LeAVJEPA: A Minimalist Architecture for Audio-Visual Self-Supervised Learning

Anonymous authors
Paper under double-blind review

Abstract

Prior audio-visual self-supervised learning methods rely on mechanisms such as EMA target encoders, prediction heads, reconstruction decoders, and contrastive losses. We introduce LeAVJEPA, the first audio-visual encoder trained under LeJEPA's collapse-free objective. A single early-fusion Vision Transformer processes audio, video, and joint audio-video inputs. Modality dropout treats a missing modality as another view of the same event, making cross-modal alignment implicit in the objective. The model aligns global embeddings with modality-specific local embeddings, and SIGReg prevents representational collapse. A controlled ablation identifies modality dropout as the key mechanism for audio-visual alignment. Despite the architectural simplicity, LeAVJEPA reaches 36.0 mAP on AudioSet-20K and 91.3% accuracy on ESC-50 under frozen evaluation. After fine-tuning, it reaches 61.1% accuracy on VGGSound, and its embeddings support zero-shot audio-visual retrieval.

1 Introduction

Self-supervised learning (SSL) has become the dominant paradigm for learning representations from unlabelled data in vision (He et al., 2022; Oquab et al., 2024), audio (Huang et al., 2022), and video (Bardes et al., 2024). Video is a natural source of paired multimodal data: every clip pairs a visual stream with an audio stream that records the same event, and their co-occurrence provides supervision at no annotation cost (Arandjelović & Zisserman, 2017; Korbar et al., 2018). The pairing is also imperfect, since sound sources move off-screen and visible objects are often silent. A good audio-visual representation should therefore draw on both modalities, yet remain useful when either one is missing or uninformative.

Existing audio-visual SSL methods combine audio and video through several different training objectives. AV-MAE (Georgescu et al., 2023), CAV-MAE (Gong et al., 2023), MAViL (Huang et al., 2023), and CAV-MAE Sync (Araujo et al., 2025) reconstruct masked spectrogram and pixel patches through decoders, most add InfoNCE terms over cross-modal negatives, and MAViL adds a second self-training stage. Alternatively, predicting in latent space focuses learning on represen-

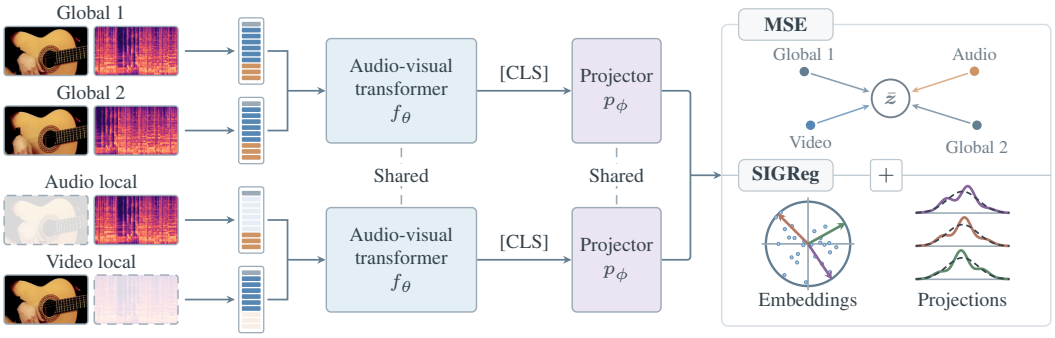


Figure 1: **LeAVJEPA connects sound and vision by dropping a modality.** One shared encoder aligns audio-only and video-only views with joint audio-video views. MSE brings embeddings together; SIGReg prevents collapse, without a teacher, predictor, or contrastive negatives.

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.

tations and avoids reconstructing low-level input detail (Assran et al., 2023; Bardes et al., 2024). MJEPA (Teotia et al., 2026) takes this route with a shared encoder, but retains an EMA target encoder and a predictor network. This raises the question of *how many of these components are necessary*. LeJEPA (Balestriero & LeCun, 2025) showed for single-modality vision that a symmetric view-invariance loss plus SIGReg, a regularizer that drives embeddings toward an isotropic Gaussian, trains stably with no teacher, predictor, stop-gradient, or negatives. However, whether the same holds for audio and video is unclear. When every view contains both modalities, the encoder can minimise the invariance loss using only the dominant modality and largely ignore the other. Learning a cross-modal representation therefore depends on views that separate the modalities, such that audio-only and video-only views of the same clip are mapped to a common embedding.

We introduce **LeAVJEPA**, a minimalist audio-visual JEPA that trains a single shared early-fusion Vision Transformer with only an invariance loss and SIGReg, treating paired audio and video as partial observations of the same event. The encoder maps each audio, video, or joint input to a sequence of tokens (Fig. 1), where global views contain both modalities and *modality dropout* makes local views audio-only or video-only. The invariance loss pulls each local-view embedding toward the centre of the global views, requiring the encoder to infer the joint representation from either modality alone. The views determine which modalities the encoder observes, while the common centre defines the representation to be predicted. Cross-modal prediction thus emerges implicitly in latent space, without a teacher, predictor, decoder, or negatives.

On VGGSound and AudioSet, LeAVJEPA reaches **61.1%** accuracy after fine-tuning and **36.0 mAP** on AudioSet-20K with a frozen attentive probe, and its embeddings support audio↔video retrieval without retrieval-specific training. These results show that the shared-encoder approach is effective, although the strongest published methods still perform better. A controlled ablation identifies modality-specific views as central to representation quality. Masked local views nearly halve linear-probe accuracy, and joint local views reduce it to the level of pretraining on audio alone.

Together, these findings support three contributions:

- **A minimalist audio-visual JEPA** that trains a single shared early-fusion encoder with a symmetric invariance loss and SIGReg, with no reconstruction decoder, EMA teacher, predictor, contrastive negatives, or multi-stage schedule (Table 1).
- **Modality dropout for partial-view learning**, which makes cross-modal prediction an implicit property of the invariance objective rather than a separate loss or network (Sec. 3.1).
- **An empirical study** that first isolates the contribution of view construction through a controlled ablation, showing that modality-specific local views substantially outperform masked, joint, audio-only, and no-local-view alternatives (Sec. 4.1), and then evaluates frozen transfer, zero-shot retrieval, and end-to-end fine-tuning with ViT-B and ViT-L backbones on VGGSound and AudioSet-2M (Sec. 4).

