# MJEPA: A Simple and Scalable Joint-Embedding Predictive Architecture for Audio-Visual Learning

Revant Teotia[1,2], Adrien Bardes[3], Michael Rabbat[4], Sumit Chopra[2], Matthew Muckley[1⋆], and Nicolas Ballas[4⋆]

[1]FAIR at Meta, New York, USA [2]New York University, New York, USA
[3]FAIR at Meta, Paris, France [4]FAIR at Meta, Montreal, Canada
rt2741@nyu.edu

**Abstract.** Self-supervised learning from large-scale video data has emerged as a dominant paradigm for visual representation learning. Since audio and visual streams naturally co-occur in video data, extending this success to jointly learn from both modalities is a natural next step, yet it remains challenging. Existing audio-visual self-supervised methods rely on modality-specific encoders and complex combinations of contrastive or reconstruction objectives, limiting cross-modal synergy and scalability. Joint Embedding Predictive Architectures (JEPAs) offer a simple, modality-agnostic alternative, but have to date been applied primarily to individual modalities. We introduce MJEPA, a joint-embedding predictive architecture for audio-visual learning that uses a single, unified encoder for both modalities. Our approach uses only a single predictive objective, applied both within and across modalities. We show that cross-modal prediction is critical: without it, a shared encoder degrades below unimodal baselines; with it, each modality's representation benefits from the other. Our frozen ViT-g model outperforms the best prior frozen baseline by over 6.8 mAP on AudioSet-20K, surpasses fully finetuned models on ESC-50 and FSD50K, and is competitive on video benchmarks despite using 10x less video data.

**Keywords:** Self-Supervised Learning · Representation Learning · Multimodal Learning

## 1 Introduction

Learning visual representations from large-scale video data has emerged as a dominant paradigm in recent years, with self-supervised methods demonstrating remarkable success across diverse downstream tasks [5]. Videos, however, are inherently multimodal: the visual stream is naturally accompanied by a rich, complementary audio signal that captures information that is often absent from

⋆ Equal supervision.

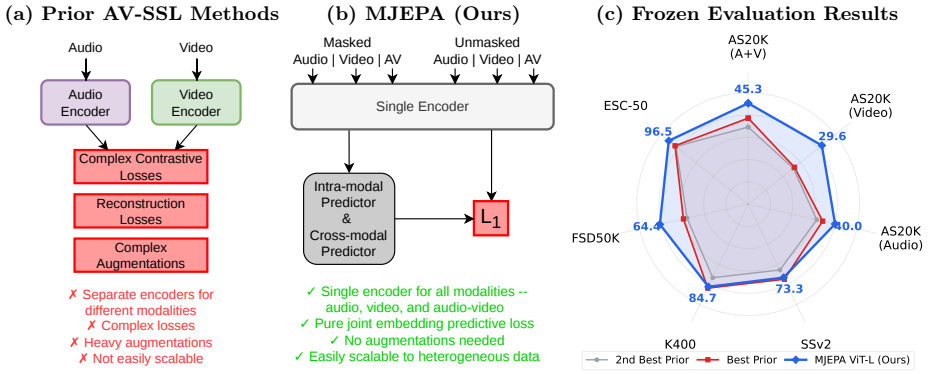


**Fig. 1: Overview of MJEPA.** (a) Prior AV-SSL methods typically combine modality-specific encoders with a mixture of contrastive losses, reconstruction objectives, and heavy augmentations. (b) MJEPA adopts a deliberately simple design: a single shared encoder processes all modalities (audio, video, and audio-video) and is trained with only joint-embedding predictive (JEPA) objective. An intra-modal predictor learns modality-specific features and a cross-modal predictor aligns representations across modalities. This design scales naturally to heterogeneous data, including mixtures of unimodal and multimodal sources. (c) Despite its simplicity, MJEPA produces highly generalizable frozen representations from a single model that match or surpass the best prior frozen features across seven audio, video, and audio-visual benchmarks—outperforming both dedicated unimodal models and specialized multimodal ones.

pixels alone. This observation motivates a natural generalization—extending successful video representation learning to jointly leverage both audio and visual modalities through their intrinsic correspondence in unlabeled video [3].

Approaches exploring audio-visual self-supervised learning (AV-SSL) tend to rely on complex, multi-component architectures: separate encoders for each modality combined with intricate loss functions that blend contrastive objectives [4,27], masked reconstruction [16], or hybrids [19,23,37]. This architectural separation and design complexity potentially limit cross-modal synergy and impede scaling—both in model size and in the use of large-scale unimodal data.

Joint-Embedding Predictive Architectures (JEPAs) [8,13] offer an alternative that is explicitly designed to be modality-agnostic: a unified framework where the same objective—predicting representations of masked regions in a learned embedding space—applies seamlessly across input types [6, 28]. Yet in practice, most existing JEPA methods have only been instantiated for individual modalities (*e.g.*, images, video, audio) leaving the multimodal setting unexplored.

We introduce **MJEPA** (Multimodal Joint-Embedding Predictive Architecture), a multimodal audio-visual self supervised learning framework that extends the JEPA paradigm to jointly learn from audio and video (Fig. 1). Our approach is built on two key design principles: (1) a **single, unified encoder** that processes audio, video, or paired audio-video inputs within a shared representation space, and (2) a **single JEPA objective** that can be applied within and across modalities. Central to our method is the concept of *cross-modal prediction*, which

enables positive transfer across modalities. Beyond standard within-modality JEPA prediction (predicting masked regions from visible context within the same modality), we introduce cross-modal prediction: predicting the representation of one modality from the other. This objective enforces strong semantic alignment between audio and video representations. Indeed, we show that naively sharing an encoder across modalities *without* cross-modal alignment actually degrades both modalities below their unimodal baselines. Cross-modal prediction, however, enables positive transfer across modalities. Combined with scaling to a 1B-parameter ViT-g and heterogeneous data mixtures, we demonstrate that this simple multimodal training learns strong audio-visual representations.

We evaluate MJEPA across audio, video, and audio-visual benchmarks to verify two core hypotheses: (1) a simple, single-encoder SSL approach learns strong audio-visual representation; (2) cross-modal prediction enables positive transfer, where each modality benefits from the other. The resulting frozen representations are competitive with or superior to task-specific finetuned models. On AudioSet-20K [15], our frozen ViT-g model outperforms the best prior frozen baseline by over 6.8 mAP on audio-video classification. On audio benchmarks, our frozen features surpass fully finetuned models on ESC-50 [34] and FSD50K [14]. On video benchmarks Kinetics-400 [26] and SSv2 [20], incorporating audio data allows our model to nearly match state-of-the-art video-based SSL despite using approximately 10× less video training data.

In summary, this paper makes the following contributions:

1. We introduce MJEPA, a simple and scalable AV-SSL framework that trains a single, unified audio-video encoder with only JEPA objectives — without reconstruction, contrastive losses, complex augmentations, or supervision.
2. We demonstrate that cross-modal prediction enables positive transfer across modalities: without it, a shared encoder underperforms unimodal baselines, while with it, each modality's representation improves the other. This synergy allows effective scaling to 1B-parameter models.
3. Extensive evaluation shows that frozen MJEPA features consistently and substantially outperform prior SSL baselines and are competitive with or superior to fully finetuned state-of-the-art models on several key benchmarks.
4. Progressive ablations show the contribution of each design component; joint encoding, shared encoder, cross-modal prediction, and scaling; demonstrating that each is essential for learning strong multimodal representations.

