MACJEPA: MISSINGNESS-ROBUST AUDIO-VISUAL RECOGNITION FROM UNTRIMMED EGOCENTRIC VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-visual models improve egocentric action recognition by exploiting complementary cues, yet typically assume that both streams remain available at inference. Existing missing-modality methods operate on trimmed, single-event clips in which a stream is entirely present or absent, whereas real sensors fail and recover within long, untrimmed observations. We redefine egocentric modality missingness as temporally localized sensor outages within untrimmed, multi-event observations, with whole-clip absence as the limiting case. We introduce **MacJEPA**, a missing-modality-robust **Ma**sked-**c**ontext query **JEPA** that recognizes visual actions and acoustic events from supplied interval queries over audio-visual context. Window-local modality dropout simulates these sensor outages during training. MacJEPA further repurposes masking in JEPA from a self-supervised pretext into a supervised robustness objective, aligning masked and clean latent representations of both multimodal content tokens and the task-conditioned queries. All objectives are optimized jointly with recognition in a single stage, requiring no test-time adaptation. Across Epic-Kitchens-100 and Epic-Sounds, a single checkpoint remains competitive under complete input and consistently surpasses published missing-modality baselines when either the dominant or auxiliary stream is removed. MacJEPA thus unifies strong full-input recognition with temporal missing-modality robustness in a single model operating on untrimmed multi-event videos.

1 INTRODUCTION

Egocentric video captures activity from the actor's perspective, and understanding it underpins assistive systems, robotics, and augmented reality (Damen et al., 2022; Grauman et al., 2022). Vision alone is a fragile signal in this setting; head motion, occlusion, and out-of-view interactions often hide evidence, and many egocentric actions remain acoustically identifiable when visual evidence is occluded or out of view. (Kazakos et al., 2019; Huh et al., 2023). Audio-visual models exploit this complementarity and define the state of the art in egocentric action recognition through convolutional or transformer-based models (Kazakos et al., 2019; 2021a; Nagrani et al., 2021; Chalk et al., 2024). Nevertheless, their gains are contingent on the availability of both streams at inference time, a premise that is often violated during deployment. For example, in real-world wearable devices, microphones or recordings can be suppressed for privacy, sensors can be duty-cycled to conserve power, or hardware can malfunction (Gong et al., 2023; Ramazanova et al., 2025a;b; Santos-Villafranca et al., 2026). The degradation is rarely graceful; losing one stream can push a multimodal model below a unimodal model trained on the surviving stream (Ramazanova et al., 2025a). This makes missing-modality resilience a deployment necessity rather than an optional refinement.

Existing egocentric missing-modality methods remain tied to atomic clips centered on a single annotated event, treating each stream as present or absent throughout the sample (Gong et al., 2023; Ramazanova et al., 2025a;b; Santos-Villafranca et al., 2026; Ma et al., 2022; Lee et al., 2023). Missingness is consequently measured across clips, which implicitly align sensor outages with action boundaries. Even methods incorporating temporal context either aggregate neighboring action-centered clips or assume complete modalities over an untrimmed stream (Kazakos et al., 2021a; Chalk et al., 2024). Deployment grants no such action boundaries, requiring systems to process background, partial actions, and multiple temporally overlapping visual and acoustic events (Damen

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.

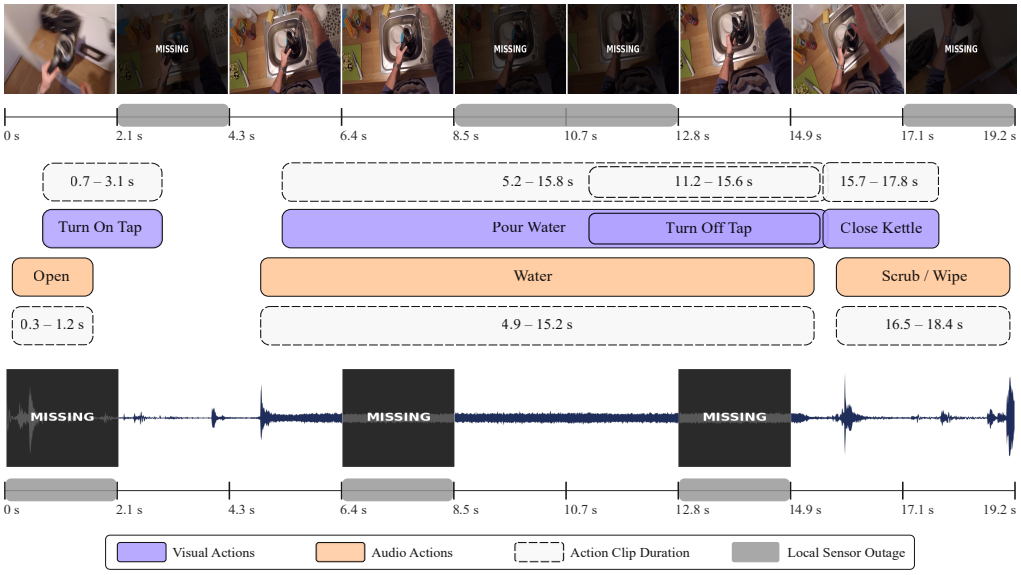

Figure 1: Local sensor outages across overlapping multimodal events in a 19.2-second untrimmed observation.

et al., 2022; Huh et al., 2023; Chalk et al., 2024). Within such continuous observations, either sensor may disappear and return, interleaving clean and corrupted evidence at different intervals. Since overlapping actions share this context, one local outage can affect recognition of several events at different times and to varying degrees. Figure 1 follows a plausible wearable-kitchen sequence where the wearer turns on the tap, pours water, turns it off, and closes the kettle, accompanied by the acoustic events *Open*, *Water*, and *Scrub/Wipe*. Initial privacy muting removes acoustic evidence for *Open* before audio returns, while a brief camera failure clips visual evidence for *Turn On Tap* before video recovers. During the overlapping *Pour Water*, *Turn Off Tap*, and *Water* events, audio is lost over 6.4–8.5 s, video over 8.5–12.8 s, and audio again over 12.8–14.9 s, repeatedly changing which stream can support recognition. These local outages therefore cause multiple overlapping events to lose different degrees of evidence, revealing what atomic clip-level binary missingness cannot represent. The operative question is no longer which modality is absent from a clip, but which is absent *when* and which events consequently lose distinguishable evidence.

We introduce **MacJEPA**, a missing modality robust **Ma**sked-**c**ontext query **J**oint-**E**mbedding **P**redictive **A**rchitecture that recognizes actions by querying supplied event intervals within untrimmed audio-visual context. During training, MacJEPA independently drops each modality over short intervals within an untrimmed observation and substitutes them with learned mask vectors. Each query must preserve its prediction from surviving temporally aligned evidence, making the missingness in Figure 1 a training condition rather than a deployment-time distribution shift. Canonical JEPA models train a masked online encoder to match clean Exponential Moving Average teacher targets in the latent space, providing an unexplored promising direction for this robustness machinery (Baevski et al., 2022; Assran et al., 2023; 2025). Moreover, aligning masked multimodal context with clean targets in a shared latent space avoids reconstructing heterogeneous audio and visual signals (Lei et al., 2025). Existing frameworks, however, use masking as a self-supervised pretext for transferable features rather than to model deployment failures (Assran et al., 2025; Teotia et al., 2026; Robson et al., 2026; Lei et al., 2025). MacJEPA repurposes JEPA's masked-to-clean consistency as a robustness objective optimized jointly with task supervision.

Our contributions are threefold. **(1)** We redefine egocentric modality missingness as an interval property of multi-event, untrimmed observations, with whole-clip binary absence as its limiting case. **(2)** We introduce interval-wise modality dropout with typed, time-stamped mask tokens, modeling local sensor outages and recoveries within one observation. **(3)** We reimagine JEPA as a task-supervised robustness objective that aligns masked and clean latent states of multimodal content and task-conditioned queries. With complete input, **MacJEPA** achieves competitive accuracies of 54.0% on Epic-Kitchens action and 58.1% on Epic-Sounds. Under total dominant-modality loss, it reaches 49.5% Epic-Kitchens verb and 50.0% Epic-Sounds sound accuracy, surpassing the strongest

published baseline by 4.2 and 9.3 points at this protocol-matched endpoint (Ramazanova et al., 2025a).

2 Related Work

Multimodal Egocentric Recognition under Missing Modalities. Audio complements egocentric vision because contact and state changes remain audible when hands or objects leave the frame (Kazakos et al., 2019; Huh et al., 2023). Convolutional designs include TBN, which fuses RGB, flow, and audio; AVSlowFast, which couples fast audio with SlowFast vision; and ASF, which adapts this design to acoustic recognition (Kazakos et al., 2019; Xiao et al., 2020; Kazakos et al., 2021b). Transformers advance this domain by learning fusion through MBT's bottleneck tokens, MTCN's neighboring clips, and M&M Mix's multiple views, while TIM queries both streams over one time-encoded untrimmed timeline (Nagrani et al., 2021; Kazakos et al., 2021a; Xiong et al., 2022; Chalk et al., 2024). These strong baselines assume continuous access to both modalities, while missing-modality methods remain sample-atomic. MMG-Ego4D evaluates withheld modality subsets, MMT learns an absent-stream token, and KARMMA distills a multimodal teacher into a dropout-trained student supporting any modality combination (Gong et al., 2023; Ramazanova et al., 2025a; Santos-Villafranca et al., 2026). MIDL instead adapts at inference while distilling full-modality predictions, whereas related work distills multimodal teachers into RGB-only students (Ramazanova et al., 2025b; Radevski et al., 2023). Fusion search and missing-aware prompting retain the same sample-wise binary premise (Ma et al., 2022; Lee et al., 2023), measuring affected action-centered samples rather than outage timing. MacJEPA unifies these lines by modeling disappearance and recovery within shared, multi-event untrimmed context, with clip-level absence as its limiting case.

Multimodal JEPAs. JEPAs predict abstract targets rather than reconstruct raw inputs, avoiding modality-specific low-level detail in heterogeneous settings (Assran et al., 2023; Baevski et al., 2022; Fei et al., 2023; Lei et al., 2025). M3-JEPA aligns modalities with a multi-gate mixture-of-experts predictor, while MJEPA uses a shared encoder for within- and cross-modal audio-video prediction (Lei et al., 2025; Teotia et al., 2026). AV-JEPA aligns global audio-visual and local unimodal views through early fusion and modality dropout, whereas VL-JEPA predicts continuous text embeddings from vision (Robson et al., 2026; Chen et al., 2026). Across these methods, masking remains a self-supervised pretext for transferable features rather than a deployment failure, and robustness to incomplete inputs is not evaluated. MJEPA leaves this direction to future work (Teotia et al., 2026). MacJEPA repurposes masking to simulate sensor outages, optimizes it jointly with supervision, and extends prediction from content tokens to task-conditioned queries.

3 Proposed Framework

Figure 2 summarizes MacJEPA. We first formulate interval-conditioned recognition over untrimmed observations and construct its query sequence in Sections 3.1 and 3.2. We then introduce our two core components: window-local modality dropout in Section 3.3 and masked-to-clean content and query JEPA, trained jointly with recognition, in Sections 3.4 and 3.5.