2 Related Work

Audio-visual self-supervised learning Early work used audio-visual correspondence (Arandjelović & Zisserman, 2017) and temporal synchrony (Korbar et al., 2018) as pretext tasks; AVID-CMA (Morgado et al., 2021), XDC (Alwassel et al., 2020), and VATT (Akbari et al., 2021) then aligned the two modalities with InfoNCE (van den Oord et al., 2018) or clustering objectives. A second line masks and reconstructs raw inputs: AV-MAE (Georgescu et al., 2023), CAV-MAE (Gong et al., 2023), MAViL (Huang et al., 2023), and CAV-MAE Sync (Araujo et al., 2025) extend MAE (He et al., 2022) to paired audio and video, in most cases with an added contrastive term. These approaches establish reconstruction and explicit alignment as effective learning signals. We study how far a common-centre objective can go when modality dropout supplies the partial observations.

Audio-only and video-only SSL backbones Single-modality methods establish how to turn sound and video into useful representations. wav2vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021) learn from waveforms; AST (Gong et al., 2021), SSAST (Gong et al., 2022), AudioMAE (Huang et al., 2022), and BEATs (Chen et al., 2023) operate on spectrograms. ViViT (Arnab et al., 2021) and TimeSformer (Bertasius et al., 2021) develop transformer architectures for video. Our architecture uses modality-specific tokenization to bring both streams into one transformer.

Table 1: Comparison of audio-visual SSL methods. LeAVJEPA is the only method that requires none of the listed components.

Method	Single shared encoder	No raw-input recon.	No contr. negatives	No decoder	No teacher / EMA	No predictor	Training objective
AV-MAE (Georgescu et al., 2023)	✗	✗	✓	✗	✓	✓	Recon.
CAV-MAE (Gong et al., 2023)	✗	✗	✗	✗	✓	✓	Recon. + contr.
MAViL (Huang et al., 2023)	✗	✗	✗	✗	✗	✓	Raw/latent recon. + contr.
CAV-MAE Sync (Araujo et al., 2025)	✗	✗	✗	✗	✓	✓	Recon. + contr.
MJEPA (Teotia et al., 2026)	✓	✓	✓	✓	✗	✗	Latent pred.
LeAVJEPA (ours)	✓	✓	✓	✓	✓	✓	**Invariance + SIGReg**

Joint-embedding SSL without reconstruction BYOL (Grill et al., 2020) and SimSiam (Chen & He, 2021) match embeddings across views without negatives, and DINO (Caron et al., 2021) and DINOv2 (Oquab et al., 2024) add multi-crop self-distillation. A complementary line replaces the asymmetric teacher with an explicit distribution-level regularizer: VICReg (Bardes et al., 2022) and Barlow Twins (Žbontar et al., 2021) constrain the embedding covariance and W-MSE (Ermolov et al., 2021) whitens it. However, these methods constrain only the second-order statistics of the embedding distribution. LeAVJEPA instead inherits SIGReg (Balestriero & LeCun, 2025), which matches the full distribution to an isotropic Gaussian using a sliced characteristic-function test.

JEPAs The Joint-Embedding Predictive Architecture (LeCun, 2022) predicts the embedding of one view from another instead of reconstructing inputs. I-JEPA (Assran et al., 2023) instantiates it with spatial masking on images; V-JEPA (Bardes et al., 2024) and V-JEPA 2 (Assran et al., 2025) extend it to video and world modelling; and LeJEPA (Balestriero & LeCun, 2025) supplies the theoretical grounding through SIGReg. MJEPA (Teotia et al., 2026) is the closest audio-visual JEPA: it shares encoder parameters across modalities but trains with an EMA target encoder and a predictor network, both of which LeAVJEPA removes.

Cross-modal alignment and retrieval CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) learn cross-modal similarity by contrasting matched and mismatched image–text pairs. AudioCLIP (Guzhov et al., 2022), ImageBind (Girdhar et al., 2023), and LanguageBind (Zhu et al., 2024) extend shared embedding spaces to audio and further modalities. These approaches train explicit cross-modal alignment. In contrast, LeAVJEPA has no such objective: audio-only and video-only views are each pulled toward the same joint embedding, but are never compared directly or ranked against negatives. Retrieval therefore tests whether cross-modal correspondence emerges without being trained explicitly (Sec. 4.3).

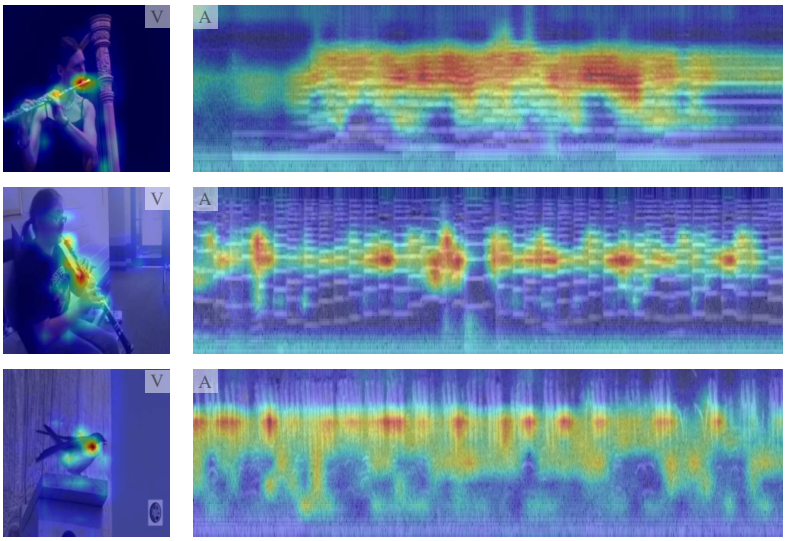
Multi-view and modality-dropout strategies SwAV (Caron et al., 2020) introduced multi-crop training with few global and many local crops, a strategy inherited by DINO and LeJEPA. AV-HuBERT (Shi et al., 2022) uses modality dropout for robustness to missing inputs. We bring these ideas together by making the missing modality define a partial view: sound alone and video alone must each approach a representation learned from their joint observation. The controlled ablation tests whether this change in view composition supplies a useful learning signal beyond ordinary cropping and masking.