## 2 Related Work

***Self-Supervised Learning for Audio and Video.*** Self-supervised learning has become the dominant paradigm for learning general-purpose representations from unimodal data. In vision, masked image modeling [22,39], joint-embedding methods [33, 38], and joint-embedding predictive architectures (JEPAs) [5, 8] have each produced strong representations, with the latter learning by predicting masked features in a latent space without pixel-level reconstruction. In audio, a parallel trajectory has unfolded: the Audio Spectrogram Transformer [18]

adapted ViTs to spectrograms, followed by self-supervised methods including data2vec [7], BEATs [10], Audio-MAE [24], and A-JEPA [13]. While these unimodal methods learn strong representations, they are trained in isolation and do not leverage the natural correspondence between audio and visual streams.

***Audio-Visual Self-Supervised Learning.*** To exploit the co-occurrence of audio and video, a variety of AV-SSL methods have been proposed [3], typically relying on separate, modality-specific encoders trained with contrastive losses, masked reconstruction, cross-modal distillation, or combinations thereof [1,4,16, 19,23,27,37]. EquiAV [27] relies heavily on data augmentations, using equivariance to encode augmentation-related information into the learned representations. The challenge of moving to a shared encoder has been noted: VATT [1], XKD [37] and AVSiam [30] all observe that unimodal representations degrade when replacing separate encoders with a shared one. In contrast, MJEPA uses a single shared encoder with only joint embedding predictive objective—requiring no data augmentations, negative samples or reconstruction—and produces representations that are strong both unimodally and multimodally, generalizing to diverse out-of-distribution benchmarks without any fine-tuning.

***Joint-Embedding Predictive Architectures and Representation Alignment.*** Our work directly extends the JEPA paradigm [28] to the multimodal domain. In JEPA, models learn abstract representations by predicting masked features in a latent space, a principle that has been successfully applied to images (I-JEPA [6]), video (V-JEPA [8]), and audio (A-JEPA [13]). We provide a detailed formulation of the JEPA objective in Sec. 3. However, most of these these prior instantiations have remained unimodal. MJEPA is the first to propose a simple, unified JEPA with a single shared encoder for audio-visual learning. The motivation for a unified model is supported by the Platonic Representation Hypothesis [25], which posits that representations trained on different modalities converge as models scale. Our cross-modal prediction objective explicitly enforces this alignment, a concept shown to be effective in generative modeling [41]. Other approaches align multiple modalities into a shared space using semantic supervision from image-text or language data [17,42], whereas MJEPA is purely self-supervised. Recent work [29] further shows that one modality can synergize the training of another, providing empirical support for the positive transfer we observe.

## 3 Preliminaries

Our methodology in Sec. 4 is presented as a progressive ablation study where we incrementally build our model and evaluate each component. To support this structure, we first introduce the core concepts and experimental setup that are used throughout the subsequent sections.

### 3.1 Joint-Embedding Predictive Architectures

Unlike generative methods that reconstruct raw inputs (*e.g.*, pixels), Joint-Embedding Predictive Architectures (JEPAs) [28] learn abstract representations by performing a prediction task entirely within a learned embedding space. The core idea is to predict the representation of a target (a complete view of an input) using only the representation of a context (a partial or corrupted view of the same input). This encourages the model to learn semantic features rather than focusing on low-level details.

The typical JEPA setup involves two main components: a context encoder, $E_\theta$, which computes representations for the corrupted input, and a predictor, $P_\phi$, which attempts to predict the target representations from the context representations. A key challenge is preventing "representation collapse", a trivial solution where the encoder outputs a constant value. This is commonly addressed using an asymmetric training setup, such as a stop-gradient (sg) operation on the target branch and using an exponential moving average (EMA) of the encoder weights to produce stable targets [6, 21].

Our work builds on the Video JEPA (V-JEPA) model [5, 8], which applies this paradigm to video by masking out large blocks of spatio-temporal tubelets. The context encoder $E_\theta$ processes the visible tubelets from a masked view $x$, while a target encoder $E_{\bar{\theta}}$ (an EMA of $E_\theta$) processes the complete, unmasked view $y$. The predictor $P_\phi$ then takes the output of the context encoder and a set of learnable mask tokens $\Delta_y$ (which specify the positions of the missing patches) to predict the target representations. The training objective minimizes the L1 distance between the predicted and target representations for the masked tokens:

$$\mathcal{L}_{\text{predict}} = \frac{1}{|M|} \sum_{i \in M} \| P_\phi(E_\theta(x), \Delta_y)_i - \text{sg}(E_{\bar{\theta}}(y))_i \|_1, \tag{1}$$

where $M$ is a set containing the masked patch indexes from the $x$ view. The loss uses a stop-gradient operator, sg, to prevent representation collapse [21] and an exponential moving average $\bar{\theta}$ of $\theta$ is used to update the weight of the target encoder $E_{\bar{\theta}}(y)$ processing the video $y$. $\mathcal{L}_{\text{predict}}$ is only applied on masked tokens, and not on the context tokens. Our goal is to extend this JEPA paradigm to jointly learn from audio and video within a single, unified architecture.

### 3.2 Frozen Evaluation Protocol

To assess the quality of learned representations, we follow the frozen evaluation protocol of [5, 8]. A lightweight attentive probe is trained on top of frozen encoder features for each downstream task. Compared to full fine-tuning, frozen evaluation provides a clearer measure of the intrinsic quality of the learned representations without task-specific adaptation, making it a more reliable assessment of self-supervised learning performance [2]. Moreover, pretrained encoders are often deployed in a frozen state in practice, making this protocol a more realistic measure of a model's general utility across diverse tasks. Architectural details of the probe are provided in the supplementary material.