3.1 Problem Setting

We formulate audio-visual recognition directly on the untrimmed timeline rather than around isolated event crops. For each modality $m \in \mathcal{M} = \{\mathrm{v}, \mathrm{a}\}$, a frozen unimodal encoder b^m produces temporally ordered features on a shared grid with stride Δ seconds and receptive field $\omega = 2\Delta$ seconds. An observation is a contiguous slice of T rows, $\boldsymbol{X}^m = [\boldsymbol{x}_0^m, \ldots, \boldsymbol{x}_{T-1}^m]$, spanning the nominal duration $S = T\Delta$ seconds. Candidate observations tile each video at one-row intervals, independently of its annotations. They may therefore begin or end within an event, include substantial background, or contain no annotated visual action. An annotation is assigned to an observation when its intersection with the observation is at least $\delta = 0.2\,\mathrm{s}$, or when the annotation is fully contained within it, following Chalk et al. (2024). A training observation contains multiple overlapping Epic-Kitchens visual action annotations and Epic-Sounds acoustic event annotations. The intervals are supplied to the model rather than inferred, so the task is interval-conditioned action recognition. To simulate temporal missingness, we partition each observation into $K = T/w$ non-overlapping windows of w consecutive rows spanning $\tau = w\Delta$ seconds.

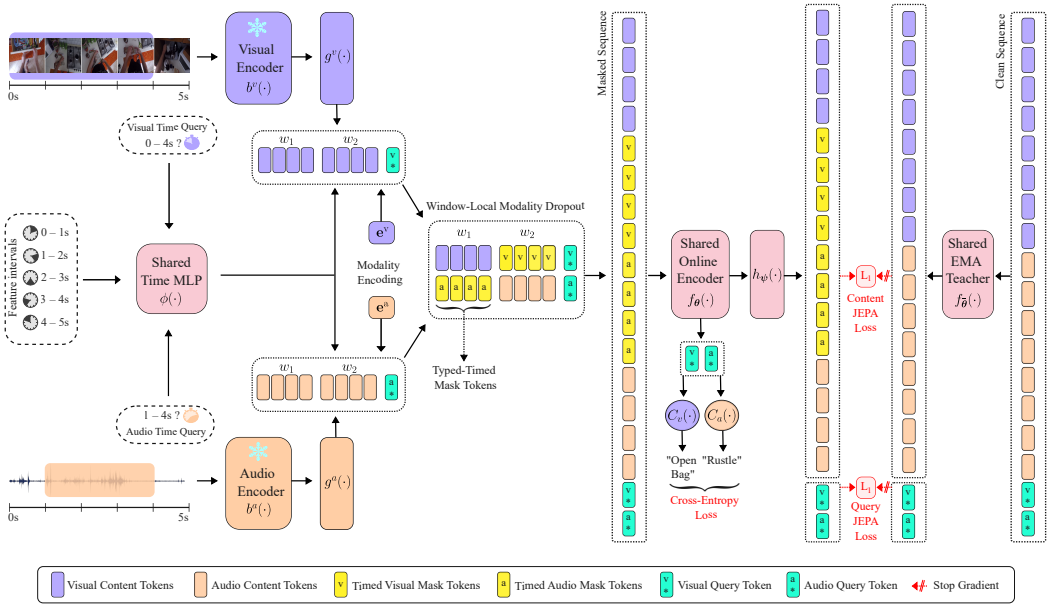

Figure 2: **MacJEPA training framework.** Window-local modality dropout corrupts the online multimodal sequence with typed, timed masks, while content and query JEPA align it with clean EMA-teacher targets under joint task supervision. Two windows and one query per stream are shown for clarity.

3.2 Interval Queries over an Untrimmed Observation

Multiple events share each observation, so time must identify both where evidence occurs and which event is being predicted. For each modality $m \in \mathcal{M}$, a modality-specific stem $g^m : \mathbb{R}^{D_m} \to \mathbb{R}^{d_c}$ applies feature dropout and projects the frozen features into a common content space. We then encode every feature and query on the same temporal coordinate system. A shared TimeMLP $\phi : \mathbb{R}^2 \to \mathbb{R}^{d_t}$ maps continuous intervals to temporal representations, providing a universal clock across both modalities (Chalk et al., 2024). Content row j is assigned $\mathbf{t}_j = [j\Delta, j\Delta + \omega]/S$, while each annotation interval is clipped to the observation and normalized to obtain $\mathbf{t}_q \in [0,1]^2$. A learned embedding $\boldsymbol{e}^m \in \mathbb{R}^d$ marks the modality of every token. Writing $j = kw + f$, the content token $\boldsymbol{u}^m_{k,f}$ at row f of window k and the token $\boldsymbol{u}_q$ for query q are:

$$\begin{aligned} \boldsymbol{u}^m_{k,f} &= \left[g^m\left(\boldsymbol{x}^m_{kw+f}\right) \,\|\, \phi(\mathbf{t}_{kw+f}) \right] + \boldsymbol{e}^m, \\ \boldsymbol{u}_q &= \left[\boldsymbol{c}_{h(q)} \,\|\, \phi(\mathbf{t}_q) \right] + \boldsymbol{e}^{m(q)}, \end{aligned} \tag{1}$$

where $\boldsymbol{c}_{h(q)}$ is a learned seed for the query head and $m(q)$ identifies its stream. We concatenate d_c and d_t, producing tokens of width $d = d_c + d_t$. The resulting sequence contains $n = 2T$ multimodal content tokens, followed by verb, noun, and action queries for each Epic-Kitchens annotation and one sound query for each Epic-Sounds annotation. The encoder restricts content tokens to attending only to content, while each query attends to all content and itself but not other queries. This keeps the encoded context independent of the query set and each prediction independent of co-processed queries, with every classifier reading only its corresponding query representation. This construction temporally locates evidence and queries but assumes complete content; we model local absence next.

3.3 Window-Local Modality Dropout with Typed, Timed Masks

We introduce window-local modality dropout to make modality availability vary within the observation. The T rows of an observation are partitioned into $K = T/w$ non-overlapping windows of w rows, and a binary mask $\boldsymbol{M} \in \{0,1\}^{K \times |\mathcal{M}|}$ specifies which modality is absent from each window. In particular, $M_{k,m} = 1$ removes all w rows of modality m from window k; it is a window-level decision, not an independent token mask. Windows are sampled independently, with probabilities $1 - \pi$, $\pi/2$, and $\pi/2$ for the complete, video-missing, and audio-missing states, respectively. Simultaneous loss of both streams is excluded so that every interval retains contemporaneous evidence from at least

one sensor. Missing features are substituted and not deleted. Let $\boldsymbol{e}^m_{\varnothing} \in \mathbb{R}^{d_c}$ be a learned absence embedding for modality m. The masked content token is:

$$\widetilde{\boldsymbol{u}}^m_{k,f} = \left[(1 - M_{k,m})g^m\left(\boldsymbol{x}^m_{kw+f}\right) + M_{k,m}\boldsymbol{e}^m_{\varnothing} \;\middle\|\; \phi(\mathbf{t}_{kw+f})\right] + \boldsymbol{e}^m. \tag{2}$$

When $M_{k,m} = 0$, this recovers the observed content token from Equation (1); when $M_{k,m} = 1$, only its projected content is replaced. The token remains in the sequence and retains its time encoding and modality tag. It therefore represents a typed, timed absence: the encoder is told which stream is unavailable and where. Furthermore, window grouping gives the simulated outage meaningful duration. Adjacent rows overlap by 50%, so masking one row would leave much of its temporal support available through its neighbors. Masking w rows creates a $\tau = w\Delta$ outage, which exceeds the $\omega = 2\Delta$ receptive field of any single feature when $w > 2$. The windows are then flattened before encoding and impose no attention boundary. Every query still attends to all $n = 2T$ content tokens, allowing it to route around a local outage and draw on evidence elsewhere in the observation. Conditional on missingness, the sampler has 3^K possible window configurations, redrawn each epoch, instead of the three states available to clip-level dropout; complete, video-missing, or audio-missing (Xiao et al., 2020; Ramazanova et al., 2025a; Santos-Villafranca et al., 2026). It therefore exposes a single event to partial evidence as sensors disappear and return. Whole-observation absence remains within its support as the limiting pattern $M_{k,m} = 1$ and $M_{k,\bar{m}} = 0$ for every k. Further details are in appendix Section A.1. We next regularize representations under these outages toward their clean counterparts.

3.4 Content and Query JEPA under Simulated Outages

Let $\mathcal{U}(\boldsymbol{X}^{\mathrm{v}}, \boldsymbol{X}^{\mathrm{a}}, \mathcal{Q}; \boldsymbol{M})$ denote the multimodal content-query sequence assembled under mask $\boldsymbol{M}$. A transformer encoder $f_{\boldsymbol{\theta}}$ processes the masked sequence, while an exponential-moving-average teacher $f_{\bar{\boldsymbol{\theta}}}$ processes its clean counterpart (Figure 2):

$$\begin{aligned}(\boldsymbol{Z}^c_{\boldsymbol{\theta}}, \boldsymbol{Z}^q_{\boldsymbol{\theta}}) &= f_{\boldsymbol{\theta}}(\mathcal{U}(\boldsymbol{X}^{\mathrm{v}}, \boldsymbol{X}^{\mathrm{a}}, \mathcal{Q}; \boldsymbol{M}))\,,\\ (\bar{\boldsymbol{Z}}^c, \bar{\boldsymbol{Z}}^q) &= f_{\bar{\boldsymbol{\theta}}}(\mathcal{U}(\boldsymbol{X}^{\mathrm{v}}, \boldsymbol{X}^{\mathrm{a}}, \mathcal{Q}; \mathbf{0}))\,.\end{aligned} \tag{3}$$

Both branches receive the same observation and queries; only the online branch contains simulated outages. The teacher receives no gradients and is updated as $\bar{\boldsymbol{\theta}} \leftarrow \alpha\bar{\boldsymbol{\theta}} + (1 - \alpha)\boldsymbol{\theta}$, providing stable, contextualized clean targets (Grill et al., 2020; Assran et al., 2023; Baevski et al., 2022). Predicting these targets in a shared latent space avoids reconstructing heterogeneous audio and visual signals in their incomparable input spaces (Lei et al., 2025). A single token-wise predictor $h_{\boldsymbol{\psi}}$, implemented as a $d \to 384 \to d$ bottleneck MLP, maps online states ($\boldsymbol{z}^c_{\boldsymbol{\theta}}$, $\boldsymbol{z}^q_{\boldsymbol{\theta}}$) towards their clean targets ($\bar{\boldsymbol{z}}^c$, $\bar{\boldsymbol{z}}^q$). It intentionally has no attention and receives no mask description beyond the typed, timed absence already encoded by $f_{\boldsymbol{\theta}}$. Predictor conditioning and capacity determine whether a JEPA preserves or removes information about a transformation (Garrido et al., 2024). Our limited, unconditioned predictor favors outage-invariant readouts and limits the train-only module from absorbing the cross-token fusion from the online encoder; more details are in appendix Section A.2.