3 Methods

LeAVJEPA maps audio-only, video-only, and joint audio-visual inputs of the same clip to a common embedding. Each local view loses a modality and learns to match the joint audio-video centre. We first construct these views, then describe the shared encoder and the objective that brings their embeddings together while preventing collapse (Fig. 1).

3.1 Constructing Joint and Partial Views

A joint view can use either sound or visual information to match another joint view. Removing one modality leaves the encoder with only the other, so an audio-only view must be matched from sound alone and a video-only view from appearance and motion alone. In both cases, the target is given by global views that retain both modalities.



Flute: The mouthpiece, where the sound is produced, is singled out rather than the whole instrument.

Oboe: Attention falls on the person's hands and the part of the oboe they are playing.

Bird: A small bird stands out against a textured background, and each of its chirps is picked out in the spectrogram.

Figure 2: **Attention finds the sound source without localization supervision.** Audio-to-video attention lands on the object producing the sound, and video-to-audio attention follows its harmonic and temporal structure. Neither behaviour is explicitly supervised; both emerge from matching single-modality views to a joint centre.

Joint views define the centre We sample two 8-second temporal crops with random offsets from each 10-second clip. Both global views ($G{=}2$) contain audio and video, with a centre crop and random horizontal flip for video and a clean mel spectrogram for audio. After encoding and projection, we average their embeddings to form the joint centre $\bar{z}$. Each clip has its own centre, connecting observations of that clip.

Modality dropout creates the partial observations We also form two local views ($K{=}2$) from the temporal crops. Even-indexed local views retain only audio, with a clean spectrogram; odd-indexed views retain only video, with standard augmentations. We drop the absent modality's tokens entirely, shortening the sequence. Here, "local" refers to a partial set of modalities rather than a shorter temporal extent. This extends LeJEPA's spatial augmentations (Balestriero & LeCun, 2025) to the modality axis, so that each local view is a partial observation of the clip restricted to a single modality.

The target is a representation of the event The embeddings of both partial views are pulled toward the joint centre. Neither view has to reconstruct the missing modality, so the audio view does not predict visual motion and the video view does not predict the waveform. Each instead predicts an embedding computed from the joint observation, which encourages the encoder to retain information that audio and video share. Sec. 4.1 tests this by varying only the local views.

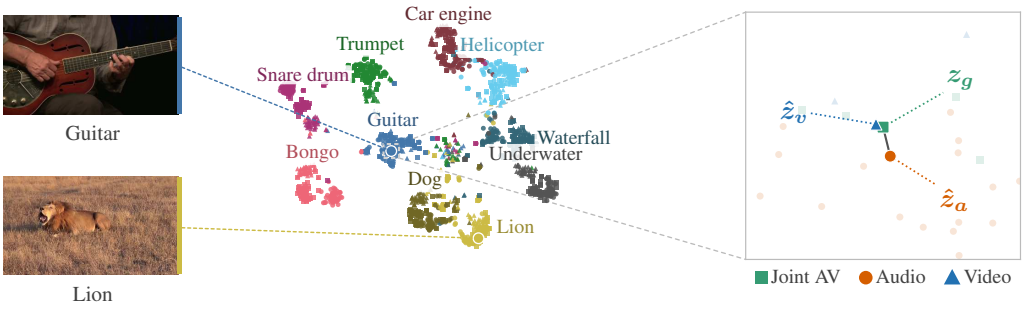
3.2 Encoding Joint and Partial Views with One Model

We project audio and video to tokens of the transformer's hidden dimension, so that a single transformer and projector process every view. Joint and partial views differ only in which tokens enter the sequence, and all are embedded in the same representation space.

Turning sound and video into tokens We resample audio to 16 kHz and convert each 8-second crop into a mel spectrogram of shape $1{\times}128{\times}801$. A 2D convolution with kernel and stride $(16, 16)$ produces $8{\times}50 = 400$ audio tokens. For video, a 3D convolution with kernel and stride $(2, 16, 16)$ maps 16 RGB frames at $224{\times}224$ resolution to $8{\times}14^2 = 1568$ tubelet tokens. Factorized positional embeddings preserve frequency and time for audio, and spatial position and time for video. Learned modality-type embeddings identify which stream each token belongs to.

Sharing computation across views We prepend a learnable `[CLS]` token and concatenate the available tokens. A joint view has the sequence

$$[\texttt{CLS};\, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{1568};\, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_{400}], \tag{1}$$



$$\mathcal{L}_{\text{inv}} = \frac{1}{G+K}\left[\sum_{g=1}^{G} \|\bar{z} - z_g\|^2 + \|\bar{z} - \hat{z}_a\|^2 + \|\bar{z} - \hat{z}_v\|^2\right]$$

Figure 3: **Semantic classes cluster, and a clip's views stay close.** t-SNE of `[CLS]` embeddings, where each clip contributes an audio-only (●), a video-only (▲), and a joint (■) point. Within each cluster, the three shapes mix: audio-only and video-only views of a class land together even though they share no input. The zoom shows the guitar sample, whose joint embedding (z_g) lies between its audio ($\hat{z}_a$) and video ($\hat{z}_v$) embeddings, as the invariance loss (bottom) encourages.

while audio-only and video-only views contain 401 and 1569 tokens, respectively. Early fusion lets the transformer attend across modalities from its first layer when both are present. When one is absent, the same weights process the surviving tokens; no separate encoder or placeholder tokens for the missing modality are needed.

For ViT-B (Dosovitskiy et al., 2021), we use 12 layers, a hidden dimension of 768, and 12 attention heads; for ViT-L, we use 24 layers, a hidden dimension of 1024, and 16 attention heads. A shared three-layer MLP projects the `[CLS]` output through $d \to 2048 \to 2048 \to 128$ dimensions, where d is the hidden dimension, with BatchNorm and GELU. This projector maps every view into the space where we compute the loss; it is not a separate cross-modal predictor. Sec. A.1 and Fig. A1 give the full architecture.