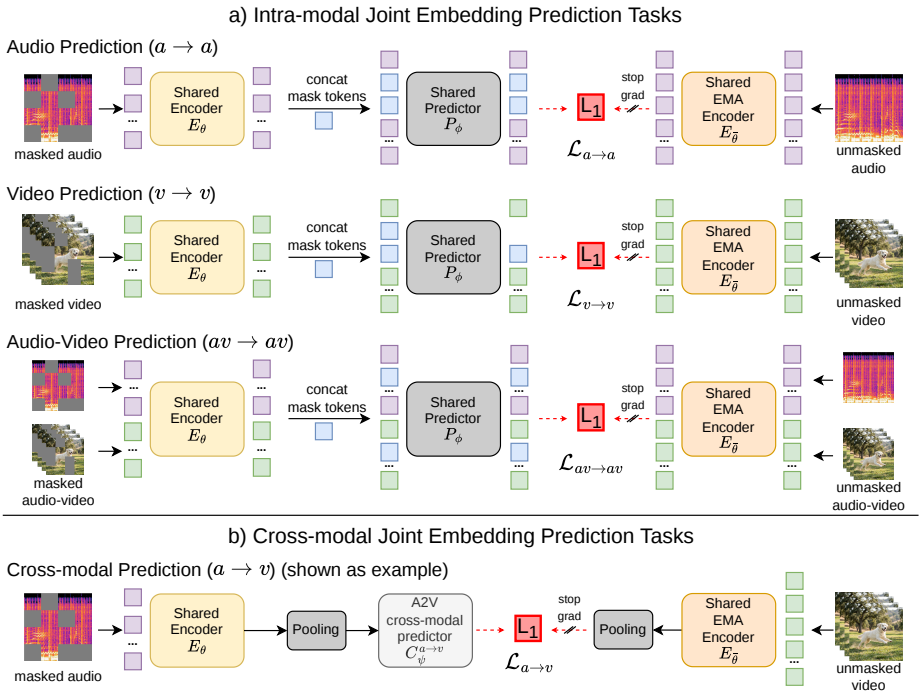


**Fig. 2: MJEPA architecture.** A single shared encoder $E_\theta$ is trained across multiple prediction tasks. **(a) Intra-modal prediction:** the encoder and a shared predictor $P_\phi$ are applied independently to three input modes—audio ($a \rightarrow a$), video ($v \rightarrow v$), and audio-video ($av \rightarrow av$)—each predicting masked token representations against targets from an EMA encoder $E_{\bar{\theta}}$. **(b) Cross-modal prediction:** shown for $a \rightarrow v$: pooled last-layer encoder features from the masked source modality are passed through a lightweight cross-modal predictor $C_\psi^{a\rightarrow v}$ to predict pooled target features of a different modality. Six such objectives ($a \leftrightarrow v$, $a \leftrightarrow av$, $v \leftrightarrow av$) enforce semantic alignment across modalities. All tasks share the same encoder and EMA encoder weights.

### 3.3 Experimental Setting

We conduct our progressive method development on the AudioSet-20K (AS20K) benchmark [15], a class-balanced 22k/19k train/eval split of AudioSet [15]. Our base model is a ViT-L [12] trained on AudioSet2M (AS2M) [15], which contains about 1.8 million 10-second video clips. For *data scaling* experiments, we additionally use the video-only VideoMix2M (VM2M) dataset [8], comprising approximately 2 million videos from HowTo100M [32], Kinetics-710 [26], and Something-Something-v2 [20], and scale our architecture to ViT-g. Further details on datasets and training are provided in the supplementary material.

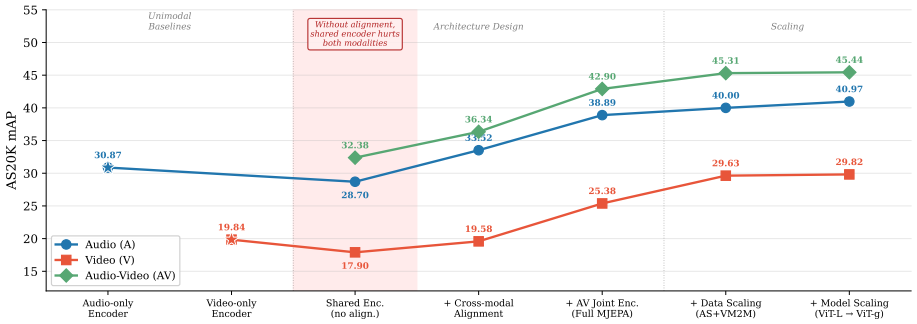


**Fig. 3: Progressive ablation of architecture and scaling on AS20K.** Starting from unimodal baselines, we first incrementally add components of our architecture (left), then show the effects of data and model scaling (right). A shared encoder without cross-modal alignment (shaded) degrades both modalities below their unimodal baselines. Cross-modal alignment, joint audio-video encoding, and scaling are each critical for learning high-quality, generalizable multimodal representations.

## 4 MJEPA: Methodology

We develop our MJEPA framework incrementally to demonstrate a key finding: while a shared encoder is a natural architecture for multimodal learning, it is not effective without an explicit cross-modal alignment objective. We begin by establishing unimodal baselines (Sec. 4.1), then demonstrate the pitfalls of a naively shared encoder (Sec. 4.2). We then show how adding our cross-modal alignment objective resolves this issue (Sec. 4.3), before introducing the full MJEPA model with joint audio-video encoding (Sec. 4.4). Finally, we demonstrate the scalability of our approach with respect to both data and model size (Sec. 4.5). Throughout this development, we report audio-only (A), video-only (V), and audio-video (A-V) mean Average Precision (mAP) on AudioSet-20K [15], a multi-label audio event classification benchmark, using the frozen evaluation protocol from Sec. 3.2. For unimodal baselines, the encoders receive only the corresponding modality's tokens. For the shared encoder variants, the same model produces features for all three settings by receiving audio, video, or concatenated audio-video tokens as input. The progressive results are shown in Fig. 3.

### 4.1 Unimodal Baselines

We first establish unimodal baselines by training two separate, specialized encoders: an audio-only model and a video-only model (see Fig. 3, left). This setup is equivalent to training separate A-JEPA [13] and V-JEPA models [8], and serves as reference points to answer the central question: can learning from audio-visual synergy improve both modalities beyond what each achieves independently? The following input representation and intra-modal prediction objective are used for these baselines and all subsequent experiments.

*Input Representation.* Audio and video are tokenized by separate linear projection layers (a 2D convolution for audio spectrograms and a 3D convolution for video tubelets) that map each modality into the model's embedding space. To distinguish modalities and encode spatial or temporal structure, we add modality-specific embeddings and separate positional embeddings for each modality. Further details on tokenization, embeddings, and masking are provided in the supplementary material.

*Intra-modal Prediction.* The primary training signal is the JEPA objective from Sec. 3, applied independently to each modality, which we refer to as the **intra-modal loss** (see Fig. 2a). For a given input mode $m \in \{a, v, av\}$, let $x^m_{\text{vis}}$ denote the visible (unmasked) tokens of modality $m$. The context encoder produces $z^m_{\text{vis}} = E_\theta(x^m_{\text{vis}})$, while the target encoder computes $y^m_{\text{mask}} = E_{\bar{\theta}}(x^m)_{\text{mask}}$ from the full, unmasked input. The predictor $P_\phi$ takes the encoded visible tokens along with learnable mask tokens $\Delta^m$ and predicts the target representations at the masked positions:

$$\mathcal{L}_{m \to m} = \| P_\phi(z^m_{\text{vis}}, \Delta^m) - \text{sg}(y^m_{\text{mask}}) \|_1 \tag{2}$$

This yields three intra-modal loss terms: $\mathcal{L}_{a \to a}$, $\mathcal{L}_{v \to v}$, and $\mathcal{L}_{av \to av}$. To encourage the learning of a rich feature hierarchy, we use multi-level prediction for the intra-modal objective, where features from multiple intermediate layers of the encoder are fused and predicted. We provide implementation details on this multi-level prediction architecture in the supplementary material.

*Training and Results.* The **Audio-only Encoder** is trained on the audio of AS2M using only the audio intra-modal loss ($\mathcal{L}_{a \to a}$), achieving 30.87 mAP on audio evaluation. Similarly, the **Video-only Encoder** is trained on the video of AS2M using only the video intra-modal loss ($\mathcal{L}_{v \to v}$), achieving 19.84 mAP on video evaluation. The relatively lower video-only score is expected, as AS20K is an audio event classification benchmark where visual information alone may not be sufficient.

### 4.2 Shared Encoder

Since audio and video are different views of the same physical phenomena, their high-level representations should inhabit a common space [25]. A natural architectural choice to encourage this is to use a **single, shared encoder** $E_\theta$ and **shared predictor** $P_\phi$ for both modalities, making parameter sharing an inductive bias toward a unified representation space.