For $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, define $d_1(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_1/d$. We regress both content and valid query states $\mathcal{Q}_{\mathrm{valid}} \subseteq \mathcal{Q}$ onto parameter-free normalized, stop-gradient teacher targets:

$$\begin{aligned}\ell_{\mathrm{ctx}} &= \frac{1}{n}\sum_{i=1}^{n} d_1\left(h_{\boldsymbol{\psi}}(\boldsymbol{z}^c_{\boldsymbol{\theta},i}), \mathrm{LN}_0(\mathrm{sg}[\bar{\boldsymbol{z}}^c_i])\right),\\ \ell_{\mathrm{qry}} &= \frac{1}{|\mathcal{Q}_{\mathrm{valid}}|}\sum_{q \in \mathcal{Q}_{\mathrm{valid}}} d_1\left(h_{\boldsymbol{\psi}}(\boldsymbol{z}^q_{\boldsymbol{\theta},q}), \mathrm{LN}_0\left(\mathrm{sg}\left[\bar{\boldsymbol{z}}^q_q\right]\right)\right),\\ \mathcal{L}_{\mathrm{JEPA}} &= \ell_{\mathrm{ctx}} + \ell_{\mathrm{qry}}.\end{aligned} \tag{4}$$

The two terms are normalized separately, giving content and query states equal total weight. Unlike masked-position objectives (Baevski et al., 2022; Assran et al., 2023; Teotia et al., 2026), ℓ_{ctx} covers all $n = 2T$ content states. Since global attention propagates corruption to observed tokens, this aligns their context-dependent representations with clean targets rather than merely filling masked positions. Crucially, content states cannot attend to queries, so $\partial\ell_{\mathrm{ctx}}/\partial\boldsymbol{c}_h = 0$. Although ℓ_{ctx} updates the shared encoder, it cannot directly regularize the learned query seed read by classifier C_h. The query term instead aligns the same task and interval under corrupted and clean context. It therefore places masked-to-clean consistency directly on the prediction representation, subsequently supervised by

cross-entropy. This encourages missing-modality invariance in task-conditioned readouts. Window-local dropout creates outage conditions to enable query attention to reroute toward surviving evidence. Content JEPA stabilizes content representations, and query JEPA preserves the prediction readout, allowing robustness to emerge as the encoder combines surviving temporal and cross-modal evidence in latent space. The teacher and predictor are discarded after training.

3.5 Learning Objective

For task heads $\mathcal{H} = \{\text{verb}, \text{noun}, \text{action}, \text{sound}\}$, let $\mathcal{Q}_h \subseteq \mathcal{Q}_{\text{valid}}$ denote the valid, unpadded queries for head h. Each classifier C_h predicts from its corresponding interval-conditioned query readouts using label-smoothed cross-entropy with $\varepsilon = 0.1$:

$$\mathcal{L}_{\text{CE}} = \sum_{h \in \mathcal{H}} \frac{1}{|\mathcal{Q}_h|} \sum_{q \in \mathcal{Q}_h} \text{CE}_\varepsilon \Big(C_h(\boldsymbol{z}^q_{\boldsymbol{\theta},q}), y_q \Big) . \tag{5}$$

We additionally apply VICReg to observed content states as a representational collapse guard, excluding substituted tokens so absence embeddings do not match real sensor statistics (Bardes et al., 2022). All objectives are optimized jointly; complete details are in appendix Sections A.4 and C.7:

$$\mathcal{L} = w_{\text{ce}}\mathcal{L}_{\text{CE}} + w_{\text{jepa}}\mathcal{L}_{\text{JEPA}} + w_{\text{vic}}\mathcal{L}_{\text{VIC}}. \tag{6}$$

4 Experiments

For all experiments, we set $\Delta = 16/30\,\text{s}$, $\omega = 32/30\,\text{s}$, $T = 36$, $w = 4$, $d_c = 768$, $d_t = 256$, $\pi = 0.7$. This gives $S = 19.2\,\text{s}$, $K = 9$, $\tau = 2.13\,\text{s}$, $n = 72$, and $d = 1024$. Following the setup in Section 4.1, we examine how MacJEPA learns missing-modality robustness through controlled design hypotheses (Section 4.2) and qualitative evidence-routing analysis (Section 4.3), then compare the same fixed checkpoint with published benchmarks under complete input and across the missingness spectrum (Section 4.4).

4.1 Experimental Setup

Data. We jointly evaluate visual actions from Epic-Kitchens-100 and acoustic events from Epic-Sounds, which annotate the same 100 hours of untrimmed egocentric video with distinct temporal intervals (Damen et al., 2022; Huh et al., 2023). Epic-Kitchens contains 97 verbs and 300 nouns; we classify the 3,806 verb–noun pairs or action labels present in train and validation set, while Epic-Sounds contains 44 sound classes. We use the official training and validation annotations. We evaluate 9,668 Epic-Kitchens validation and 8,045 Epic-Sounds validation segments.

Setup. We concatenate frozen Omnivore Swin-B and VideoMAE-L features into a $2{,}048$-D visual input and use $2{,}304$-D Auditory SlowFast R50 audio features (Girdhar et al., 2022; Tong et al., 2022; Kazakos et al., 2021b). Each 19.2 s observation has 36 aligned rows grouped into nine 2.13 s windows. 768-D content and 256-D time representations form 1,024-D tokens for a 6-layer, 8-head transformer with MLP ratio 2. Training uses AdamW for 50 epochs with batch size 256, learning rate 3×10^{-4}, weight decay 10^{-4}, two-epoch warm-up, and cosine decay. During training, each observation window independently drops one uniformly chosen modality with probability 0.7. Evaluation uses two missingness protocols. **Nested per-window temporal missingness.** To evaluate robustness to local sensor disappearance and recovery within an observation, controlled analyses in Section 4.2 and appendix Section B use fixed nested masks that remove the selected modality from each observation window with probability $r \in \{0, 0.25, 0.50, 0.75, 1\}$. At $r = 0.75$, this affects 6.75 of nine windows on average, equivalent to 27 aligned rows or 14.4 s of the validation observations. **Observation-level missingness.** For comparison with prior work in Section 4.4.2, we follow MMT and MIDL to define missingness; a fraction r of validation observations loses the selected modality for the full $19.2\,\text{s}$, while the remainder stay complete (Ramazanova et al., 2025a;b). Following Chalk et al. (2024), we average logits across overlapping observations containing each annotation and report annotation-level top-1 accuracy. Video is dominant for Epic-Kitchens and audio for Epic-Sounds, with the reverse defining weaker-modality removal. All results use the same epoch-50 checkpoint. Appendix Sections B and C details feature extraction, implementation, optimization, and ablations of drop probabilities, window and observation lengths, feature backbones, and encoder depth.

Table 1: Controlled component analysis under dominant-modality removal. Top-1 accuracy (%) for Epic-Kitchens action and Epic-Sounds sound reported. Follows nested per-window temporal missingness protocol.

Ablation	Epic-Kitchens action, V drop					Epic-Sounds sound, A drop				
	0	25	50	75	100	0	25	50	75	100
Unimodal reference	13.0	13.0	13.0	13.0	13.0	41.4	41.4	41.4	41.4	41.4
H1: Encoder + CE	53.6	51.6	45.7	27.8	4.1	56.5	55.1	53.5	45.2	26.3
H2: H1 + window-local modality dropout	52.9	51.7	48.7	44.0	16.3	55.5	54.4	53.8	51.0	46.7
H3: H2 + content JEPA	53.6	53.1	50.8	45.2	18.2	56.1	55.5	**55.0**	51.5	48.4
H4: H3 + query JEPA (**MacJEPA**)	**54.0**	**53.7**	**52.2**	**47.3**	**19.5**	**58.1**	**56.0**	54.3	**52.1**	**50.0**

4.2 What Produces Missing-Modality Robustness in MacJEPA?

Table 1 isolates each component's robustness contribution under dominant-modality removal following the nested per-window temporal missingness defined in Section 4.1, fixing the features, encoder, query interface, and training schedule. Following prior work, the unimodal references use only the surviving weaker stream, testing whether robust multimodal inference remains preferable (Ramazanova et al., 2025a;b). On Epic-Kitchens, verbs, e.g., *Open*, describe temporally extended interactions, while nouns, e.g., *Bag*, identify manipulated objects from fewer salient frames, and their composition defines actions, e.g, *Open Bag*. On Epic-Sounds, the sound head identifies acoustic events, e.g., *Rustle*. **(1) Missingness training is necessary.** H1 sees only complete observations, leaving its absence embeddings untrained. Under dominant-stream loss, audio may disclose an Epic-Kitchens verb but rarely its visually grounded noun. Similarly, recovering an Epic-Sounds sound from vision requires learning that correspondence. H1 therefore falls from 53.6 to 4.1 on Epic-Kitchens and from 56.5 to 26.3 on Epic-Sounds, respectively 8.9 and 15.1 points below the unimodal references, confirming the fragility of conventional multimodal models (Ramazanova et al., 2025a;b) **(2) Temporal outage exposure improves robustness.** Window-local modality dropout with learned mask tokens makes absence temporally addressable. Since verb cues unfold across an interval, masking encourages H2 to integrate surviving windows rather than depend on one frame. Nouns often rely on fewer salient object-defining frames, while the sound head loses roughly one-third of its acoustic evidence, making clean supervision for them less reliable. H2 consequently raises the 100% endpoints by 12.2 and 20.4 points but lowers clean action and sound accuracy by 0.7 and 1.0 points, reflecting the robustness–clean-accuracy trade-off (Ramazanova et al., 2025a). **(3) Content JEPA restores contextual consistency.** Since global attention propagates an outage beyond the replaced tokens, dense content JEPA aligns all corrupted content representations with clean targets, stabilizing the multimodal context. For action recognition, it restores the noun recognition object cues weakened by dropout, without discarding the temporal integration learned for verbs, although Table 1 does not isolate noun and verb effects separately. Consequently, H3 recovers H1's clean Epic-Kitchens action ceiling and comes within 0.4 points on Epic-Sounds sound, while further improving on the 100% endpoints of H2 by 1.9 and 1.7 points. **(4) Query JEPA aligns what classifiers read.** Content JEPA cannot directly constrain interval-conditioned queries, since content tokens never attend to them. Query JEPA aligns the masked and clean classifier readouts, preserving the verb–noun composition of visual action and the cross-modal evidence supporting sound. H4 improves over H3 by 0.4 and 2.0 points under complete input and by 1.3 and 1.6 points under total dominant-stream loss. Together, dropout simulates the condition for attention to reroute toward surviving evidence, content JEPA stabilizes the multimodal context, and query JEPA preserves the task decision.

4.3 MacJEPA Queries Reroute Attention Around Local Outages

Figure 3 follows four interval queries within one 19.2 s validation observation; each curve shows the fraction of a query's attention assigned to each 2.13 s visual or audio interval, averaged across encoder layers and heads. With complete input, the queries attend to distinct temporal regions, and all four predictions match their ground truth. The corruption removes video over 0–2.1 s and 12.8–17.1 s, and audio over 4.3–6.4 s, 8.5–10.7 s, and 17.1–19.2 s, disrupting both the early *Pour* and later *Drink* queries. Attention responds locally: the early queries shift visual mass from 0–2.1 s to 2.1–4.3 s, while the *Pour* sound query moves its audio peak from 8.5–10.7 s to the surviving 6.4–8.5 s interval. The later queries redistribute visual mass from 12.8–17.1 s to their neighboring intervals, while their audio peaks move from the missing final interval to 14.9–17.1 s for *Drink Water* and 12.8–14.9 s for

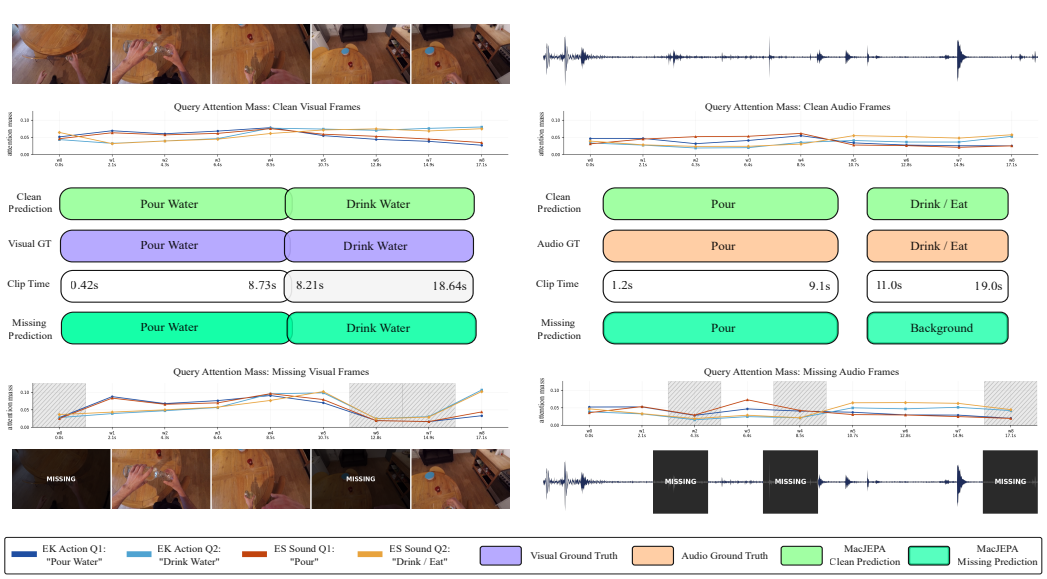

Figure 3: Query attention over visual (left) and audio (right) windows before (above) and after (below) local outages. Hatched regions are missing, boxes compare clean and corrupted predictions against ground truth.