3.3 Matching Views without Collapse

We want observations of the same clip to agree while retaining differences between clips. However, matching views alone would allow the encoder to return the same vector for every input. The LeJEPA objective (Balestriero & LeCun, 2025) addresses both requirements by combining an invariance loss with distributional regularization. Let $\boldsymbol{z}_i = p_\phi(f_\theta(x_i)) \in \mathbb{R}^{d_z}$ denote a projected view embedding and $\bar{\boldsymbol{z}} = \frac{1}{G}\sum_{g=1}^{G} \boldsymbol{z}_g$ the mean of the global embeddings for the same clip. We minimize

$$\mathcal{L} = (1-\lambda)\underbrace{\frac{1}{G+K}\sum_{i=1}^{G+K} \|\bar{\boldsymbol{z}} - \boldsymbol{z}_i\|^2}_{\text{invariance}} + \lambda \underbrace{\text{SIGReg}(\{\boldsymbol{z}_i\})}_{\text{regularization}}, \tag{2}$$

where λ balances agreement and distributional regularization. The invariance term is written for one clip and averaged over the minibatch; SIGReg regularizes embeddings across clips.

A learned target without a teacher The invariance term draws all views toward their clip's centre, and gradients also flow through the global embeddings that form it. The centre is therefore learned alongside the partial views, rather than being fixed or supplied by a second encoder. All views update the shared encoder and projector through the same loss, without a stop-gradient, EMA teacher, or cross-modal predictor. The centre can consequently adapt to information that the partial observations can express.

Preventing collapse with SIGReg SIGReg (*Sketched Isotropic Gaussian Regularization*) supplies the pressure against collapse by matching the distribution of embeddings across clips to an isotropic Gaussian. It projects embeddings onto M random unit directions and compares each projection's empirical characteristic function with the Gaussian target through the Epps–Pulley test. Resampling directions at each step constrains the distribution across different slices. For a fixed number of

directions and evaluation points, the computation is linear in batch size and can aggregate the characteristic-function statistics across devices.

LeJEPA motivates the Gaussian target through its analysis of downstream probes: anisotropic covariance can amplify estimation error for linear probes, while an isotropic Gaussian minimizes integrated squared bias for the nonlinear probes considered in that analysis. These results motivate our regularizer under LeJEPA's assumptions; they do not guarantee optimal performance on arbitrary audio-visual tasks. The two terms act at different levels: invariance brings a clip's views together, while SIGReg preserves spread across clips. Their separate ablations in Sec. 4.1 test whether both are needed.

From a common target to cross-modal alignment No loss term compares audio-only and video-only embeddings directly. Both are instead pulled toward the same joint centre, and since $\|\boldsymbol{z}_a - \boldsymbol{z}_v\| \leq \|\boldsymbol{z}_a - \bar{\boldsymbol{z}}\| + \|\bar{\boldsymbol{z}} - \boldsymbol{z}_v\|$, this also brings them close to each other. Retrieval additionally requires each embedding to lie closer to its partner than to other clips. SIGReg encourages this by spreading embeddings across clips, but the objective does not enforce it directly. Fig. 3 illustrates the resulting neighbourhoods, and Sec. 4.3 evaluates retrieval quantitatively.

3.4 Monitoring the Representation

We monitor VGGSound features with a linear probe on the `[CLS]` token and an attentive probe that pools patch tokens through one cross-attention query (Bardes et al., 2024; Teotia et al., 2026). During pretraining and fine-tuning, these probes receive detached features, so their losses cannot update the encoder. We fit AudioSet probes after pretraining with the backbone frozen, following standard audio-visual evaluation (Gong et al., 2023; Araujo et al., 2025). App. B gives the training configurations.

4 Experiments

The evaluation first tests whether modality dropout is responsible for the improvement in representation quality (Sec. 4.1). We then examine whether the learned features remain useful under frozen transfer (Sec. 4.2), align audio and video well enough for cross-modal retrieval (Sec. 4.3), and improve further with end-to-end fine-tuning (Sec. 4.4).

Datasets and models AudioSet (Gemmeke et al., 2017) provides approximately two million clips with 527 multi-label sound classes; VGGSound (Chen et al., 2020) provides approximately 184k training clips across 309 event classes. The ablations pretrain a ViT-B on VGGSound under a fixed short budget. The transfer, retrieval, and fine-tuning experiments use a ViT-B and a ViT-L pretrained on AudioSet-2M with two global and two modality-specific local views and λ=0.05. App. B lists the full training configurations. All results are from single runs, except where noted.

Evaluation protocols Each table states whether it uses a frozen linear probe, a frozen attentive probe, or end-to-end fine-tuning. The probes measure information accessible in the pretrained features (Sec. 3.4); fine-tuning additionally lets supervised gradients adapt the backbone.

4.1 Learning from Modality-Specific Partial Views

The first test separates the contribution of the views from that of the encoder and loss. If ordinary joint multi-crop views were sufficient, the model could learn useful audio-visual features without ever losing a modality. We pretrain a ViT-B on VGGSound with the same encoder, hyperparameters, and training budget, varying the local-view construction and including a no-local-view control. Fig. 4 compares the resulting online probe test accuracies; Table A4 in Sec. C.4 lists the values with top-5 accuracy, loss values, and embedding statistics.

Low loss does not imply useful cross-modal features Joint local views remain close to the audio-only baseline, while global views alone perform no better (Fig. 4, left). The weak features do not arise from collapse: these variants achieve low invariance loss and near-unit embedding standard deviation (Table A4). They instead reflect a limitation of the task, since when every view contains both modalities, the views can agree through one modality alone without the encoder relating it to the

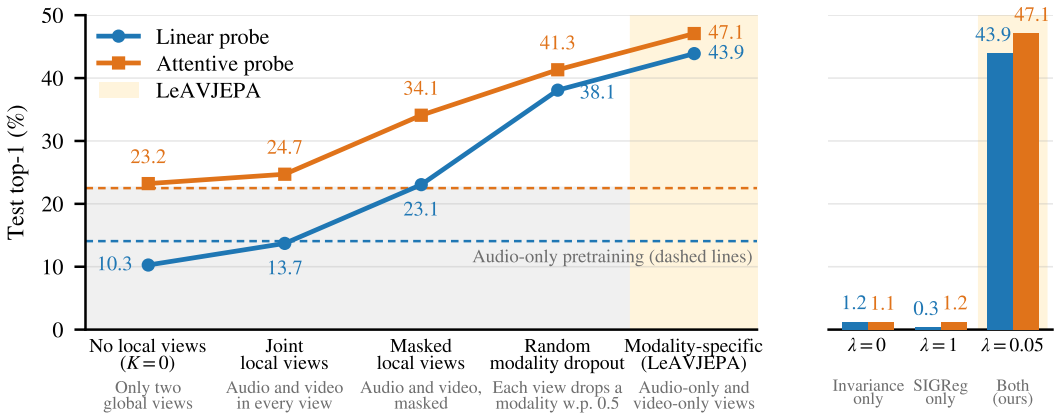


Figure 4: **Modality-specific views improve cross-modal learning in LeAVJEPA.** Left: joint local views remain near the audio-only baselines (dashed lines). Random modality dropout improves probe accuracy, and making every local view audio-only or video-only yields the largest gain. Right: either loss term alone fails to learn useful features.

other. The results are consistent with this shortcut. Distributional regularity rules out collapse, but it does not show that the features combine information from audio and video.