We test this hypothesis by training a single shared encoder-predictor pair on both audio and video from AS2M simultaneously. The model is trained with only the two intra-modal losses ($\mathcal{L}_{a \to a}$ and $\mathcal{L}_{v \to v}$), applied to their respective modalities. Audio and video are encoded independently and never jointly; there is no explicit cross-modal alignment objective.

As shown in Fig. 3 (shaded region), this naive approach is not effective. While the audio-video evaluation benefits from seeing both modalities at probe time

(32.38 mAP), the individual modality performances *drop below* their unimodal baselines: audio falls from 30.87 to 28.70 mAP and video from 19.84 to 17.90 mAP. Without an explicit alignment objective, the shared encoder weights are pulled in two directions — one for each modality, leading to degraded representations for both modalities. This finding is consistent with prior work that also observed unimodal performance degradation when moving to a shared encoder [1, 30, 37] and highlights the central challenge our work addresses.

### 4.3 Shared Encoder + Cross-modal Alignment

Recent work suggests that different modalities naturally converge towards a shared statistical model of reality [25], and that explicitly enforcing this alignment acts as a powerful learning signal that can synergize training across modalities [29, 41]. Motivated by this, we add two cross-modal prediction objectives ($\mathcal{L}_{a\to v}$, $\mathcal{L}_{v\to a}$) to the shared encoder from the previous step, which retains its intra-modal losses ($\mathcal{L}_{a\to a}$, $\mathcal{L}_{v\to v}$). The model is now trained with four losses in total, enforcing both modality-specific coherence and cross-modal semantic alignment. As illustrated in Fig. 2b, we realize this through a set of lightweight cross-modal predictors, $C_\psi^{m_1\to m_2}$, which are simple MLPs that take the mean-pooled, last-layer output of the masked source modality to predict the mean-pooled, last-layer output of the unmasked target modality. The **cross-modal loss** for a source modality $m_1$ and target modality $m_2$ is:

$$\mathcal{L}_{m_1\to m_2} = \|C_\psi^{m_1\to m_2}(\text{pool}(z_{\text{vis}}^{m_1})) - \text{sg}(\text{pool}(y^{m_2}))\|_1 \tag{3}$$

where $z_{\text{vis}}^{m_1} = E_\theta(x_{\text{vis}}^{m_1})$ is the context encoder's *last-layer* output from the *masked* source input, $y^{m_2} = E_{\bar{\theta}}(x^{m_2})$ is the target encoder's *last-layer* output from the *full, unmasked* target input, and pool$(\cdot)$ denotes mean pooling over tokens. In contrast to the multi-level intra-modal prediction, the cross-modal predictors operate only on last-layer features, since low-level features tend to be modality-specific (*e.g.*, spectral patterns in audio, textures in video), whereas high-level features encode semantic concepts shared across modalities and are thus the appropriate target for cross-modal alignment. We use global pooling because there is no natural token-level correspondence between audio and video, making a pooled, holistic representation the appropriate level for cross-modal alignment. Note that audio and video are still encoded independently in this configuration, *i.e.* the model does not process joint audio-video inputs.

As shown in Fig. 3, this single addition of cross-modal alignment resolves the conflict observed in the previous step and substantially improves performance across all three settings: audio climbs to 33.52 mAP (surpassing the unimodal baseline by +2.65), video recovers to 19.58 mAP, and audio-video reaches 36.34 mAP. This confirms that cross-modal alignment is the critical ingredient that enables a shared encoder to benefit from multimodal data.

### 4.4 Full MJEPA: Joint Audio-Video Encoding

Building on the shared encoder with cross-modal alignment from the previous step, we now enable the encoder to also process concatenated audio-video tokens

as a third input mode (see Fig. 2a), adding the joint intra-modal loss ($\mathcal{L}_{av\to av}$) and the remaining four cross-modal losses ($\mathcal{L}_{a\leftrightarrow av}$, $\mathcal{L}_{v\leftrightarrow av}$), for a total of six cross-modal losses: $\mathcal{L}_{a\leftrightarrow v}$, $\mathcal{L}_{a\leftrightarrow av}$, and $\mathcal{L}_{v\leftrightarrow av}$. This allows the encoder to learn from the full audio-visual context during pre-training, with all representations aligned through the cross-modal prediction objectives. As shown in Fig. 3, the result is another substantial boost across the board: audio reaches 38.89 mAP (+8.02 over the unimodal baseline), video reaches 25.38 mAP (+5.54), and audio-video reaches 42.90 mAP. The final training objective is the unweighted sum of all nine loss terms: three intra-modal ($\mathcal{L}_{a\to a}$, $\mathcal{L}_{v\to v}$, $\mathcal{L}_{av\to av}$) and six cross-modal ($\mathcal{L}_{a\leftrightarrow v}$, $\mathcal{L}_{a\leftrightarrow av}$, $\mathcal{L}_{v\leftrightarrow av}$). This is our base **MJEPA ViT-L** model.

### 4.5 Full MJEPA + Scaling

A key advantage of our unified architecture is its ability to efficiently leverage heterogeneous datasets. Because MJEPA uses a shared encoder that operates in audio-only, video-only, and audio-video modes, incorporating additional unimodal data requires no architectural changes—one simply feeds in video-only samples and computes $\mathcal{L}_{v\to v}$. This is naturally enabled by our shared encoder design, whereas prior AV-SSL methods that rely on separate encoders and paired data may require non-trivial modifications to incorporate unimodal sources. In practice, we train on both paired audio-visual and video-only data simultaneously by partitioning compute across the two data sources; implementation details are provided in the supplementary material.

Starting from the base ViT-L trained on AS2M (∼5K hours of paired audio-video data), we augment training with the video-only VM2M dataset [8], adding ∼2M videos (∼136K hours) from standard video benchmarks (+ **Data Scaling**). As shown in Fig. 3, this significantly improves all metrics (mAP): audio from 38.89 to 40.00, video from 25.38 to 29.63, and audio-video from 42.90 to 45.31. Notably, video-only data improves not just video but also audio representations, confirming the positive cross-modal transfer enabled by our alignment objectives. We then scale from ViT-L (300M) to ViT-g (1B parameters) (+ **Model Scaling**), yielding a final boost to 40.97 (A), 29.82 (V), and 45.44 (AV), demonstrating that MJEPA scales effectively with both data and model size.

## 5 Main Results

We now evaluate our full MJEPA models against state-of-the-art self-supervised methods on a comprehensive set of audio, video, and audio-video classification benchmarks. All evaluations use the frozen attentive probe protocol described in Sec. 3.2, providing a fair and consistent comparison across methods. The key takeaway from our results is that MJEPA's simple, unified architecture consistently learns superior frozen representations that outperform prior specialized models, and in several cases, even match or exceed fully finetuned models. We first evaluate audio-video classification on AudioSet-20K (Sec. 5.1), then audio-only performance on AS20K, ESC-50, and FSD50K (Sec. 5.2), and finally video-only performance on Kinetics-400 and SSv2 (Sec. 5.2). We additionally validate

**Table 1: AudioSet-20K audio-visual evaluation.** We report mAP (↑) for audio-only (A), video-only (V), and audio-video (A-V) inputs. All frozen results use the same attentive probing protocol; finetuned results are from respective papers. **Bold**: best frozen; underline: best overall. MJEPA achieves frozen state of the art across all three settings, with frozen video-only results surpassing all finetuned baselines.