Table 2: Full-modality validation top-1 accuracy (%). MacJEPA uses a single saved checkpoint for both Epic-Kitchens and Epic-Sounds. Best results are bold, and second-best results are underlined.

(a) Epic-Kitchens

Audio-Visual Benchmarks	*xp*	Verb	Noun	Action
TBN	224p	66.0	47.2	36.7
MMT	224p	64.0	57.3	42.8
MBT	224p	64.8	58.0	43.4
MTCN	336p	70.7	62.1	49.6
M&M	420p	72.0	<u>66.3</u>	53.6
TIM	224p	**77.1**	**67.2**	**57.5**
MacJEPA (ours)	224p	<u>74.0</u>	64.5	<u>54.0</u>

(b) Epic-Sounds

Benchmarks	Mod.	Sound
SSAST	A	53.5
ASF	A	53.8
MMT	A,V	55.2
TIM	A	55.7
DIFFSED	A	56.9
TIM	A,V	**58.3**
MacJEPA (ours)	A,V	<u>58.1</u>

Drink/Eat. Under this attention redistribution, both Epic-Kitchens predictions and the Epic-Sounds *Pour* prediction remain unchanged; only *Drink/Eat* changes to *Background* since substantial late multimodal evidence is removed. In this example, we visualize how query attention mass redistributes away from local outages toward nearby surviving evidence, while three of the four task readouts remain unchanged relative to clean input.

4.4 Comparison with Prior Work

4.4.1 Full-Modality Recognition

Table 2 establishes two findings under complete input. **(1.)** On Epic-Kitchens, Table 2(a), MacJEPA achieves 74.0 verb, 64.5 noun, and 54.0 action accuracy using 224p inputs, ranking second on verb and action while outperforming TBN, MBT, MTCN, and M&M on both metrics despite the latter using 420p inputs (Kazakos et al., 2019; Nagrani et al., 2021; Kazakos et al., 2021a; Xiong et al., 2022; Chalk et al., 2024). The MMT row, the only missing-modality framework in this comparison, reports its modal-complete MBT baseline rather than its adapted model (Ramazanova et al., 2025a). **(2.)** On Epic-Sounds, Table 2(b), MacJEPA reaches 58.1 sound accuracy, only 0.2 points below audio-visual TIM, 1.2 points above the strongest audio-only model, DIFFSED, and 2.9 points above the modal-complete MMT baseline (Chalk et al., 2024; Bhosale et al., 2024; Ramazanova et al., 2025a). It also surpasses SSAST, ASF, and audio-only TIM (Gong et al., 2022; Kazakos et al., 2021b; Chalk et al., 2024).

Table 3: Dominant-modality robustness. Top-1 accuracy (%) for Epic-Kitchens verb under missing video and Epic-Sounds sound under missing audio. Follows observation-level missingness protocol.

(a) Epic-Kitchens verb: Video dropped

Benchmarks	Missing rate (%)				
	0	25	50	75	100
Unimodal (A only)	40.0	40.0	40.0	40.0	40.0
MIDL-LTA	63.7	58.4	52.4	46.7	41.4
MMT	63.4	59.0	54.6	50.0	45.3
MacJEPA (ours)	**74.0**	**67.8**	**61.6**	**55.6**	**49.5**

(b) Epic-Sounds sound: Audio dropped

Benchmarks	Missing rate (%)				
	0	25	50	75	100
Unimodal (V only)	41.4	41.4	41.4	41.4	41.4
MIDL-LTA	54.9	46.8	39.5	32.6	26.0
MMT	56.3	52.3	48.6	44.6	40.7
MacJEPA (ours)	**58.1**	**55.9**	**53.9**	**51.8**	**50.0**

Table 4: Auxiliary-modality robustness. Top-1 accuracy (%) for Epic-Kitchens verb under missing audio and Epic-Sounds sound under missing video. Follows observation-level missingness protocol.

(a) Epic-Kitchens verb: Audio dropped

Benchmarks	Missing rate (%)				
	0	25	50	75	100
Unimodal (V only)	63.2	63.2	63.2	63.2	63.2
MIDL-LTA	63.7	61.7	59.5	57.3	55.3
MacJEPA (ours)	**74.0**	**73.5**	**73.2**	**72.9**	**72.5**

(b) Epic-Sounds sound: Video dropped

Benchmarks	Missing rate (%)				
	0	25	50	75	100
Unimodal (A only)	46.5	46.5	46.5	46.5	46.5
MIDL-LTA	55.0	53.4	52.0	51.0	49.4
MacJEPA (ours)	**58.1**	**56.9**	**55.6**	**54.4**	**53.2**

4.4.2 MISSING-MODALITY ROBUSTNESS

Tables 3 and 4 adopt the observation-level missingness of Section 4.1 to align with published benchmarks, which score trimmed single-event clips using separate Epic-Kitchens and Epic-Sounds models. Our MacJEPA model is instead jointly evaluated on both annotation sets within each 19.2 s untrimmed validation video, containing on average 4.55 Epic-Kitchens and 4.04 Epic-Sounds annotations. **Robustness to Dominant Missing Modalities.** Following prior work, the unimodal references are trained only on the auxiliary surviving stream (Ramazanova et al., 2025a;b). Table 3 reveals three findings. **(1)** MacJEPA leads both benchmarks at every missing rate, indicating that its queries can increasingly rely on surviving evidence even at this observation-level missingness protocol. **(2)** At 75% missingness, it exceeds MMT by 5.6 and 7.2 points; on Epic-Kitchens and Epic-Sounds respectively, where MIDL-LTA has already fallen below the 41.4 unimodal reference on Epic-Sounds. **(3)** At the 100% endpoint, MacJEPA leads the strongest prior method by 4.2 points on Epic-Kitchens and 9.3 points on Epic-Sounds. MacJEPA therefore achieves the highest reported absolute accuracy across the dominant-modality missingness sweep, including complete stream loss. **Robustness to Auxiliary Missing Modalities.** Following prior work, our unimodal references use only the dominant surviving stream. (Nagrani et al., 2021; Ramazanova et al., 2025a). Table 4 yields two findings. **(1)** MacJEPA loses only 1.5 points when audio is removed from all observations on Epic-Kitchens and 4.9 points when all video is removed on Epic-Sounds, compared with drops of 8.4 and 5.6 points for MIDL-LTA, respectively (Ramazanova et al., 2025b). **(2)** At 100% missingness, MacJEPA exceeds MIDL-LTA by 17.2 points on Epic-Kitchens and 3.8 points on Epic-Sounds, while remaining 9.3 and 6.7 points above the corresponding dominant-stream unimodal references. Supported by our component and attention analyses, a single fixed MacJEPA checkpoint remains competitive with established audio-visual models under complete input, withstands local outages under our nested per-window temporal missingness protocol, and outperforms prior missing-modality methods under their observation-level missingness protocol across dominant and auxiliary removal, including complete stream loss, without test-time adaptation.

5 CONCLUSION

MacJEPA redefines egocentric missingness as local sensor outages within untrimmed, multi-event observations, with whole-clip absence as the limiting case, and addresses it by introducing typed interval-wise dropout and supervised content-query JEPA. A single checkpoint remains competitive with complete input and leads prior methods at every reported dominant- and auxiliary-stream missing rate on Epic-Kitchens and Epic-Sounds. Future work can extend this framework to temporal action detection under missingness and additional egocentric streams such as optical flow and IMU.

AI USE STATEMENT

In this work, generative AI tools were used only to polish the manuscript's language, including grammar correction, sentence rephrasing, and improvements to readability and clarity. They were not used to generate data, develop theoretical models or conceptual frameworks, formulate mathematical claims, assist with proofs, propose or refine hypotheses, design or evaluate the methodology or experiments, implement methods, translate, clean or reformat data, conduct qualitative or thematic analysis, or interpret results. The authors developed all research concepts and ideas and conducted the analyses. The authors reviewed all AI-assisted text for accuracy, originality, and compliance with ethical standards, and take full responsibility for the final manuscript and all of its claims.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. No new human-subject or animal experimentation was conducted. All datasets were obtained and used in accordance with their respective usage guidelines. We did not seek to identify participants or collect or release additional personal information. We considered potential privacy, bias, and security concerns and remain committed to transparent and responsible reporting throughout the research process.

REPRODUCIBILITY STATEMENT

We provide a code repository containing the MacJEPA scripts, implementation, and reproduction instructions in a zipped code file in our supplementary materials. Section 4.1 specifies the experimental and evaluation protocols, while Section C details feature extraction, model construction, optimization, and implementation choices. Together, these resources document the complete pipeline required to reproduce our results.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. doi: 10.48550/arXiv.2506.09985.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1298–1312, 2022.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1728–1738, 2021.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Swapnil Bhosale, Sauradip Nag, Diptesh Kanojia, Jiankang Deng, and Xiatian Zhu. DiffSED: Sound event detection with denoising diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 792–800, 2024. doi: 10.1609/aaai.v38i2.27837.

Jacob Chalk, Jaesung Huh, Evangelos Kazakos, Andrew Zisserman, and Dima Damen. TIM: A time interval machine for audio-visual action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18153–18163, 2024. doi: 10.1109/CVPR52733.2024.01719.

Delong Chen, Mustafa Shukor, Théo Moutakanni, Willy Chung, Lei Yu, Tejaswi Kasarla, Allen Bolourchi, Yann LeCun, and Pascale Fung. VL-JEPA: Joint embedding predictive architecture for vision-language. In *International Conference on Learning Representations (ICLR)*, 2026.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 721–725, 2020. doi: 10.1109/ICASSP40776.2020.9053174.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision*, 130(1):33–55, 2022. doi: 10.1007/s11263-021-01531-2.

Zhengcong Fei, Mingyuan Fan, and Junshi Huang. A-JEPA: Joint-embedding predictive architecture can listen. *arXiv preprint arXiv:2311.15830*, 2023. doi: 10.48550/arXiv.2311.15830.