Dropping a modality makes the task cross-modal When a local view contains only audio or only video, the encoder must match the joint centre from that modality alone. Random modality dropout improves both probes, and consistently making each local view audio-only or video-only yields the strongest features. The agreement between linear and attentive evaluation suggests that the benefit is not specific to one type of probe. Since these comparisons retain the same encoder, objective, and training budget, they identify view composition as a consequential design choice. Merely supplying both modalities does not confer the benefit obtained by learning from partial observations.

Both loss terms are necessary The loss ablations distinguish the roles of alignment and regularization (Fig. 4, right). Removing SIGReg (λ=0) collapses the embedding and reduces classification toward chance. Removing invariance (λ=1) preserves embedding spread but also fails to learn useful features. Neither agreement alone nor distributional regularity alone is sufficient under this training protocol. The full objective makes the modality-specific task learnable while avoiding its trivial solution. Additional comparisons find similar performance with separate encoders and limited sensitivity to the tested SIGReg weights and local-view counts (Sec. C.4), suggesting that the gain does not depend on one architectural configuration or a narrowly selected hyperparameter value.

4.2 Frozen Transfer to AudioSet-20K and ESC-50

The ablation shows that modality dropout produces better features on the pretraining dataset. We next test whether those features transfer to other tasks with the encoder frozen, so that downstream labels can train a probe but cannot change the representation. Table 2 follows the frozen protocol of MJEPA (Teotia et al., 2026): an attentive probe on frozen AudioSet-pretrained features, evaluated on AudioSet-20K with audio-only (A), video-only (V), and audio-video (A-V) inputs, and on ESC-50 (Piczak, 2015). Baseline rows are the re-evaluations reported by MJEPA under the same protocol.

The frozen results show that both modalities contribute to the joint representation (Table 2). Audio-only input is stronger than video-only input, yet combining them improves over audio alone for both model sizes. The larger backbone improves across audio, video, and joint evaluation, so its advantage is not restricted to a single input type. Transfer to ESC-50 further shows that features learned through the joint objective remain useful when only audio is available downstream.

LeAVJEPA is competitive with several separate-encoder methods, although both model sizes remain below MJEPA. This supports the utility of the simpler objective, which reaches these results without MJEPA's EMA teacher or predictor. The probe follows the evaluation protocol of MJEPA (Teotia et al., 2026), and Sec. C.3.1 documents its settings.

Table 2: Frozen transfer with an attentive probe. AudioSet-20K mAP for audio-only (A), video-only (V), and audio-video (A-V) inputs, and ESC-50 accuracy. Baseline frozen results and parameter counts are as reported by Teotia et al. (2026); the corresponding fully fine-tuned results are collected in Table A3 (Sec. C.3). Best result per column in bold. *Not reported in the MJEPA paper; provided by the authors through personal correspondence.

				AudioSet-20K (mAP)			ESC-50
	Method	Params	Pre-train data	A	V	A-V	Acc.
Separate encoders	CAV-MAE (Gong et al., 2023)	170M	IN+AS	19.38	18.14	34.59	77.5
	MAViL (Huang et al., 2023)	170M	IN+AS	30.00	–	–	90.8
	EquiAV (Kim et al., 2024)	170M	IN+AS	34.25	18.60	38.60	93.2
	CAV-MAE Sync (Araujo et al., 2025)	170M	IN+AS	21.66	16.20	28.50	89.2
	XKD (Sarkar & Etemad, 2024)	170M	AS	–	–	–	93.6
Shared encoder	MJEPA ViT-B* (Teotia et al., 2026)	86M	AS	37.54	20.18	40.68	–
	MJEPA ViT-L (Teotia et al., 2026)	300M	AS	**38.89**	**25.38**	**42.90**	**95.2**
	LeAVJEPA ViT-B (ours)	87M	AS	27.79	14.82	30.98	88.0
	LeAVJEPA ViT-L (ours)	304M	AS	31.56	18.25	36.03	91.3

Table 3: AudioSet cross-modal retrieval on the CAV-MAE balanced evaluation subset (1,725 clips; 1,722 available for our evaluation), recall in percent. Baseline rows are as reported in the respective papers and all train an explicit cross-modal alignment term or predictor; LeAVJEPA rows are zero-shot cosine retrieval unless marked. Best baseline result per column in bold. †MJEPA retrieves through its jointly trained cross-modal predictor with an L1 nearest-neighbour search rather than direct embedding similarity. ‡Contrastive alignment head trained on frozen features; not zero-shot.

	Audio → video			Video → audio		
Method	R@1	R@5	R@10	R@1	R@5	R@10
CAV-MAE (Gong et al., 2023)	15.1	34.0	43.0	18.8	39.5	50.1
CAV-MAE Sync (Araujo et al., 2025)	27.9	52.4	62.2	**35.2**	**58.3**	**67.6**
EquiAV (Kim et al., 2024)	**29.6**	**53.7**	**63.1**	30.1	53.3	62.9
MJEPA ViT-L† (Teotia et al., 2026)	29.3	**53.7**	61.8	30.6	53.4	61.3
LeAVJEPA ViT-B (ours)	10.5	28.9	36.6	11.4	28.2	36.8
LeAVJEPA ViT-L (ours)	19.1	40.5	50.4	18.7	42.1	51.9
LeAVJEPA ViT-L + InfoNCE head‡ (ours)	24.4	46.9	57.7	25.6	48.0	57.1

4.3 Cross-Modal Retrieval without a Contrastive Loss

Classification can succeed without placing paired audio and video close enough for retrieval. We therefore test a more specific consequence of the common-centre objective: can a single-modality embedding identify its partner among other clips? We encode each clip twice, from its audio tokens alone and from its video tokens alone, and rank candidates by cosine similarity of the projected `[CLS]` features. We evaluate on the balanced AudioSet retrieval subset released with CAV-MAE (Gong et al., 2023), the pool also used by CAV-MAE Sync, EquiAV, and MJEPA.

