**	**0.9966**
T2. Reconstruction target				
Semantic target	0.7393	0.7121	0.8613	0.9615
Detached visual target	**0.8687**	**0.8264**	**0.9247**	**0.9966**
T3. Contrastive negative handling				
Standard InfoNCE	0.6973	0.7381	0.8808	**0.9976**
Multi-positive	0.7871	0.8085	0.9076	0.9886
Same-task exclusion	**0.8687**	**0.8264**	**0.9247**	0.9966

Table 5: **Activity/view probes and controlled distractors on AE2.** Probe scores are fractions; distractor accuracies are percentages. Near-chance linear viewpoint decoding does not prove invariance; fixed-candidate accuracy is distinct from frame mAP@10.

Variant	Task probe↑	View probe→ 0.5	Semantic distractor (%)↑	Temporal distractor (%)↑
Recon-only	0.5667	0.5335	40.04	74.97
Align-only	0.6443	0.4995	61.67	71.66
Recon+Align	0.8333	0.4887	89.97	71.74
AnchorMAE (ours)	**0.9876**	**0.4997**	**98.33**	**75.04**

terms in T3. These are the intended controls; the supplied material does not include the per-run records needed to audit all of them.

The supplied T1 and T2 results are consistent with the benefit of frame-level transfer and preserving visual reconstruction targets; they do not establish that semantic targets universally cause representation collapse. In T3, same-task exclusion has higher F1, mAP@10, and progression R^2, while standard InfoNCE has marginally higher headline τ (0.9976 versus 0.9966). T3 is described as varying positive/negative assignment under otherwise matched training settings, but its run-level batch records are unavailable. The one-positive exclusion variant removes non-paired same-task clips from the denominator rather than treating them as negatives.

Representation and Controlled-Distractor Diagnostics. Clip alignment alone does not prove that visual frames encode task identity; phase-correspondence metrics do not detect cross-task confusion. Table 5 therefore reports task and viewpoint linear probes and separate, fixed-candidate semantic and temporal distractor tests. For each query, cosine similarity must rank a correct same-task/same-phase opposite-view frame above the corresponding distractor. Semantic distractors come from different tasks; temporal distractors are same-task but different-phase frames. The reported protocol uses frozen visual-only embeddings, held-out queries, and fixed candidate-galleries for each comparison; exact query/gallery manifests are not included in the supplied supplement. The gap between task-probe and distractor results for Recon+Align and the full model is consistent with the contribution of frame-sensitive task supervision beyond combining reconstruction and clip-level alignment. The temporal-distractor difference is smaller: 75.04% versus 71.74% for Recon+Align, and only 0.07 percentage points above Recon-only. The view probe approaches chance for AnchorMAE, although a nonlinear classifier may still recover viewpoint cues. We use only the standard AE2 nearest-neighbor Kendall's τ for temporal order and do not present an additional ambiguously defined within-task score.

6 Qualitative Analysis

Figure 2 compares opposite-view retrievals for the supplied egocentric *Pour Liquid* query. The displayed annotations identify a Recon-only cross-task retrieval (Break Eggs) and same-task phase mismatches for Align-only and Recon+Align; the full-model example is annotated as the query's



Figure 2: **Qualitative retrieval comparison.** Top-1 opposite-view retrievals from four models for the same ego query. The annotated results illustrate a cross-task error, same-task phase mismatches, and a matching task/phase example. Annotations are for display only, never inference inputs.

Pouring phase. The panel is illustrative: the annotations must be checked against source records, and selected examples cannot establish error prevalence or causation. Tables 3 and 4 report the complementary ablation evidence, while Table 5 reports separate fixed-candidate task/phase diagnostics.

7 Conclusion

AnchorMAE couples masked visual reconstruction and task-level semantic supervision through shared-encoder dual pathways. A training-only semantic pathway applies pre-/post-encoder anchoring and transfers detached frame-level supervision to the deployable visual pathway, while reconstruction retains pre-semantic targets. The visual pathway alone runs at inference without task text. AE2 and Ego-Exo4D results, reported objective ablations, and task/phase diagnostics examine both activity discrimination and temporal correspondence. Gains vary by retrieval rank, and the temporal-distractor improvement over Recon-only is modest. Generalization beyond fixed prompts and the studied activities remains an important direction.