Quentin Garrido, Mahmoud Assran, Nicolas Ballas, Adrien Bardes, Laurent Najman, and Yann LeCun. Learning and leveraging world models in visual representation learning. *arXiv preprint arXiv:2403.00504*, 2024. doi: 10.48550/arXiv.2403.00504.

Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16102–16112, 2022. doi: 10.1109/CVPR52688.2022.01563.

Xinyu Gong, Sreyas Mohan, Naina Dhingra, Jean-Charles Bazin, Yilei Li, Zhangyang Wang, and Rakesh Ranjan. MMG-Ego4D: Multi-modal generalization in egocentric action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6481–6491, 2023. doi: 10.1109/CVPR52729.2023.00627.

Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. SSAST: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 10699–10709, 2022. doi: 10.1609/aaai.v36i10.21315.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5842–5850, 2017.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, and Jitendra Malik. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18973–18990, 2022. doi: 10.1109/CVPR52688.2022.01842.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 21271–21284, 2020.

Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6047–6056, 2018.

Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. EPIC-SOUNDS: A large-scale dataset of actions that sound. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10096198.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. EPIC-Fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5491–5500, 2019. doi: 10.1109/ICCV.2019.00559.

Evangelos Kazakos, Jaesung Huh, Arsha Nagrani, Andrew Zisserman, and Dima Damen. With a little help from my temporal context: Multimodal egocentric action recognition. In *British Machine Vision Conference (BMVC)*, 2021a.

Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-fast auditory streams for audio recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 855–859. IEEE, 2021b.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14943–14952, 2023.

Hongyang Lei, Xiaolong Cheng, Qi Qin, Dan Wang, Fan Kun, Huazhen Huang, Qingqing Gu, Yetao Wu, Zhonglin Jiang, Yong Chen, and Luo Ji. M3-JEPA: Multimodal alignment via multi-gate MoE based on the joint-embedding predictive architecture. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, 2025.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18177–18186, 2022.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 14200–14213, 2021.

Gorjan Radevski, Dusan Grujicic, Matthew Blaschko, Marie-Francine Moens, and Tinne Tuytelaars. Multimodal distillation for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5213–5224, 2023. doi: 10.1109/ICCV51070.2023.00481.

Merey Ramazanova, Alejandro Pardo, Humam Alwassel, and Bernard Ghanem. Exploring missing modality in multimodal egocentric datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 75–85, 2025a. doi: 10.1109/CVPRW67362.2025.00013.

Merey Ramazanova, Alejandro Pardo, Bernard Ghanem, and Motasem Alfarra. Test-time adaptation for combating missing modalities in egocentric videos. In *International Conference on Learning Representations (ICLR)*, 2025b.

Benjamin Robson, Santeri Mentu, Wenshuai Zhao, and Arno Solin. AV-JEPA: Extending LeJEPA to audio-visual self-supervised learning. In *ICML 2026 Workshop on Machine Learning for Audio*, 2026. Non-archival.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Maria Santos-Villafranca, Dustin Carrión-Ojeda, Alejandro Perez-Yus, Jesus Bermudez-Cameo, Jose J. Guerrero, and Simone Schaub-Meyer. Multimodal knowledge distillation for egocentric action recognition robust to missing modalities. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2026. doi: 10.48550/arXiv.2504.08578. To appear.

Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 567–576, 2015.

Revant Teotia, Adrien Bardes, Michael Rabbat, Sumit Chopra, Matthew Muckley, and Nicolas Ballas. MJEPA: A simple and scalable joint-embedding predictive architecture for audio-visual learning. *arXiv preprint arXiv:2606.25225*, 2026. doi: 10.48550/arXiv.2606.25225.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 10078–10093, 2022. doi: 10.52202/068431-0732.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual SlowFast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. doi: 10.48550/arXiv.2001.08740.

Xuehan Xiong, Anurag Arnab, Arsha Nagrani, and Cordelia Schmid. M&M mix: A multimodal multiview transformer ensemble. *arXiv preprint arXiv:2206.09852*, 2022. doi: 10.48550/arXiv.2206.09852.

APPENDIX

A MACJEPA FRAMEWORK: EXTENDED DISCUSSION

The interval coordinates, token construction, encoder architecture, attention mask, and task heads follow the formulation in Section 3 and are specified fully in Section C. In brief, each observation is represented by $n = 2T$ time-encoded audio-visual content tokens followed by interval-conditioned verb, noun, action, and sound queries. Content tokens attend globally to content, while each query attends to all content and itself but not to other queries. In this section, we primarily focus on MacJEPA's two central mechanisms: representing temporally local sensor outages and aligning their resulting content and task readouts with clean latent targets. Additionally, we detail the task-specific prediction heads we used, the VICReg loss and the complete learning objective of our framework in this section.

A.1 WINDOW-LOCAL MODALITY DROPOUT AND TYPED, TIMED MASKING

MacJEPA represents availability with a binary mask $\boldsymbol{M} \in \{0,1\}^{K\times|\mathcal{M}|}$, where $M_{k,m} = 1$ means that modality m is unavailable throughout window k. When window-local corruption is applied, each window independently draws one of three states:

$$s_k \sim \text{Categorical}\left(1-\pi,\ \frac{\pi}{2},\ \frac{\pi}{2}\right), \qquad s_k \in \{0, \text{v}, \text{a}\}, \tag{7}$$

with

$$M_{k,m} = \mathbf{1}[s_k = m]. \tag{8}$$

The state $s_k = 0$ preserves both streams, while $s_k = \text{v}$ and $s_k = \text{a}$ remove vision and audio, respectively. Removing both is excluded, ensuring that every window retains contemporaneous evidence from at least one sensor. For either modality,

$$\Pr(M_{k,m} = 1) = \frac{\pi}{2}. \tag{9}$$

Since each window contributes equal numbers of visual and audio tokens, $\pi/2$ is also the expected fraction of substituted tokens in the complete multimodal sequence. Across K windows, the sampler supports

$$3^K \tag{10}$$

patterns of disappearance and recovery, redrawn each epoch. These patterns are not equiprobable because the complete, video-missing, and audio-missing states occur with probabilities $1-\pi$, $\pi/2$, and $\pi/2$, respectively. Whole-observation absence remains within this support when one modality is removed from every window, recovering clip-level missingness as a limiting case.

Let $\boldsymbol{e}^m_\varnothing \in \mathbb{R}^{d_c}$ be the learned absence embedding for modality m. The masked content token is

$$\widetilde{\boldsymbol{u}}^m_{k,f} = \left[(1-M_{k,m})g^m\left(\boldsymbol{x}^m_{kw+f}\right) + M_{k,m}\boldsymbol{e}^m_\varnothing \ \middle\|\ \phi(\mathbf{t}_{kw+f})\right] + \boldsymbol{e}^m. \tag{11}$$

The projected feature is replaced, but its sequence position is not deleted. The token retains its temporal encoding and modality tag. The learned vector $\boldsymbol{e}^m_\varnothing$ identifies which stream is missing, while the complete token $\widetilde{\boldsymbol{u}}^m_{k,f}$ additionally identifies where that absence occurs. Visual and audio use separate absence embeddings. Since corresponding rows share identical temporal coordinates, a common absence vector would give both streams the same substituted content component. Modality-specific vectors make the missing stream explicit before attention is applied.

Windowing also gives the corruption meaningful duration. We choose w such that

$$\tau = w\Delta > \omega. \tag{12}$$

Because adjacent feature supports overlap, removing one row would leave much of its temporal evidence available through neighboring features. Masking w consecutive rows instead creates an outage whose duration exceeds the receptive field of an individual feature and cannot be recovered from feature overlap alone.

Importantly, substitution does not explicitly block attention to missing positions. For query q and content token i, one attention head computes

$$\alpha_{qi} = \mathrm{softmax}_i\left(\frac{(W_Q \boldsymbol{z}_q^q)^\top (W_K \widetilde{\boldsymbol{u}}_i^c)}{\sqrt{d_h}}\right), \qquad \mathbf{o}_q = \sum_{i=1}^{n} \alpha_{qi} W_V \widetilde{\boldsymbol{u}}_i^c, \tag{13}$$

where W_Q, W_K, and W_V are the attention projections and d_h is the head width. A typed, timed absence token remains visible as a key and value, allowing the query to recognize that a particular source is unavailable at that time. Masked-to-clean query alignment then encourages the query to reduce its dependence on unavailable evidence and redistribute attention toward the surviving modality or neighboring temporal context. This rerouting is learned from the robustness objective rather than imposed by an attention mask. The same substitution also changes content self-attention. Every observed content token can attend to absence tokens and surviving evidence across the observation. Dense content JEPA constrains these context-dependent content states toward their clean targets, while query JEPA refines the resulting task-specific attention readout. Mask tokens therefore serve two connected roles: they stabilize content reasoning around an outage and, more directly, enable each query to reroute toward evidence that remains informative. Finally, windows define only the support of corruption. They are flattened before encoding and impose no attention boundaries. Every query retains access to all $n = 2T$ content tokens across the complete $S = T\Delta$ observation.

A.2 EMA Teacher, Predictor, and JEPA

For observation $\mathbf{X} = (\boldsymbol{X}^{\mathrm{v}}, \boldsymbol{X}^{\mathrm{a}})$, query set $\mathcal{Q}$, and sampled mask $\boldsymbol{M}$, the online and target branches compute

$$\begin{aligned} (\boldsymbol{Z}_{\boldsymbol{\theta}}^c, \boldsymbol{Z}_{\boldsymbol{\theta}}^q) &= f_{\boldsymbol{\theta}}(\mathcal{U}(\mathbf{X}, \mathcal{Q}; \boldsymbol{M}))\,, \\ (\bar{\mathbf{Z}}^c, \bar{\mathbf{Z}}^q) &= f_{\bar{\boldsymbol{\theta}}}(\mathcal{U}(\mathbf{X}, \mathcal{Q}; \mathbf{0}))\,. \end{aligned} \tag{14}$$

Both branches receive the same observation and interval queries. The online encoder sees typed, timed absence tokens, whereas the target encoder sees the complete multimodal sequence. The target states therefore describe how the same content and task-specific readouts would have been represented under full evidence. The target encoder remains in evaluation mode, receives no gradient, and follows the online parameters through an exponential moving average:

$$\begin{aligned} \bar{\boldsymbol{\theta}}_t &= \alpha_t \bar{\boldsymbol{\theta}}_{t-1} + (1 - \alpha_t)\boldsymbol{\theta}_t, \\ \alpha_t &= 0.996 + (0.9999 - 0.996)\min\left(\frac{t}{T_{\mathrm{tot}}}, 1\right). \end{aligned} \tag{15}$$

The momentum is scheduled over optimization steps and reaches 0.9999 at the end of training. Model buffers are copied from the online encoder rather than averaged. This EMA and stop-gradient construction provides slowly evolving contextual targets without requiring negative pairs (Grill et al., 2020; Assran et al., 2023; Baevski et al., 2022).