Direct cosine retrieval is well above chance in both directions and improves consistently with the larger backbone (Table 3). This shows that the model can identify a paired observation even when the query and candidate share no input modality. Nevertheless, the strongest methods trained with explicit cross-modal alignment remain ahead, indicating that matching a joint centre leaves room to improve cross-modal alignment.

To test whether the frozen features support stronger retrieval with a learned head, we fit a contrastive (InfoNCE) head on 300 AudioSet training clips per class and select the best head on a held-out validation set. The head improves retrieval in both directions without updating the encoder.

Table 4: Audio-visual classification on VGGSound (top-1, end-to-end fine-tuning). Published baselines all use separate audio and video encoders, except AVSiam, and are listed above the rule; our AudioSet-pretrained shared-encoder ViT-B (87M) and ViT-L (304M) below it. Best result per block in bold.

Method	Params	Init.	Pretraining	Top-1
AV-MAE (Georgescu et al., 2023)	172M	–	VGGS	64.2
AV-MAE (Georgescu et al., 2023)	611M	–	VGGS	65.0
AVSiam (Lin & Bertasius, 2024)	100M	IN-21K SL	AS-2M	64.9
AVSiam (Lin & Bertasius, 2024)	332M	IN-21K SL	AS-2M	67.1
CAV-MAE (Gong et al., 2023)	164M	IN-SSL	AS-2M	65.5
MAViL (Huang et al., 2023)	172M	IN-SSL	AS-2M	**67.1**
EquiAV (Kim et al., 2024)	172M	IN-SSL	AS-2M	**67.1**
LeAVJEPA (Linear)	87M	–	AS-2M	58.3
LeAVJEPA (Attentive)	87M	–	AS-2M	60.4
LeAVJEPA (Linear)	304M	–	AS-2M	59.9
LeAVJEPA (Attentive)	304M	–	AS-2M	**61.1**

4.4 End-to-End Fine-Tuning on VGGSound

We finally allow supervision to adapt the entire encoder, testing downstream performance under the end-to-end protocol used by most published audio-visual SSL comparisons. Sec. C.3 documents the training settings and an additional comparison with CAV-MAE Sync under a matched frozen probe.

Attentive pooling improves over the linear head for both model sizes (Table 4), showing that the choice of classification head still matters after fine-tuning. However, the advantage of the larger backbone is less pronounced here than in frozen transfer or retrieval. This contrast shows why frozen and fine-tuned performance should be reported separately, since fine-tuning can compensate for differences in representation quality that frozen evaluation exposes.

The published baselines retain an advantage after fine-tuning, with a similar ordering on AudioSet-20K (Table A3). Differences in initialization, objectives, and optimization prevent attributing this gap to any individual component. The shared-encoder approach therefore supports useful transfer, while matching the strongest fully adapted systems remains an open challenge.

5 Discussion and Conclusion

LeAVJEPA shows that the construction of partial views is central to learning useful audio-visual representations. Treating sound and vision as incomplete observations of the same event allows a shared encoder to learn cross-modal correspondence through a symmetric objective. This extends the role of view construction in self-supervised learning, since withholding a modality determines which input the encoder must use to predict a representation learned from both modalities. Agreement between views is informative only insofar as the views define a useful learning task. Modality-specific observations substantially improve the learned features under the same encoder and loss, demonstrating that the information retained in each view deserves explicit consideration when designing multimodal objectives. The contribution of LeAVJEPA is to establish that useful alignment and transfer can already emerge without a dedicated cross-modal predictor, reconstruction decoder, or teacher network. More broadly, our results show that a single encoder, a symmetric objective, and views that each contain only one modality are enough to learn strong cross-modal representations, offering a simple, scalable foundation for future multimodal models.

Limitations and future work Our evaluation relies mainly on sound-centric benchmarks, leaving the encoder's broader visual capabilities unresolved. Published baselines also use different training protocols, so performance gaps are hard to attribute. Moreover, matching partial observations assumes that they contain sufficient shared information. Off-screen sounds and unrelated visual activity offer relevant tests of when that assumption no longer holds. A natural next step is to extend modality dropout beyond audio and video, for example to text or depth, and to evaluate on video-centric tasks. With the method reduced to one encoder and two loss terms, another is to make it cheaper to train, for example with fewer views, tokens, or steps.

AI USE STATEMENT

We used generative AI tools to edit the text of this paper, including grammar, phrasing, and condensing prose; to write and debug parts of the training, evaluation, and plotting code, and to set up and manage experiments; to assist with parts of the experimental design; and to help search for and summarise some of the related work. We did not use generative AI tools to generate the core research idea or method. We have reviewed all AI-assisted work. AI-generated code and experiment configurations were verified and tested by the authors before use in any reported experiment, and the authors made all final decisions on experimental design. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only publicly available research datasets (AudioSet, VGGSound, and ESC-50) under their existing terms of use. It involves no human subjects, no new data collection, and no release of new data. The authors declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

Anonymous source code is provided as supplementary material, covering the model, data pipeline, pretraining, fine-tuning, probing, and retrieval, together with the job scripts for the reported runs. The code and pretrained checkpoints will be released publicly upon acceptance. Sec. 3 describes the view construction, encoder, and objective, and App. A specifies the encoder architecture and probe readouts. App. B and Secs. B.2 and B.3 list the full pretraining, fine-tuning, and probe configurations, including optimiser settings, batch sizes, and hardware. All experiments use public datasets.

REFERENCES

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 24206–24221, 2021.

Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 9758–9770, 2020.

Relja Arandjelović and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 609–617, 2017.

Edson Araujo, Andrew Rouditchenko, Yuan Gong, Saurabhchand Bhati, Samuel Thomas, Brian Kingsbury, Leonid Karlinsky, Rogerio Feris, James R. Glass, and Hilde Kuehne. 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 (CVPR)*, pp. 18794–18803, 2025.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6816–6826, 2021.