| | | | AudioSet20K mAP ↑ | | |
|---|---|---|---|---|---|
| Method | Params | Pre-train data | A | V | A-V |
| *Frozen evaluation* | | | | | |
| CAV-MAE [19] | 170M | IN+AS | 19.38 | 18.14 | 34.59 |
| MAViL [23][a] | 170M | IN+AS | 30.00 | – | – |
| EquiAV [27] | 170M | IN+AS | 34.25 | 18.60 | 38.60 |
| CAV-MAE Sync [4] | 170M | IN+AS | 21.66 | 16.20 | 28.50 |
| MJEPA ViT-L (ours) | 300M | AS | 38.89 | 25.38 | 42.90 |
| MJEPA ViT-L + data scaling (ours) | 300M | AS+VM2M | 40.00 | 29.63 | 45.31 |
| MJEPA ViT-g + data scaling (ours) | 1B | AS+VM2M | **40.97** | **29.82** | **45.44** |
| *Full-model finetuning (reported)* | | | | | |
| CAV-MAE [19] | 170M | IN+AS | 37.70 | 19.80 | 42.00 |
| MAViL [23] | 170M | IN+AS | 41.80 | 24.80 | 44.90 |
| EquiAV [27] | 170M | IN+AS | 42.40 | 25.70 | 46.60 |

[a] MAViL does not release pretrained weights; the paper only reports frozen audio-only result.

the quality of our cross-modal alignment through a retrieval experiment in the supplementary material, providing further evidence that our predictive objectives learns semantically meaningful cross-modal correspondences.

### 5.1 Audio-video frozen evaluation on AudioSet20K.

***Setting.*** We evaluate on AudioSet-20K [15] using the frozen evaluation protocol from Sec. 3.2. For baselines with separate audio and video encoders, we concatenate their output tokens before feeding them to the probe for the A-V setting. To ensure a fair comparison, we evaluate all baselines with our attentive probe, which we found to be consistently stronger than linear probing (*e.g.*, improving CAV-MAE Sync [4] from a reported 8.7 to 21.66 mAP). Notably, all baselines use ImageNet-1K [36] pretraining, whereas MJEPA is trained from scratch.

***Comparison with baselines.*** As shown in Table 1, our method sets a new state of the art for frozen audio-visual representations. Our base MJEPA ViT-L model, trained only on AudioSet, already surpasses the previous best frozen model, EquiAV [27], by a large margin: +4.6 mAP on audio-only, +6.8 on video-only, and +4.3 on audio-video evaluation. Augmenting the pre-training with additional video data (AS+VM2M) further improves performance across all three modalities, demonstrating that the improved video representation from our unified encoder also enhances audio and audio-video features through cross-modal prediction objective. This data-scaled MJEPA ViT-L model produces frozen features that are highly competitive with fully finetuned prior work. Its frozen

**Table 2: Audio-only frozen evaluation.** We report AS20K and FSD50K mAP (↑), and ESC-50 accuracy (↑). **Bold**: best frozen result; underline: best overall. MJEPA sets a new state of the art for frozen audio representations, even surpassing several fully finetuned methods.

| Method | Encoder Params | Modality | Pre-train data | AS20K | ESC-50 | FSD50K |
|---|---|---|---|---|---|---|
| **Frozen feature probing** | | | | | | |
| AJEPA [13] | 85M | A | AS | 18.0 | 74.4 | 43.9 |
| SSLAM [2] | 88M | A | AS | 31.4 | 93.2 | 57.9 |
| SPEAR [40]* | 600M | A | 197k hrs mix | 11.9 | 89.4 | 57.1 |
| Dasheng-1.2B [11]* | 1.2B | A | 272k hrs mix | 31.6 | 92.2 | 56.9 |
| CAV-MAE [19] | 170M | A+V | IN+AS | 19.4 | 77.5 | 46.1 |
| MAViL [23] | 170M | A+V | IN+AS | 30.0 | 90.8 | – |
| XKD [37] | 170M | A+V | AS | – | 93.6 | 51.5 |
| EquiAV [27] | 170M | A+V | AS | 34.3 | 93.2 | 57.9 |
| CAV-MAE Sync [4] | 170M | A+V | AS | 21.7 | 89.2 | 55.5 |
| MJEPA ViT-L (ours) | 300M | A+V | AS | 38.9 | 95.2 | 63.9 |
| MJEPA ViT-L + data scaling (ours) | 300M | A+V | AS+VM2M | 40.0 | 96.8 | 65.5 |
| MJEPA ViT-g + data scaling (ours) | 1B | A+V | AS+VM2M | **40.9** | **96.9** | **65.8** |
| **Full-model finetuning (reported)** | | | | | | |
| AJEPA [13] | 85M | A | AS | 38.4 | 96.3 | – |
| SSLAM [2] | 88M | A | AS | 40.9 | 96.2 | – |
| SPEAR [40]* | 600M | A | 197k hrs mix | 39.4 | – | – |
| CAV-MAE [19] | 170M | A+V | IN+AS | 37.7 | – | – |
| MAViL [23] | 170M | A+V | IN+AS | 41.8 | 94.4 | – |
| XKD [37] | 170M | A+V | AS | – | 96.5 | 58.5 |
| EquiAV [27] | 170M | A+V | AS | 42.4 | 96.0 | 62.6 |

* SPEAR [40] and Dasheng [11] use significantly more pre-training data (197k and 272k hours respectively) compared to the 5k hours of audio in AudioSet. SPEAR also uses Dasheng [11] and WavLM [9] as teachers.

video-only performance (29.63 mAP) surpasses that of all reported finetuned baselines, including EquiAV (25.70 mAP). Finally, by scaling the model size to ViT-g, MJEPA further closes the gap to the overall state of the art, achieving an audio-video mAP of 45.44, within 1.2 points of the fully finetuned EquiAV, while maintaining superior frozen video-only performance. These results demonstrate that a simple JEPA-only objective can produce highly transferable representations that scale effectively with both data and model size.

### 5.2 Audio-only frozen evaluation.

***Setting.*** We evaluate frozen audio representations on three tasks: multi-label sound event classification on AS20K [15] and FSD50K [14], and single-label environmental sound classification on ESC-50 [34], using their standard splits. Notably, FSD50K evaluation labels are considered clean due to an extensive multi-stage human validation process, making it a more reliable benchmark than AS20K. Following Section 3.2, we train an attentive probe over frozen features. To ensure a fair comparison, we re-evaluated all publicly available baselines using our probe, which we found to be significantly stronger than the linear probes reported in many prior works, a finding also supported by recent analysis [35] (*e.g.*, improving SSLAM [2] AS20K mAP from reported 16.9 to 31.4). For mod-

**Table 3: Video-only frozen evaluation.** We report top-1 accuracy (↑) on Kinetics-400 (appearance-focused) and Something-Something v2 (SSv2; motion-focused). **Bold**: best result per column. Despite using 10× less video data than VJEPA2, MJEPA achieves comparable performance by leveraging cross-modal alignment with audio.

| Method | Encoder Params | Pre-train data | K400 | SSv2 |
|---|---|---|---|---|
| VJEPA ViT-L [8] | 300M | VM2M | 80.8 | 69.5 |
| VJEPA ViT-H [8] | 600M | VM2M | 82.0 | 71.4 |
| VJEPA2 ViT-L [5] | 300M | VM22M | 85.1 | 73.7 |
| VJEPA2 ViT-g [5] | 1B | VM22M | **86.6** | **75.3** |
| MJEPA ViT-L (ours) | 300M | AS | 75.2 | 58.6 |
| MJEPA ViT-L (ours)[a] | 300M | VM2M | 80.6 | 69.8 |
| MJEPA ViT-L + data scaling (ours) | 300M | AS+VM2M | 84.7 | 73.3 |
| MJEPA ViT-g + data scaling (ours) | 1B | AS+VM2M | 85.0 | 73.9 |

[a] Trained with video-only intra-modal loss ($\mathcal{L}_{v\to v}$) for comparison with VJEPA.

els without public weights (MAViL [23], XKD [37]) or where our probing yielded lower scores than reported (SPEAR [40] on ESC-50 and FSD50K), we conservatively report their published numbers.