A single token-wise predictor is shared by content and query states:

$$h_{\boldsymbol{\psi}}(\mathbf{z}) = W_2\,\mathrm{GELU}(W_1\,\mathrm{LN}(\mathbf{z}) + \mathbf{b}_1) + \mathbf{b}_2, \qquad d \to 384 \to d. \tag{16}$$

The predictor is a bottleneck rather than an expansion and contains no attention. It cannot gather evidence across tokens after encoding. Cross-modal and temporal recovery must therefore occur within the online encoder, which is the component retained at inference. The predictor also receives no separate mask-location sequence. Its only knowledge of missingness comes through the typed, timed absence states already processed by the encoder. A mask-conditioned or attention-based predictor could learn to invert the corruption itself, allowing the retained encoder to preserve outage-specific representations (Garrido et al., 2024). Limiting the predictor instead encourages stability in the task-conditioned readouts. The content states may still retain information about which sensor failed. Teacher targets are detached and normalized using a parameter-free LayerNorm. For compactness, define

$$d_1(\mathbf{u}, \mathbf{v}) = \frac{1}{d}\|\mathbf{u} - \mathbf{v}\|_1, \qquad \widehat{\mathbf{z}} = \mathrm{LN}(\mathrm{sg}[\bar{\mathbf{z}}])\,.$$

Layer normalization removes target scale as a trivial degree of freedom, while the mean absolute distance follows latent-prediction objectives such as V-JEPA (Assran et al., 2025). Let $v_q \in \{0, 1\}$

identify valid, unpadded query entries. The content and query objectives are

$$\begin{aligned}\ell_{\text{ctx}} &= \frac{1}{n}\sum_{i=1}^{n} d_1\left(h_{\boldsymbol{\psi}}(\boldsymbol{z}^c_{\boldsymbol{\theta},i}), \widehat{\boldsymbol{z}}^c_i\right), \\ \ell_{\text{qry}} &= \frac{\sum_{q=1}^{3N_v+N_a} v_q\, d_1\left(h_{\boldsymbol{\psi}}(\boldsymbol{z}^q_{\boldsymbol{\theta},q}), \widehat{\boldsymbol{z}}^q_q\right)}{\sum_{q=1}^{3N_v+N_a} v_q}, \\ \mathcal{L}_{\text{JEPA}} &= \ell_{\text{ctx}} + \ell_{\text{qry}}.\end{aligned} \tag{17}$$

The content loss is dense over all $n = 2T$ states, rather than restricted to substituted positions. Transformer representations depend on their context, so a local outage can also alter observed tokens elsewhere in the sequence. Dense content JEPA aligns this entire corrupted context with its clean counterpart instead of treating the objective as masked-position reconstruction. The query term aligns the same task and interval under corrupted and clean evidence. Content and query losses are normalized separately before summation, giving the two sets equal aggregate weight despite their different sizes. Since an observation typically contains fewer query states than content states, an individual query consequently receives greater average weight. This emphasis is deliberate because the classifiers operate directly on query representations. Content alignment cannot replace query alignment. The attention mask prevents content tokens from reading queries, and therefore

$$\frac{\partial \ell_{\text{ctx}}}{\partial \boldsymbol{c}_h} = 0, \qquad \forall h \in \mathcal{H}. \tag{18}$$

The content loss still updates the shared encoder, but it has no direct gradient path to the learned query seeds. Query JEPA closes this path by constraining the exact interval-conditioned representation consumed by each classifier. It is therefore the term that directly refines query attention toward evidence that survives an outage.

A.3 Task-Specific Prediction Heads

Each classifier reads only its corresponding query representation. Verb, noun, and sound use independent linear heads

$$C_h(\mathbf{z}) = W_h \,\text{Drop}_{0.3}(\mathbf{z}) + \mathbf{b}_h, \qquad h \in \{\text{verb}, \text{noun}, \text{sound}\}, \tag{19}$$

with 97, 300, and 44 outputs, respectively. For the 3,806 Epic-Kitchens action classes, we use a compositional classifier. If $v(c)$ and $n(c)$ denote the verb and noun associated with action c, its score is

$$S_c(\mathbf{z}) = a_{v(c)}(\mathbf{z}) + b_{n(c)}(\mathbf{z}) + I_c, \qquad c \in \{1, \ldots, 3806\}. \tag{20}$$

Here, $a : \mathbb{R}^d \to \mathbb{R}^{97}$ and $b : \mathbb{R}^d \to \mathbb{R}^{300}$ belong exclusively to the action head. They are not shared with the standalone verb and noun classifiers. The learned residual I_c, initialized to zero, captures compatibility specific to each verb–noun pair. This factorization shares evidence across the long-tailed action vocabulary. Rare actions benefit from observations of their constituent verbs and nouns, while I_c preserves pair-specific information. The four heads otherwise share no parameters and interact only through the common encoder.

A.4 VICReg and Complete Objective

For each head, classification is the mean label-smoothed cross-entropy over its valid queries, with smoothing $\varepsilon = 0.1$. A head with no valid query in a batch contributes an exact zero. The four task losses are combined as

$$\mathcal{L}_{\text{CE}} = \mathcal{L}^{\text{verb}}_{\text{CE}} + \mathcal{L}^{\text{noun}}_{\text{CE}} + w_{\text{act}}\mathcal{L}^{\text{action}}_{\text{CE}} + \lambda_A \mathcal{L}^{\text{sound}}_{\text{CE}}, \qquad \lambda_A = 0.1. \tag{21}$$

The sound weight is explicit because visual and acoustic queries have different counts and label spaces, while the action coefficient controls the contribution of the larger compositional action vocabulary.

We additionally apply VICReg to observed online content states (Bardes et al., 2022). Let $\mathbf{Z} \in \mathbb{R}^{N \times d}$ contain all non-substituted content states in the batch and let $\widetilde{\mathbf{Z}} = \mathbf{Z} - \mathbb{E}[\mathbf{Z}]$. The regularizers are

$$\begin{aligned} \sigma_j &= \sqrt{\mathrm{Var}(\widetilde{\mathbf{Z}}_{:,j}) + 10^{-4}}, \\ \mathcal{L}_{\mathrm{var}} &= \frac{1}{d} \sum_{j=1}^{d} \max(0, 1 - \sigma_j), \\ \mathbf{C} &= \frac{\widetilde{\mathbf{Z}}^\top \widetilde{\mathbf{Z}}}{N-1}, \qquad \mathcal{L}_{\mathrm{cov}} = \frac{1}{d} \sum_{i \neq j} C_{ij}^2. \end{aligned} \tag{22}$$

For compactness, we denote the complete weighted VICReg contribution by

$$w_{\mathrm{vic}} \mathcal{L}_{\mathrm{VIC}} \equiv \mu \mathcal{L}_{\mathrm{var}} + \nu \mathcal{L}_{\mathrm{cov}}. \tag{23}$$

The variance term prevents individual dimensions from collapsing, while the covariance term discourages redundant dimensions. Substituted tokens are excluded because their learned absence embeddings are constant content values. Including them would force these special tokens to reproduce the statistics of genuine sensor observations. Both terms are computed in fp32 for numerical stability. The complete scheduled objective is

$$\boxed{\mathcal{L} = w_{\mathrm{ce}} \mathcal{L}_{\mathrm{CE}} + w_{\mathrm{jepa}} \mathcal{L}_{\mathrm{JEPA}} + w_{\mathrm{vic}} \mathcal{L}_{\mathrm{VIC}}} \tag{24}$$

with $\lambda_A = 0.1$, and $\nu = 1$. The schedules for w_{ce}, w_{act}, w_{jepa}, μ, and the clean-observation probability are specified in Section C.7 and Table 11. All objectives are active from the beginning of training and optimized jointly. MacJEPA never reconstructs raw audio or video. Window-local dropout defines the unavailable evidence, dense content JEPA stabilizes the resulting multimodal context, and query JEPA preserves the representation used for prediction. The EMA teacher and predictor are discarded after training, leaving only the online encoder and task heads at inference.

B Additional Ablations

Beyond the controlled component analysis in Table 1, we examine MacJEPA's sensitivity to its training and architectural choices. Each variant is trained from scratch for 50 epochs and evaluated using its final epoch-50 online checkpoint, with only the factor under study changed. Following the nested per-window temporal missingness protocol defined in Sections C.1 and 4.1, the first four ablations report annotation-level top-1 accuracy at rates $r \in \{0, 0.25, 0.50, 0.75, 1\}$ for Epic-Kitchens action under video removal and Epic-Sounds sound under audio removal, alongside the corresponding unimodal references. We study the training-time modality-drop probability, outage-window duration, observation length, and encoder depth, followed by an Epic-Kitchens action ablation of the frozen visual features. Together, these studies isolate how outage exposure, temporal geometry, model capacity, and input representation shape missing-modality robustness

B.1 Effect of Training-Time Modality-Drop Probability

We isolate the window-level corruption probability

$$\pi = \Pr\left(\sum_{m \in \mathcal{M}} M_{k,m} = 1 \right),$$

where $M_{k,m} = 1$ denotes that modality m is removed from window k. Within observations selected for corruption, a window remains complete with probability $1 - \pi$ and loses either modality with probability $\pi/2$. We vary only $\pi \in \{0.3, 0.5, 0.7, 0.9\}$, keeping the observation length $T = 36$, window size $w = 4$, six-layer encoder, frozen feature backbones, clean-observation schedule, learning objectives, and optimization fixed. MacJEPA uses $\pi = 0.7$. Table 5 exposes a clear robustness–availability trade-off. On Epic-Kitchens, increasing π from 0.3 to 0.9 raises accuracy under complete video loss from 13.7 to 21.5, but reduces clean and 75%-missing accuracy from 54.1 to 53.6 and from

Table 5: Effect of the training-time modality-drop probability π. We report top-1 accuracy (%) across per-window missing rates for Epic-Kitchens action under video removal and Epic-Sounds sound under audio removal. $\pi = 0.7$ is the MacJEPA setting. Follows nested per-window temporal missingness protocol.

Training setting	Epic-Kitchens action, V drop					Epic-Sounds sound, A drop				
	0	25	50	75	100	0	25	50	75	100
Unimodal reference	13.0	13.0	13.0	13.0	13.0	41.4	41.4	41.4	41.4	41.4
$\pi = 0.3$	**54.1**	**53.9**	**53.6**	**49.7**	13.7	57.4	56.7	55.4	53.1	47.8
$\pi = 0.5$	53.9	**53.9**	53.2	48.8	16.9	57.7	**57.2**	**55.9**	**54.1**	**50.1**
$\pi = 0.9$	53.6	53.2	51.8	46.5	**21.5**	**58.1**	**57.2**	55.8	53.6	**50.1**
$\pi = 0.7$ (MacJEPA)	54.0	53.7	52.2	47.3	19.5	**58.1**	56.0	54.3	52.1	50.0

Table 6: Effect of outage granularity w. We report top-1 accuracy (%) across per-window missing rates for Epic-Kitchens action under video removal and Epic-Sounds sound under audio removal. $w = 4$ is the MacJEPA setting. Follows nested per-window temporal missingness protocol.

Window setting	Epic-Kitchens action, V drop					Epic-Sounds sound, A drop				
	0	25	50	75	100	0	25	50	75	100
Unimodal reference	13.0	13.0	13.0	13.0	13.0	41.4	41.4	41.4	41.4	41.4
$w = 2$ (18×1.07 s)	53.6	53.6	**53.1**	**50.9**	14.8	57.7	56.9	**56.1**	**54.1**	48.6
$w = 3$ (12×1.60 s)	53.5	53.5	52.4	48.6	16.9	57.1	56.6	55.7	**54.1**	49.3
$w = 6$ (6×3.20 s)	53.7	53.3	51.4	44.5	**22.3**	57.9	**57.3**	55.9	53.7	**50.0**
$w = 4$ (9×2.13 s; MacJEPA)	**54.0**	**53.7**	52.2	47.3	19.5	**58.1**	56.0	54.3	52.1	**50.0**

49.7 to 46.5. Frequent training outages therefore improve reliance on the surviving audio stream, while excessive corruption weakens the visual evidence available for learning action composition. The effect is milder on Epic-Sounds: complete-loss accuracy rises from 47.8 at $\pi = 0.3$ to 50.1 at $\pi = 0.5$ and then saturates, while clean accuracy remains within 57.4–58.1. We retain $\pi = 0.7$ since it exposes most eligible windows to an outage while preserving enough complete cross-modal evidence to learn fusion, providing a balanced policy across tasks and missing rates rather than optimizing a single evaluation endpoint.