AI USE STATEMENT

Generative AI tools were used solely to aid, polish and paraphrase the writing of this manuscript, specifically for language editing and improving clarity and readability. Grammarly was used for this purpose. All AI-assisted text was reviewed and revised by the authors. The authors take full responsibility for the final content of the paper, including all technical claims, mathematical formulations, experimental protocols, citations, and reported results.

ETHICS STATEMENT

This work studies representation learning for egocentric and exocentric video using established research datasets and does not introduce a new human-subject study. The method may be relevant to applications involving wearable cameras and human activity understanding, which can raise privacy, consent, and surveillance concerns when deployed in real-world settings. Our experiments are restricted to research benchmarks and are intended to study cross-view visual representation learning rather than to identify individuals or infer sensitive personal attributes. We encourage downstream users to consider dataset licenses, consent and privacy requirements, demographic and environmental biases, and the potential consequences of deploying video-understanding systems in safety- or privacy-sensitive settings.

REPRODUCIBILITY STATEMENT

Section 3 specifies AnchorMAE's visual and training-only semantic pathways, shared temporal encoder, masked reconstruction, task-aware alignment, frame-level semantic transfer, and visual-only inference procedure. Section 4 reports benchmark metrics and experimental comparisons, while Section 5 presents reported mechanism ablations and representation diagnostics. The appendix provides additional mathematical details and the specified training objectives (Appendix A.1), further ablation results (Appendix A.3), definitions and interpretation of evaluation controls (Appendix A.2), and illustrative qualitative analyses (Appendix A.4). Complete per-run checkpoint/configuration manifests, effective-negative counts, numerical hardware-profiling measurements, and anonymous source code are not included in the present materials.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Youngchae Chee, Hosu Lee, Sungjune Park, Junho Kim, and Yong Man Ro. PRISM: predictive recomposition via semantic latent decomposition for view-invariant video representation learning. *CoRR*, 2026.

Minghao Chen, Fangyun Wei, Chong Li, and Deng Cai. Frame-wise action representations for long videos via sequence contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13801–13810, 2022.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1801–1810, 2019.

David Fan, Jue Wang, Shuai Liao, Zhikang Zhang, Vimal Bhat, and Xinyu Li. Text-guided video masked autoencoder. In *Computer Vision – ECCV 2024*, volume 15063 of *Lecture Notes in Computer Science*, pp. 282–298, 2024.

Yuying Ge, Yixiao Ge, Xihui Liu, Jinpeng Wang, Jianping Wu, Ying Shan, Xiaohu Qie, and Ping Luo. MILES: visual BERT pre-training with injected language semantics for video-text retrieval. In *Computer Vision – ECCV 2022*, volume 13695 of *Lecture Notes in Computer Science*, pp. 691–708, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zachary Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, María Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Dutt Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J. Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina González, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran K. Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbeláez, Gedas Bertasius, Dima Damen, Jakob J. Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard A. Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shout, and Michael Wray. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19383–19400, 2024.

Isma Hadji, Konstantinos G. Derpanis, and Allan D. Jepson. Representation learning via global temporal alignment and cycle-consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11068–11077, 2021.

Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, Chengfei Cai, Hongfa Wang, Dima Damen, Bernard Ghanem, Wei Liu, and Mike Zheng Shou. Egocentric video-language pretraining. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. Viewpoint rosetta stone: Unlocking unpaired ego-exo videos for view-invariant representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15802–15812, 2025.

Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Bootstrap your own views: Masked ego-exo modeling for fine-grained view-invariant video representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13661–13670, 2025.

Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. EgoVLPv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5285–5297, 2023.

Qitong Wang, Long Zhao, Liangzhe Yuan, Ting Liu, and Xi Peng. Learning from semantic alignment between unpaired multiviews for egocentric video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3307–3317, 2023.

Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13525–13536, 2024.

Zihui Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53688–53710, 2023.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6586–6597, 2023.