***Results and Analysis.*** As shown in Table 2, MJEPA establishes a new state of the art for frozen representations on general audio tasks. Our base ViT-L model, trained only on AudioSet, already outperforms all prior audio-only and audio-visual SSL methods on AS20K, ESC-50, and FSD50K. This demonstrates that our unified audio-visual architecture learns superior audio features, due to the learned semantic correspondence between modalities by the intra-modal and cross-modal JEPA objectives. This hypothesis is further supported by our data scaling results: augmenting the pre-training with additional video-only data (AS+VM2M) monotonically improves performance on all three audio tasks, demonstrating positive transfer from video to audio through our shared encoder and cross-modal alignment. Notably, the frozen representations from our scaled models are competitive with, and in some cases surpass, fully fine-tuned models. Our frozen MJEPA ViT-g achieves 96.9% accuracy on ESC-50 and 65.8 mAP on FSD50K, outperforming the best finetuned results reported from XKD [37] (96.5%) and EquiAV [27] (62.6%), respectively.

### 5.3 Video-only frozen evaluation.

***Setting.*** We evaluate the quality of the learned frozen video representations on two standard video understanding benchmarks: Kinetics-400 (K400) [26], which focuses on appearance understanding, and Something-Something-v2 (SSv2) [20], a more challenging benchmark for motion understanding. Following the protocol of VJEPA [8] and VJEPA2 [5], we report top-1 accuracy using an attentive probe trained on frozen features with 16 frames, 8 segments × 3 spatial views for K400 and 2 segments × 3 spatial views for SSv2. Our comparison focuses on state-of-the-art video-only SSL methods, as prior AV-SSL methods are typically trained only on AudioSet and not on video-centric datasets.

***Results and Analysis.*** Table 3 highlights the benefit of learning from aligned audio-visual data. We first note that MJEPA trained solely on AudioSet [15], without any video-specific data, already achieves reasonable video recognition performance (75.2% on K400, 58.6% on SSv2), indicating that the audio-visual correspondence in AudioSet provides a useful learning signal for video understanding. When trained on video-only VM2M data using only the $\mathcal{L}_{v\to v}$ objective, our ViT-L model achieves 80.6% on K400 and 69.8% on SSv2, on par with the dedicated video-only VJEPA model (80.8% / 69.5%), demonstrating that our unified architecture does not compromise video representation quality. The true advantage of our approach becomes evident when we combine both data sources. By training on AS+VM2M, the K400 performance of our ViT-L model jumps by over 4 points to 84.7%, significantly surpassing the VM2M-only baseline. This improvement stems from the cross-modal JEPA objective, which aligns representations between modalities, acting as a complementary learning signal that enriches the learned video representations. Consequently, our ViT-L model trained on AS+VM2M nearly matches the performance of VJEPA2 ViT-L on both K400 (84.7% vs. 85.1%) and SSv2 (73.3% vs. 73.7%), despite VJEPA2 being pre-trained on VM22M [5], a dataset with approximately 10× more video data. Scaling to ViT-g further improves performance to 85.0% on K400 and 73.9% on SSv2, underscoring the scalability of our approach.

## 6 Conclusion

In this work, we introduced MJEPA, a simple and scalable framework for audio-visual self-supervised learning. Our approach departs from the prevailing trend of complex, multi-component architectures by demonstrating that a single, unified encoder trained with only a JEPA objective is sufficient to learn powerful representations. We showed that while naively sharing an encoder degrades performance, the addition of a cross-modal prediction objective is the critical ingredient that enables positive transfer, allowing each modality to synergistically improve the other and enabling the framework to scale effectively with both model size and heterogeneous data. The resulting frozen representations set a new state of the art for audio-visual SSL, and their strong performance against fully finetuned models underscores their generality and robustness, a key advantage over more task-specific, adapted representations.

Our work opens promising avenues for future research. The surprising effectiveness of our simple, global-pooled cross-modal predictors suggests that more expressive cross-modal prediction mechanisms could learn even richer alignment across modalities. While this work focuses on audio and video, MJEPA serves as a successful proof-of-concept for a truly modality-agnostic architecture, paving the way for unified models in domains like medical imaging (*e.g.*, CT, MRI, X-ray) or robotics (*e.g.*, vision, proprioception) that can flexibly learn from co-occurring data and are robust to missing modalities at test time.

## References


1. Akbari, H., Yuan, L., Qian, R., Chuang, W.H., Chang, S.F., Cui, Y., Gong, B.: Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 24206–24221. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper_files/paper/2021/file/cb3213ada48302953cb0f166464ab356-Paper.pdf
2. Alex, T., Atito, S., Mustafa, A., Awais, M., Jackson, P.J.B.: SSLAM: Enhancing self-supervised models with audio mixtures for polyphonic soundscapes. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=odU59TxdiB
3. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
4. Araujo, E., Rouditchenko, A., Gong, Y., Bhati, S., Thomas, S., Kingsbury, B., Karlinsky, L., Feris, R., Glass, J.R.: Cav-mae sync: Improving contrastive audio-visual mask autoencoders via fine-grained alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
5. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Komeili, M., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., Arnaud, S., Gejji, A., Martin, A., Robert Hogan, F., Dugas, D., Bojanowski, P., Khalidov, V., Labatut, P., Massa, F., Szafraniec, M., Krishnakumar, K., Li, Y., Ma, X., Chandar, S., Meier, F., LeCun, Y., Rabbat, M., Ballas, N.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 (2025)
6. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15619–15629 (2023)
7. Baevski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M.: data2vec: A general framework for self-supervised learning in speech, vision and language (2022), https://arxiv.org/abs/2202.03555
8. Bardes, A., Garrido, Q., Ponce, J., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: Revisiting feature prediction for learning visual representations from video. arXiv:2404.08471 (2024)
9. Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., Wei, F.: Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing **16**(6), 1505–1518 (Oct 2022). https://doi.org/10.1109/jstsp.2022.3188113, http://dx.doi.org/10.1109/JSTSP.2022.3188113
10. Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., Che, W., Yu, X., Wei, F.: BEATs: Audio pre-training with acoustic tokenizers. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 5178–5193. PMLR (23–29 Jul 2023), https://proceedings.mlr.press/v202/chen23ag.html
11. Dinkel, H., Yan, Z., Wang, Y., Zhang, J., Wang, Y., Wang, B.: Scaling up masked audio encoder learning for general audio classification. In: Interspeech 2024 (2024)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.:

An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
13. Fei, Z., Fan, M., Huang, J.: A-jepa: Joint-embedding predictive architecture can listen (2024), https://arxiv.org/abs/2311.15830
14. Fonseca, E., Favory, X., Pons, J., Font, F., Serra, X.: FSD50K: an open dataset of human-labeled sound events. IEEE/ACM Transactions on Audio, Speech, and Language Processing **30**, 829–852 (2022)
15. Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776–780 (2017). https://doi.org/10.1109/ICASSP.2017.7952261
16. Georgescu, M.I., Fonseca, E., Ionescu, R.T., Lucic, M., Schmid, C., Arnab, A.: Audio-Visual Masked Autoencoders. In: ICCV (2023)
17. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: CVPR (2023)
18. Gong, Y., Chung, Y.A., Glass, J.: AST: Audio Spectrogram Transformer. In: Proc. Interspeech 2021. pp. 571–575 (2021). https://doi.org/10.21437/Interspeech.2021-698
19. Gong, Y., Rouditchenko, A., Liu, A.H., Harwath, D., Karlinsky, L., Kuehne, H., Glass, J.R.: Contrastive audio-visual masked autoencoder. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=QPtMRyk5rb
20. Goyal, R., Kahou, S.E., Michalski, V., Materzyńska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., Memisevic, R.: The "something something" video database for learning and evaluating visual common sense (2017), https://arxiv.org/abs/1706.04261
21. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., et al.: Bootstrap your own latent: A new approach to self-supervised learning. In: Advances in Neural Information Processing Systems. vol. 33, pp. 21271–21284 (2020)
22. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377 (2021)
23. Huang, P.Y., Sharma, V., Xu, H., Ryali, C., Fan, H., Li, Y., Li, S.W., Ghosh, G., Malik, J., Feichtenhofer, C.: MAVil: Masked audio-video learners. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=OmTMaTbjac
24. Huang, P.Y., Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., Metze, F., Feichtenhofer, C.: Masked autoencoders that listen. In: NeurIPS (2022)
25. Huh, M., Cheung, B., Wang, T., Isola, P.: Position: The platonic representation hypothesis. In: Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (eds.) Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 20617–20642. PMLR (21–27 Jul 2024), https://proceedings.mlr.press/v235/huh24a.html
26. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The kinetics human action video dataset (2017), https://arxiv.org/abs/1705.06950
27. Kim, J., Lee, H., Rho, K., Kim, J., Chung, J.S.: EquiAV: Leveraging equivariance for audio-visual contrastive learning. In: Salakhutdinov, R., Kolter, Z.,

Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (eds.) Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 24327–24341. PMLR (21–27 Jul 2024), https://proceedings.mlr.press/v235/kim24v.html

28. LeCun, Y., et al.: A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. Open Review **62**(1), 1–62 (2022)
29. Lee, J.J., Yoon, S.W.: Can one modality model synergize training of other modality models? In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=5BXWhVbHAK
30. Lin, Y.B., Bertasius, G.: Siamese vision transformers are scalable audio-visual learners. arXiv preprint arXiv:2403.19638 (2024)
31. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019), https://arxiv.org/abs/1711.05101
32. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In: ICCV (2019)
33. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=a68SUt6zFt, featured Certification
34. Piczak, K.J.: ESC: Dataset for Environmental Sound Classification. In: Proceedings of the 23rd Annual ACM Conference on Multimedia. pp. 1015–1018. ACM Press (2015). https://doi.org/10.1145/2733373.2806390, http://dl.acm.org/citation.cfm?doid=2733373.2806390
35. Rauch, L., Heinrich, R., Ghaffari, H., Miklautz, L., Moummad, I., Sick, B., Scholz, C.: Unmute the patch tokens: Rethinking probing in multi-label audio classification. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=FbY5Co2NWk
36. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge (2015), https://arxiv.org/abs/1409.0575
37. Sarkar, P., Etemad, A.: Xkd: Cross-modal knowledge distillation with domain alignment for video representation learning (2022)
38. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: Dinov3 (2025), https://arxiv.org/abs/2508.10104
39. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: NeurIPS (2022)
40. Yang, X., Yang, Y., Jin, Z., Cui, Z., Wu, W., Li, B., Zhang, C., Woodland, P.: Spear: A unified ssl framework for learning speech and audio representations (2026), https://arxiv.org/abs/2510.25955
41. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. In: International Conference on Learning Representations (2025)

42. Zhu, B., Lin, B., Ning, M., YAN, Y., Cui, J., HongFa, W., Pang, Y., Jiang, W., Zhang, J., Li, Z., Zhang, C., Li, Z., Liu, W., Li, Y.: Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In: Kim, B., Yue, Y., Chaudhuri, S., Fragkiadaki, K., Khan, M., Sun, Y. (eds.) International Conference on Learning Representations. vol. 2024, pp. 9588–9608 (2024), https://proceedings.iclr.cc/paper_files/paper/2024/file/2862ccf01e3843c81623b246895bcc45-Paper-Conference.pdf

## Supplementary Material

This supplementary material provides model and implementation details (Sec. 7), training details (Sec. 8), and an additional cross-modal retrieval evaluation (Sec. 9).

## 7 Model and Implementation Details

Fig. 4 provides a detailed view of the MJEPA architecture, complementing the high-level overview in Fig. 2 of the main paper. We describe the components in detail below.

*Encoder Architecture.* The MJEPA encoder follows the architecture of the vision transformer [12], and we propose two sizes: ViT-L (300M) and ViT-g (1B).

*Input Tokenization and Embeddings.* As shown in Fig. 4, inputs are tokenized by modality-specific layers—a 2D convolution for audio spectrograms and a 3D convolution for video tubelets—that project patches or tubelets into the shared latent embedding space. To inform the shared encoder of the input structure, we add modality-specific learnable embeddings and positional embeddings. We use 2D sincos absolute positional embeddings for audio tokens (time, frequency) and 3D sincos for video tokens (time, height, width).

*Predictor Architectures.* The shared intra-modal predictor is a narrow ViT (384 embedding dimension) that receives multi-level features from the encoder. Intermediate features from layers {5, 11, 17, 23} for ViT-L, and {9, 19, 29, 39} for ViT-g, are concatenated along the embedding dimension and fused via a 2-layer MLP before being passed to the predictor. The predictor then adds its own set of modality-specific positional and modality embeddings before processing the sequence. These positional embedding are same as encoder positional embedding - 2D sincos absolute positional embeddings for audio tokens and 3D sincos for video tokens. In contrast, the six cross-modal predictors are simple 3-layer MLPs that operate only on the mean-pooled, last-layer features of the encoder.

*Attentive Probe Architecture.* Our attentive probe consists of a learnable query token that attends to all encoder output tokens via cross-attention, followed by 4 transformer blocks and a linear classification head.

## 8 Training Details

*Optimizer and Precision.* All models are trained with the AdamW optimizer [31] with $\beta_1 = 0.9$ and $\beta_2 = 0.995$. We use bfloat16 precision for all training and evaluation.