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.

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

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 12449–12460, 2020.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

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

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

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pp. 813–824, 2021.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 9912–9924, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725, 2020.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pp. 5178–5193, 2023.

Xinlei Chen and Kaiming He. Exploring simple Siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15745–15753, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pp. 3015–3024, 2021.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017.

Mariana-Iuliana Georgescu, Eduardo Fonseca, Radu Tudor Ionescu, Mario Lucic, Cordelia Schmid, and Anurag Arnab. Audiovisual masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16098–16108, 2023.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190, 2023.

Yuan Gong, Yu-An Chung, and James Glass. AST: Audio spectrogram transformer. In *Interspeech*, pp. 571–575, 2021.

Yuan Gong, Cheng-I Jeff 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.

Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. Contrastive audio-visual masked autoencoder. In *International Conference on Learning Representations (ICLR)*, 2023.

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.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. AudioCLIP: Extending CLIP to image, text and audio. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 976–980, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2022.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 28708–28720, 2022.

Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, Haoqi Fan, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, and Christoph Feichtenhofer. MAViL: Masked audio-video learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 20371–20393, 2023.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pp. 4904–4916, 2021.

Jongsuk Kim, Hyeongkeun Lee, Kyeongha Rho, Junmo Kim, and Joon Son Chung. EquiAV: Leveraging equivariance for audio-visual contrastive learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pp. 24327–24341, 2024.

Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pp. 7763–7774, 2018.

Yann LeCun. A path towards autonomous machine intelligence. OpenReview preprint, version 0.9.2, 2022. URL https://openreview.net/forum?id=BZ5a1r-kVsf. Version 0.9.2, 2022-06-27.

Yan-Bo Lin and Gedas Bertasius. Siamese vision transformers are scalable audio-visual learners. In *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science, pp. 303–321. Springer, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12470–12481, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Karol J. Piczak. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 1015–1018, 2015.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pp. 8748–8763, 2021.

Pritam Sarkar and Ali Etemad. XKD: Cross-modal knowledge distillation with domain alignment for video representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14875–14885, 2024.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *International Conference on Learning Representations (ICLR)*, 2022.

Revant Teotia, Adrien Bardes, Michael Rabbat, Sumit Chopra, Matthew J. Muckley, and Nicolas Ballas. MJEPA: A simple and scalable joint-embedding predictive architecture for audio-visual learning. In *Computer Vision – ECCV 2026, Part II*, Lecture Notes in Computer Science, pp. 97–114. Springer, 2026.

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

Jure Žbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow Twins: Self-supervised learning via redundancy reduction. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pp. 12310–12320, 2021.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. LanguageBind: Extending video-language pretraining to N-modality by language-based semantic alignment. In *International Conference on Learning Representations (ICLR)*, 2024.

APPENDICES

App. A describes the encoder and the linear and attentive probes. App. B documents pretraining, fine-tuning, and probe training. App. C presents training curves, qualitative analyses of the learned representations, additional benchmark comparisons, and ablation and hyperparameter studies.

A Method Details **15**

A.1 Encoder Architecture 15

A.2 Linear and Attentive Probes 15

B Implementation Details **16**

B.1 Pretraining Configuration 16

B.2 Fine-Tuning Configuration 17

B.3 Probe Training and Evaluation 17

C Additional Results **18**

C.1 Training Curves 18

C.1.1 Embedding-Quality Curves 18

C.1.2 VGGSound Pretraining Curves 18

C.1.3 AudioSet Pretraining Curves 19

C.1.4 Fine-Tuning Curves 19

C.2 Qualitative Analysis of Learned Representations 20

C.2.1 Semantic Structure of the Embedding Space 20

C.2.2 Feature PCA of Video Patch Tokens 21

C.3 Additional Benchmark Comparisons 22

C.3.1 Evaluation and Training-Budget Comparisons 22

C.4 Ablations and Hyperparameter Sensitivity 23

C.4.1 Hyperparameter Sensitivity 24

C.5 Loss-Component and Pipeline Ablations 25

C.5.1 Removing the SIGReg term (λ=0) 25

C.5.2 Removing the invariance term (λ=1) 26

C.5.3 Removing modality dropout and masking 27

A METHOD DETAILS

A.1 ENCODER ARCHITECTURE

Fig. A1 shows the full LeAVJEPA encoder used for every view (global and local) during pretraining. Video frames of shape $3\times16\times224\times224$ are tokenized by a $2\times16\times16$ Conv3D tubelet into 1568 video tokens; the mel spectrogram of shape $1\times128\times801$ is tokenized by a 16×16 Conv2D into 400 audio tokens. Both streams receive factorized positional embeddings (spatial + temporal for video; frequency + time for audio) and a learnable modality-type embedding (ID 0 for video, ID 1 for audio). A learnable `[CLS]` token is prepended and the resulting 1969-token sequence is passed through a 12-layer ViT-Base (d=768, 12 heads, FlashAttention-2). On local views, the tokens of the dropped modality are omitted from the sequence entirely (`[CLS]` plus 1568 video or 400 audio tokens), so the same encoder handles joint, audio-only, and video-only inputs without architectural changes. The `[CLS]` output is taken as the clip embedding and fed to the projection MLP for the LeJEPA loss (Fig. 1).

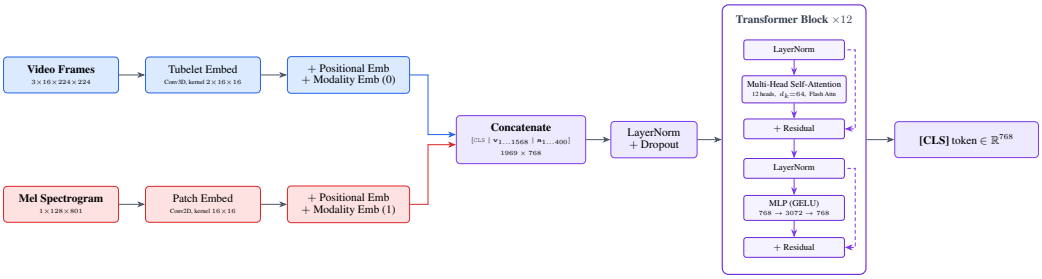


Figure A1: LeAVJEPA early-fusion encoder. Video tubelets and mel-spectrogram patches are embedded, summed with factorized positional and modality-type embeddings, concatenated with a `[CLS]` token, and processed by a 12-layer ViT-Base. The same encoder is used for every view, including the modality-dropout local views, where the absent modality contributes no tokens and the sequence is correspondingly shorter.