B.2 Effect of Outage Granularity or Window Length

We isolate the number of feature rows w grouped into one outage window. For a fixed observation of $T = 36$ rows, this determines the number and duration of independently masked intervals as

$$K = \frac{T}{w}, \qquad \tau = w\Delta.$$

We vary only $w \in \{2, 3, 4, 6\}$, producing outages of 1.07, 1.60, 2.13, and 3.20 s. The observation set, $\pi = 0.7$, clean-observation schedule, frozen features, six-layer encoder, objectives, and optimization remain fixed. MacJEPA uses $w = 4$, giving nine 2.13 s windows per observation. At fixed T, changing w also changes $K = T/w$ and hence the 3^K possible train-time availability patterns. Fine windows produce more independently located short outages, while coarse windows expose the model to fewer but longer failures. Table 6 shows that outage duration controls the type of robustness learned while leaving clean recognition comparatively stable. Short 1.07 s windows are easiest to bridge using neighboring temporal and cross-modal evidence, giving the strongest intermediate robustness. At 75% video missingness, they reach 50.9 on Epic-Kitchens, compared with 44.5 for 3.20 s windows. Longer training outages instead prepare the model for sustained sensor failure, raising complete-video-loss accuracy from 14.8 to 22.3 as w increases from 2 to 6. The same distinction is weaker but remains visible on Epic-Sounds, where short windows lead at intermediate rates and the coarser settings perform best under complete audio loss. We retain $w = 4$ because its 2.13 s outage exceeds the 1.07 s feature receptive field while preserving nine independently addressable intervals. This setting achieves the strongest clean accuracy on both tasks and balances partial- and complete-outage robustness without optimizing for either dataset or missingness endpoint.

Table 7: Effect of observation length T. We report top-1 accuracy (%) across per-window missing rates for Epic-Kitchens action under video removal and Epic-Sounds sound under audio removal. $T = 36$ is the MacJEPA setting. Follows nested per-window temporal missingness protocol.

Observation setting	Epic-Kitchens action, V drop					Epic-Sounds sound, A drop				
	0	25	50	75	100	0	25	50	75	100
Unimodal reference	13.0	13.0	13.0	13.0	13.0	41.4	41.4	41.4	41.4	41.4
$T = 20$ (10.67 s, $K = 5$)	53.7	53.6	52.7	47.5	**22.0**	57.7	57.0	55.9	53.6	49.5
$T = 28$ (14.93 s, $K = 7$)	53.7	53.6	**52.9**	**47.7**	19.6	**58.1**	**57.6**	**56.1**	**53.9**	49.9
$T = 44$ (23.47 s, $K = 11$)	53.6	53.3	52.5	**47.7**	18.4	56.6	56.3	55.0	53.0	49.2
$T = 36$ (19.20 s, $K = 9$; MacJEPA)	**54.0**	**53.7**	52.2	47.3	19.5	**58.1**	56.0	54.3	52.1	**50.0**

B.3 Effect of Observation Length

We isolate the number of feature rows T in each observation. With feature stride Δ and fixed window size $w = 4$, the observation span and number of outage windows are

$$S = T\Delta, \qquad K = \frac{T}{w}.$$

We vary only $T \in \{20, 28, 36, 44\}$, corresponding to observation spans of 10.67, 14.93, 19.20, and 23.47 s. The window size $w = 4$, $\pi = 0.7$, clean-observation schedule, frozen features, six-layer encoder, objectives, and optimization remain fixed. Observation indices are regenerated for each length, while multi-view averaging preserves annotation-level evaluation with essentially unchanged denominators. MacJEPA uses $T = 36$.

Because w remains fixed, varying T also changes $K = T/w$. At missing rate r, the expected outage therefore covers rK windows, or rS seconds, so longer observations provide both more context and more possible outage locations. Table 7 shows that wider temporal context does not uniformly improve robustness. The 14.93 s observation performs best through most intermediate missing rates, reaching 52.9 and 47.7 on Epic-Kitchens at 50% and 75%, and 57.6, 56.1, and 53.9 on Epic-Sounds from 25% to 75%. This span provides useful neighboring evidence without introducing as much unrelated background or additional events. The shortest observation instead leads under complete video loss on Epic-Kitchens, while extending the context to 23.47 s provides no consistent benefit. Longer context therefore increases available evidence but also makes attention resolve more temporally distant and potentially irrelevant content.

We retain $T = 36$ because its 19.2 s span captures multiple overlapping events and nine independently masked intervals while achieving the strongest clean Epic-Kitchens accuracy, matching the strongest clean Epic-Sounds result, and leading under complete audio loss. Shorter observations win isolated missingness columns, but none provides the same balance between complete-input recognition and robustness across both tasks.

B.4 Effect of Encoder Capacity

We isolate the transformer depth L_{enc} and vary only $L_{\text{enc}} \in \{2, 4, 6, 8\}$. The model width remains 1,024 with eight attention heads and an MLP ratio of two. For every depth, drop-path increases linearly from 0 to 0.1, preserving the same terminal regularization rate. The observation length $T = 36$, window size $w = 4$, $\pi = 0.7$, clean-observation schedule, frozen features, objectives, and optimization remain fixed. MacJEPA uses $L_{\text{enc}} = 6$.

Table 8 shows no monotonic relationship between capacity and missingness robustness. Increasing the model from 21.2M to 71.6M parameters changes clean Epic-Kitchens and Epic-Sounds accuracy by only 0.3 and 0.2 points, respectively. The eight-layer model leads at intermediate Epic-Kitchens missingness and under complete audio loss on Epic-Sounds, whereas the two-layer model is strongest at intermediate Epic-Sounds rates. The six-layer model instead leads clean recognition on both tasks and complete video-loss recognition on Epic-Kitchens. Even the two-layer variant retains approximately 99% of its clean accuracy using only 39% of the parameters, indicating that robustness is produced primarily by the training objective rather than encoder scale. We retain $L_{\text{enc}} = 6$ because it balances capacity and computation while giving the strongest clean performance on both tasks and competitive robustness across the full sweep, rather than optimizing an isolated missingness rate.

Table 8: Effect of encoder depth L_{enc}. We report top-1 accuracy (%) across per-window missing rates for Epic-Kitchens action under video removal and Epic-Sounds sound under audio removal. Total parameter counts are shown in parentheses. Follows nested per-window temporal missingness protocol.

Encoder setting	Epic-Kitchens action, V drop					Epic-Sounds sound, A drop				
	0	25	50	75	100	0	25	50	75	100
Unimodal reference	13.0	13.0	13.0	13.0	13.0	41.4	41.4	41.4	41.4	41.4
$L_{\text{enc}} = 2$ (21.2M)	53.4	53.2	52.1	46.7	18.2	57.6	**57.0**	**56.2**	**54.0**	48.8
$L_{\text{enc}} = 4$ (38.0M)	53.6	53.3	51.9	47.5	19.3	57.2	56.6	55.4	53.6	50.0
$L_{\text{enc}} = 8$ (71.6M)	53.7	53.5	**52.4**	**47.9**	18.3	57.8	**57.0**	56.0	53.8	**50.5**
$L_{\text{enc}} = 6$ (54.8M; MacJEPA)	**54.0**	**53.7**	52.2	47.3	**19.5**	**58.1**	56.0	54.3	52.1	50.0

Table 9: Robustness across frozen visual feature sources. We report Epic-Kitchens action top-1 accuracy (%) under per-window video removal. Omnivore and VideoMAE are concatenated in the MacJEPA setting. Follows nested per-window temporal missingness protocol.

Visual features	Video missing rate (%)				
	0	25	50	75	100
Unimodal reference (A only)	13.0	13.0	13.0	13.0	13.0
Omnivore only ($D_{\text{v}} = 1024$)	52.7	52.4	51.2	45.3	**19.6**
VideoMAE only ($D_{\text{v}} = 1024$)	52.5	51.7	49.6	43.4	18.3
Omnivore + VideoMAE ($D_{\text{v}} = 2048$; MacJEPA)	**54.0**	**53.7**	**52.2**	**47.3**	19.5

B.5 Robustness across Visual Feature Sources

We next isolate the frozen visual representation $\boldsymbol{X}^{\text{v}}$ by ablating the frozen unimodal visual encoder b^v. Let $\mathbf{o}_j$ and $\mathbf{m}_j$ denote the 1,024-dimensional Omnivore Swin-B (Girdhar et al., 2022) and VideoMAE-L (Tong et al., 2022) features at row j. We compare

$$\boldsymbol{x}_j^{\text{v}} \in \{\mathbf{o}_j,\ \mathbf{m}_j,\ [\mathbf{o}_j \| \mathbf{m}_j]\}\,,$$

giving visual dimensions $D_{\text{v}} \in \{1024, 1024, 2048\}$. Only the visual stem input width changes. Auditory SlowFast (Kazakos et al., 2021b) features, $T = 36$, $w = 4$, $\pi = 0.7$, the six-layer encoder, clean-observation schedule, objectives, and optimization remain fixed. MacJEPA concatenates Omnivore and VideoMAE features. Table 9 shows that MacJEPA's robustness is not tied to one visual backbone. Both single-backbone variants remain well above the audio-only reference through most of the sweep, while concatenating their representations improves action accuracy whenever visual evidence remains. It raises clean accuracy by 1.3 points over Omnivore and 1.5 points over VideoMAE, with gains of 2.0 and 3.9 points at 75% video missingness. Under complete video loss, all variants converge within 1.3 points because their visual features are no longer available, and Omnivore's 0.1-point advantage over the combined model is marginal. We therefore retain the concatenated representation because it provides the best balance of clean and partial-outage recognition while preserving near-best complete-loss robustness, rather than selecting a visual backbone from an endpoint at which vision is entirely absent.

C Extended Implementation Details

We provide the evaluation protocol, complete data deatils, feature extraction, sequence construction, architecture, and training used for every reported result. For all experiments, we set $(\Delta, \omega, T, w, d_c, d_t, \delta, \pi) = (16/30\,\text{s}, 32/30\,\text{s}, 36, 4, 768, 256, 0.2\,\text{s}, 0.7)$, giving $(S, K, \tau, n, d) = (19.2\,\text{s}, 9, 2.13\,\text{s}, 72, 1024)$.

C.1 Evaluation Protocol

We evaluate the epoch-50 online model rather than the EMA teacher. Observations advance by one feature row, so each annotation appears in multiple overlapping views. Following Chalk et al. (2024), we average its logits across all views before computing annotation-level top-k accuracy. We use two missingness protocols with identical clean ($r = 0$) and complete-loss ($r = 1$) endpoints.

Nested per-window temporal missingness. Controlled analyses in Sections B and 4.2 use the nested per-window temporal missingness protocol designed to test robustness to local sensor disappearance and recovery. Each observation–window pair receives a fixed pseudorandom rank, thresholded at $r \in \{0, 0.25, 0.50, 0.75, 1\}$, so windows removed at one rate remain removed at higher rates. At $r = 0.75$, the selected modality is absent from 6.75 of nine windows, or 14.4 s, on average. Overlapping observations receive independent ranks, so a physical moment is removed from approximately a fraction r of its views.