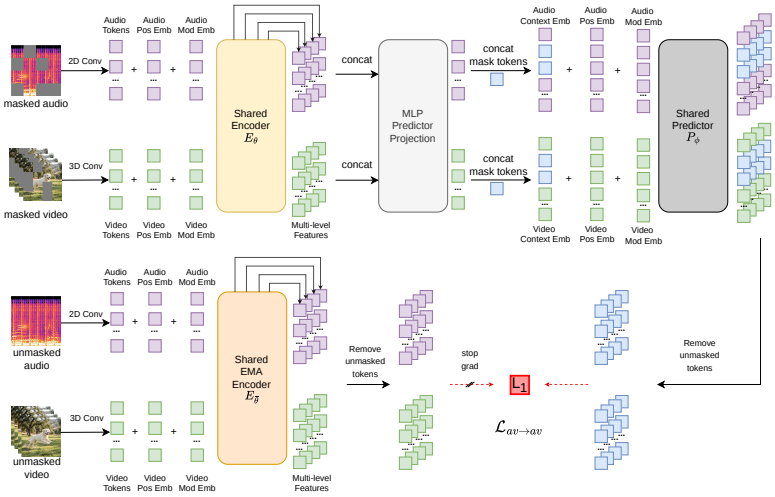


**Fig. 4: Detailed architecture of MJEPA**, illustrated for the joint audio-video intra-modal prediction ($\mathcal{L}_{av\to av}$). The audio-only ($\mathcal{L}_{a\to a}$) and video-only ($\mathcal{L}_{v\to v}$) cases follow the same pipeline with a single modality branch. **Top-left:** Masked audio and video inputs are tokenized by modality-specific projection layers (2D Conv for audio, 3D Conv for video), and augmented with modality-specific positional and modality embeddings before being fed to the shared context encoder $E_\theta$. **Top-middle:** Multi-level features from intermediate encoder layers are concatenated along the embedding dimension and projected back to the model dimension via a MLP. **Top-right:** The projected context tokens are concatenated with learnable mask tokens, added with predictor-specific modality embeddings and positional embeddings, and processed by the shared predictor $P_\phi$ to produce multi-level predictions for the masked positions. **Bottom:** The target branch processes the full, unmasked inputs through the EMA encoder $E_{\bar{\theta}}$ and extracts multi-level features at the masked positions. The $L_1$ loss is computed between the predicted and target representations with a stop-gradient on the target. Cross-modal predictors are not shown; they operate on pooled last-layer features as described in Sec. 4.3 of the main paper.

*Preprocessing and Augmentation.* We follow standard data preprocessing practices. For audio, we adopt the pipeline from AST [18], converting each 10-second clip into a $1024 \times 128$ log-mel spectrogram, which is then normalized using a mean of -4.268 and a standard deviation of 4.569. For video, we use random resized cropping with a scale of [0.3, 1.0] and an aspect ratio of [0.75, 1.35]. The crop size is 224 for the base ViT-L model and 256 for the scaled models.

*Masking Strategy.* We use modality-specific masking. For audio, we randomly mask 75-80% of patches. For video, our base ViT-L model uses a single masking strategy of 8 small blocks (15% of the frame) that span the full temporal dimension. For the scaled models (ViT-L and ViT-g), we adopt the more aggressive multi-block strategy from VJEPA2 [5], applying a combination of eight smaller masks (15% of the frame) and two larger masks (70% of the frame) per

**Table 4: Key training hyperparameters for MJEPA models.** AS2M denotes the audio-visual AudioSet-2M dataset, while VM2M denotes the video-only VideoMix2M dataset. The scaled models are pre-trained with a constant learning rate and then undergo a cooldown phase with an increased frame count and a linear learning rate decay.

| Hyperparameter | Base ViT-L | Scaled ViT-L | Scaled ViT-g |
|---|---|---|---|
| Datasets | AS2M | AS2M + VM2M | AS2M + VM2M |
| Total Batch Size | 512 | 2560 (512 AV + 2048 V) | 2880 (576 AV + 2304 V) |
| Crop Size | 224 | 256 | 256 |
| LR Schedule | Cosine decay (6e-4 to 1e-6) | Constant (5.25e-4) | Constant (5.25e-4) |
| EMA Schedule | Cosine decay (0.999 to 1.0) | Constant (0.999) | Constant (0.999) |
| Pre-training Steps | 150k | 150k | 150k |
| Cooldown Steps | - | 12k | 12k |
| Cooldown Frames | - | 64 | 64 |

iteration. For combined audio-video inputs, masking is applied independently to each stream.

*Training Schedule.* The base ViT-L model is trained on AS2M for 150k steps with a batch size of 512. It uses a cosine learning rate decay from 6e-4 to 1e-6 with a 60-epoch warmup, and a cosine EMA schedule from 0.999 to 1.0. The scaled models (ViT-L and ViT-g on AS2M+VM2M) are pre-trained for 150k steps with a constant learning rate of 5.25e-4 and a fixed EMA of 0.999, both with a 40-epoch warmup. This is followed by a 12k-step cooldown phase where the input is increased to 64 frames and the learning rate is linearly decayed to zero.

*Distributed Training for Data Scaling.* When training with additional video-only data, we partition the workload across GPUs: a subset of GPUs processes paired audio-visual samples, while the remaining GPUs process video-only samples, computing only $\mathcal{L}_{v \to v}$. Gradients from all GPUs are combined via all-reduce before a single optimizer step. The sampling weights for the video-only datasets (Kinetics-710, SSv2, HowTo100M) follow those used in V-JEPA [8]. A scaling factor of 5.0, determined via hyperparameter search, is applied to the video-only loss to balance gradient magnitudes across the two data sources.

## 9 Cross-modal Retrieval Evaluation

To further validate the quality of the cross-modal alignment learned by MJEPA, we evaluate the cross-modal predictors on a retrieval task. Unlike prior methods that rely on explicit contrastive objectives designed for retrieval, MJEPA is trained with a pure predictive objective.

Given a set of audio-video pairs, we first extract audio and video features independently using our shared MJEPA encoder — audio tokens are fed to obtain

**Table 5: Cross-modal retrieval on AudioSet.** We report recall at $K$ (R@1/5/10; ↑) for audio→video and video→audio retrieval. Despite using no contrastive objective, MJEPA achieves retrieval performance competitive with state-of-the-art contrastive methods, validating the quality of our cross-modal alignment.

| | Audio→Video R@K ↑ | | | Video→Audio R@K ↑ | | |
|---|---|---|---|---|---|---|
| Method | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| CAV-MAE [19] | 15.1 | 34.0 | 43.0 | 18.8 | 39.5 | 50.1 |
| CAV-MAE Sync [4] | 27.9 | 52.4 | 62.2 | **35.2** | **58.3** | **67.6** |
| EquiAV [27] | **29.6** | **53.7** | **63.1** | 30.1 | 53.3 | 62.9 |
| MJEPA (ours) | 29.3 | **53.7** | 61.8 | 30.6 | 53.4 | 61.3 |

audio features, and video tokens are fed to obtain video features. For audio-to-video retrieval, we pass the audio features through the cross-modal predictor $C_\psi^{a\to v}$ to generate a predicted video representation, and retrieve the nearest ground-truth video representation from the pool using L1 distance, consistent with our training objective. The same procedure applies for video-to-audio retrieval using $C_\psi^{v\to a}$. We compute distances on pre-final layer norm features. Baselines typically use cosine similarity, which is natural for their contrastive losses. Following prior work [4, 19, 27], we evaluate on the AudioSet retrieval subset.

As shown in Table 5, MJEPA, without any contrastive objective, achieves retrieval performance that is highly competitive with state-of-the-art contrastive methods such as EquiAV [27]. This is a notable finding, as prior work has shown that removing the contrastive objective can cause retrieval performance to collapse to random chance [19]. While methods like CAV-MAE Sync [4], which use explicit temporal alignment and a contrastive loss, show stronger video-to-audio performance, our results demonstrate that the JEPA predictive objective alone is sufficient to learn strong semantic alignment between modalities.