Observation-level missingness. For comparison with prior work in Section 4.4.2, we follow the observation-level missingness protocol of MMT and MIDL: a fraction r of validation observations loses the selected modality for the full 19.2 s, while the remainder stay complete (Ramazanova et al., 2025a;b). For fixed MacJEPA, this corresponds to the endpoint mixture $\mathrm{Acc}(r) = (1 - r)\mathrm{Acc}(0) + r\mathrm{Acc}(1)$. Thus, intermediate rates vary outage incidence across observations rather than outage duration within them.

MMT and MIDL classify one annotation from a trimmed 2 s single-event clip and evaluate Epic-Kitchens and Epic-Sounds with separate models, whereas our model jointly queries both annotation sets within each 19.2 s untrimmed observation containing multiple overlapping audio-visual events. Each forward pass contains 8.59 annotations on average, comprising 4.55 Epic-Kitchens and 4.04 Epic-Sounds annotations. Epic-Kitchens targets occupy a mean/median 16.7%/9.8% of the observation, compared with 21.4%/6.4% for Epic-Sounds, whose median event spans only about 1.2 s. Despite fewer queries, Epic-Sounds presents the denser attribution problem: 49.4% temporally overlap another sound event, compared with 40.5% on Epic-Kitchens. Since query-to-query attention is blocked, each query must resolve its target independently from shared content. Longer context aids outage recovery but introduces distractor suppression and multi-instance attribution absent from clip-based evaluation. Thus, TIM matches our observation protocol, while MMT and MIDL match our observation-level missingness protocol. No prior reference matches both.

Missing video and audio are evaluated separately. Video is dominant for Epic-Kitchens and audio for Epic-Sounds, with the reverse defining auxiliary-modality removal. We report annotation-level top-1 accuracy. All Epic-Kitchens heads share one valid-annotation set, and action accuracy uses the compositional head's own argmax rather than the conjunction of separate verb and noun predictions.

C.2 DATASET

Epic-Kitchens-100 contains 100 hours of unscripted, head-mounted recordings from 45 kitchens, comprising 700 untrimmed videos, approximately 20 million frames, and 90,000 fine-grained action segments (Damen et al., 2022). Each action is annotated by its temporal extent, verb, and noun. The label space contains 97 verbs, 300 nouns, and 3,806 verb–noun action classes present across the training and validation sets. Epic-Sounds adds temporally localized acoustic annotations to the same underlying videos (Huh et al., 2023). It contains 78.4K categorized audible events across 44 classes, together with 39.2K non-categorized sound segments. We use only the categorized training and validation annotations. The acoustic intervals and labels are defined independently of the visual actions, so the two annotation streams may differ in duration, overlap, or occur without a corresponding event in the other modality. We use the Epic-Kitchens and Epic-Sounds annotations distributed with TIM, preserving its video-disjoint splits and class numbering (Chalk et al., 2024). We evaluate 9,668 Epic-Kitchens and 8,045 Epic-Sounds validation annotations.

C.3 INDEXING AND OBSERVATION CONSTRUCTION

Each observation contains 36 feature rows and advances by one row, or $16/30$ s, along the untrimmed video. Consecutive observations consequently overlap by 35 rows. An annotation is assigned when its overlap with the observation is at least 0.2 s or when it is fully contained within the observation, following Chalk et al. (2024). Its interval is clipped to the observation and normalized by the 19.2 s observation duration. Training observations contain, on average, 5.48 Epic-Kitchens and 5.28 Epic-Sounds queries; validation observations contain 4.55 and 4.04, respectively. Visual and acoustic queries are padded independently, with validity masks excluding padded entries from classification, query JEPA, and multi-view accumulation. Our feature-complete index contains 487 training and 136 validation videos, producing 485,786 training and 83,964 validation observations.

Table 10: Frozen feature-extraction settings.

	Omnivore Swin-B	VideoMAE-L	Auditory SlowFast R50
Output width	1,024	1,024	2,304
Input support	32 frames	16 frames, tubelet 2	1.0 s waveform
Input size	224×224	224×224	128×200 log-mel
Preprocessing	ImageNet normalization	ImageNet normalization	24-kHz mono audio
Extraction precision	fp16 autocast	fp16 autocast	fp32

C.4 Frozen Feature Extraction

All feature backbones are frozen and run offline. Omnivore Swin-B is pretrained on ImageNet, Kinetics, and SUN RGB-D (Girdhar et al., 2022; Russakovsky et al., 2015; Kay et al., 2017; Song et al., 2015). The VideoMAE-L branch follows VideoMAE/InternVideo pretraining on Kinetics, Something-Something V2, AVA, and WebVid2M (Tong et al., 2022; Wang et al., 2022; Kay et al., 2017; Goyal et al., 2017; Gu et al., 2018; Bain et al., 2021). Both visual extractors are then fine-tuned on Epic-Kitchens-100 labels. Auditory SlowFast R50 is initialized from VGGSound and fine-tuned on Epic-Sounds labels (Kazakos et al., 2021b; Chen et al., 2020; Huh et al., 2023). Further pretraining, fine-tuning, and extraction details are provided in the respective extractor papers (Girdhar et al., 2022; Tong et al., 2022; Wang et al., 2022; Kazakos et al., 2021b). Each visual row concatenates 1,024-dimensional Omnivore and VideoMAE features, forming a 2,048-dimensional input. Omnivore samples 32 frames, resizes the short side to 256 pixels, and takes a 224×224 center crop. VideoMAE samples 16 frames using a 224×224 crop. Both apply ImageNet normalization, and their outputs are concatenated without learned fusion. Audio rows use 2,304-dimensional Auditory SlowFast features. Audio is sampled at 24 kHz, with 24,000 samples per feature and a 12,800-sample hop that matches the $\Delta = 16/30$ s visual-grid stride. The log-mel input contains 128 frequency bins and 200 temporal frames, using a 10-ms analysis window and 5-ms hop. Auditory SlowFast uses a slow–fast ratio of four. All streams are extracted on the same temporal grid. Each row is assigned the encoded support $\omega = 32/30$ s, producing 50% overlap between adjacent rows. Although the physical audio support is 1.0 s, the common encoded extent gives corresponding audio and visual rows identical temporal coordinates. This compact grid represents 19.2 s of context with 72 content tokens. Table 10 summarizes the extraction settings.

C.5 Temporal Encoding and Sequence Assembly

Each 19.2 s observation contains 36 rows grouped into nine four-row windows. Every 2.13 s window contributes four visual and four audio tokens, giving 72 content positions. Content row j uses the normalized interval $[j\Delta, j\Delta + \omega]/S$. Its endpoint is not clipped and the final row reaches $37/36$. Query intervals are clipped to the observation and normalized to $[0, 1]^2$. A shared TimeMLP with widths $2 \to 256 \to 256 \to 256$ uses three ReLU layers followed by LayerNorm. It encodes all content and query intervals and is the model's only positional signal. Visual and audio stems independently apply feature dropout of 0.5, project to 768 dimensions, and apply GELU and LayerNorm. Concatenating the projected feature with its 256-dimensional time encoding produces a 1,024-dimensional token, after which a learned modality embedding is added. Mask substitution occurs after the stem and before temporal concatenation. Each modality has a learned 768-dimensional absence vector initialized from a truncated normal distribution with standard deviation 0.02. Within each window, four visual tokens precede four audio tokens. The flattened sequence is ordered as

$$[72 \text{ content} \mid N_v \text{ verb} \mid N_v \text{ noun} \mid N_v \text{ action} \mid N_a \text{ sound}].$$

Four learned 768-dimensional query seeds, initialized with the same scale, represent verb, noun, action, and sound. Visual-task queries reuse the visual modality embedding, while sound queries reuse the audio embedding.

C.6 Encoder, Attention, and Task Readouts

The encoder contains six pre-normalized transformer blocks of width 1,024, with eight attention heads and a 2,048-dimensional GELU MLP. Attention and MLP dropout are 0.3. Drop-path increases linearly from 0 to 0.1 across the blocks and is applied to both residual branches. A final LayerNorm produces the encoded states, and sequence-level dropout is set to zero. Every sequence row can

Table 11: Training-objective schedule.

Epoch	w_{ce}	w_{act}	w_{jepa}	μ	p_{clean}
1	0.20	0.33	2.20	4.40	0.150
2	0.40	0.67	1.90	3.80	0.163
3	0.60	1.00	1.60	3.20	0.175
4	0.80	1.00	1.30	2.60	0.188
5–50	1.00	1.00	1.00	2.00	0.200

attend to all 72 content columns. Query columns are visible only to themselves. Content therefore attends globally to content but not to queries, while each query attends to all content and itself but not to other queries. Windows are flattened before encoding and impose no attention boundaries. Verb, noun, and sound readouts use dropout of 0.3 followed by independent linear classifiers with 97, 300, and 44 outputs. The compositional action head in Section A.3 predicts 3,806 action classes. The four heads share no parameters. The encoder contains 50.40M parameters, the modality stems 3.35M, the heads 0.87M, and the TimeMLP 0.13M. The complete inference model has 54.76M parameters. Content and query JEPA share one training-only predictor consisting of LayerNorm and a $1024 \to 384 \to 1024$ GELU MLP without attention. It adds 0.79M parameters and is optimized jointly with the online model. The EMA teacher is a non-trainable copy kept in evaluation mode. The predictor and teacher are discarded at inference.

C.7 Mask Sampling and Objective Schedule

Training masks are generated independently for each observation and epoch using the global seed, epoch, video row, and observation start. The probability of retaining an observation entirely clean increases from 0.15 to 0.20 during the first five epochs. In every remaining observation, each window independently drops one uniformly selected modality with probability 0.7. Audio and video are never removed simultaneously. Classification uses label smoothing $\varepsilon = 0.1$ and sound weight $\lambda_A = 0.1$. Content and query JEPA use $\lambda_q = 1$. VICReg is applied only to observed content states and computed in fp32 with variance target $\gamma = 1$, numerical constant 10^{-4}, and covariance weight $\nu = 1$. All objectives are optimized jointly from the first step. Their coefficients reach their final values during the first five epochs, as listed in Table 11. The action-specific ramp limits the early contribution of the 3,806-way action head. There is no separate self-supervised stage or frozen-encoder probe.

C.8 Optimization and Regularization

We train every model for 50 epochs with AdamW using $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 10^{-4}, batch size 256, and gradient clipping at 1.0 over the online model and predictor. The training index contains 1,897 full batches per epoch with `drop_last=True`, giving 94,850 optimization steps. The learning rate warms linearly from 10^{-6} to 3×10^{-4} over the first 3,794 steps, corresponding to two epochs. It then follows cosine decay to 10^{-6} without a constant-rate hold period. Training uses fp16 autocast and gradient scaling, while VICReg is explicitly evaluated in fp32. Feature dropout is 0.5, encoder and head dropout are 0.3, and stochastic depth increases from 0 to 0.1. The EMA momentum increases linearly from 0.996 to 0.9999 over all 94,850 steps. Teacher buffers are copied from the online branch rather than averaged. We use seed 0 and eight persistent data-loading workers with pinned memory and shuffled training observations. Training completed without non-finite batches. All reported full-modality, robustness, component, and additional-ablation results use the final epoch-50 online checkpoint. The same checkpoint evaluates Epic-Kitchens and Epic-Sounds, and no early stopping or task-specific checkpoint selection is used.