A.2 LINEAR AND ATTENTIVE PROBES

The linear probe applies LayerNorm and a linear classifier to the `[CLS]` token. The attentive probe instead pools encoder patch tokens through cross-attention. Concretely, the attentive probe maintains a single learnable query that attends to the full sequence of encoder patch tokens (audio and video, excluding `[CLS]`) through a multi-head cross-attention layer (d=768, 12 heads), followed by LayerNorm and a linear classifier that produces the class logits. Unlike the linear probe, which is forced to summarize the sequence through the `[CLS]` token alone, the attentive probe adaptively aggregates the most informative audio and video patches per sample, providing a more expressive probe of the frozen backbone (Bardes et al., 2024; Teotia et al., 2026).

B Implementation Details

B.1 Pretraining Configuration

Table A1 lists the AudioSet-2M pretraining configurations of the ViT-B and ViT-L backbones. No online probes are attached during AudioSet pretraining; AudioSet probes are fitted post hoc on the frozen backbone (Sec. B.3).

Table A1: AudioSet-2M pretraining configurations. Settings spanning both columns are shared.

Setting	ViT-B	ViT-L
Data		
Dataset	AudioSet-2M (unbalanced), ∼1.9M clips	
Clip length	8 s (from a 10 s clip, random offset)	
Architecture		
Encoder	12L, 768d, 12h	24L, 1024d, 16h
Parameters	87M	304M
Audio embed	Conv2D, kernel/stride 16×16	
Video embed	Conv3D, kernel/stride 2×16×16	
Sequence length	1969 (1 CLS + 1568 video + 400 audio)	
Projector	768→2048→2048→128	1024→2048→2048→128
LeJEPA training procedure		
Global views G	2	
Local views K	2 (1 audio-only, 1 video-only)	
λ (SIGReg weight)	0.05	
Optimization		
Optimizer	AdamW (Loshchilov & Hutter, 2019)	
Learning rate	5×10^{-4}	
Weight decay	0.05	
LR schedule	Linear warmup (15%) + cosine to 10^{-6}	
Gradient clipping	5.0	
Mixed precision	bf16	
Gradient checkpointing	No	Yes
Batch size (per GPU)	40	128
Effective batch size	320	1024
Compute		
GPUs	8× NVIDIA H200	8× NVIDIA B300
Parallelism	DDP	
Wall-clock time	∼192 h (to epoch 57)	∼160 h (to epoch 66)

B.2 Fine-Tuning Configuration

Table A2 lists the VGGSound fine-tuning configuration used for both AudioSet-pretrained backbones in Table 4. The two runs share the recipe and effective batch size.

Table A2: VGGSound fine-tuning configuration for the AudioSet-pretrained ViT-B and ViT-L backbones. Settings spanning both columns are shared.

Setting	ViT-B	ViT-L
Probes on top of pretrained backbone		
Linear classifier	LayerNorm + linear on `[CLS]` (309 classes)	
Attentive classifier	1 query, 12-head cross-attention	1 query, 16-head cross-attention
Optimization		
Optimizer	AdamW	
Head learning rate	2×10^{-4}	
Backbone learning rate	2×10^{-5} ($0.1\times$ head LR)	
Weight decay	0.05	
LR schedule	Linear warmup (5%) + cosine to 10^{-7}	
Gradient clipping	1.0	
Label smoothing	0.1	
Mixed precision	bf16	
Batch size (per GPU)	80	40
Effective batch size	320	
Epochs	20	
Data and augmentation		
Views	Single clip, no SSL views or masking	
SpecAugment	Frequency mask 24, time mask 80	
Mixup α	0.2	
Random resized crop	Min. scale 0.5	
Compute		
GPUs	$4\times$ NVIDIA H200	$8\times$ NVIDIA H200
Parallelism	DDP	

B.3 Probe Training and Evaluation

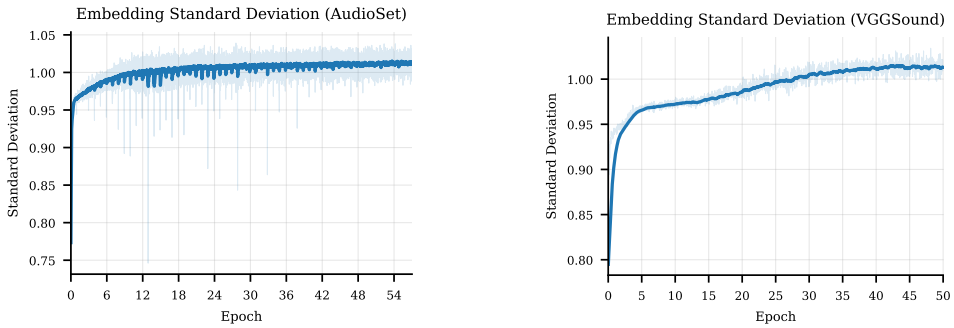
During VGGSound pretraining and fine-tuning, two classification probes are trained on frozen backbone features (gradients detached from the encoder): a **linear probe** (LayerNorm + linear classifier on the `[CLS]` token) and an **attentive probe** that pools information across patch tokens via cross-attention. We do not attach these probes during AudioSet pretraining; AudioSet probes used at evaluation time are trained post-hoc on top of the frozen AudioSet-pretrained backbone (Sec. 4.2). Both probes are trained with cross-entropy loss (label smoothing 0.1) using a separate learning rate (10^{-3}) and no weight decay, following the LeJEPA evaluation protocol of frozen-backbone probing, which is also standard practice in audio-visual SSL (Gong et al., 2023; Araujo et al., 2025).

C Additional Results

C.1 Training Curves

C.1.1 Embedding-Quality Curves

Fig. A2 tracks the per-dimension standard deviation of the projected embeddings over training for both pretraining runs. In each case the value rises rapidly toward $\sim$1.01 (from $\sim$0.8 on VGGSound and $\sim$0.8 on AudioSet), confirming that SIGReg drives the embedding distribution toward the target isotropic Gaussian $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ at both VGGSound and AudioSet scales.

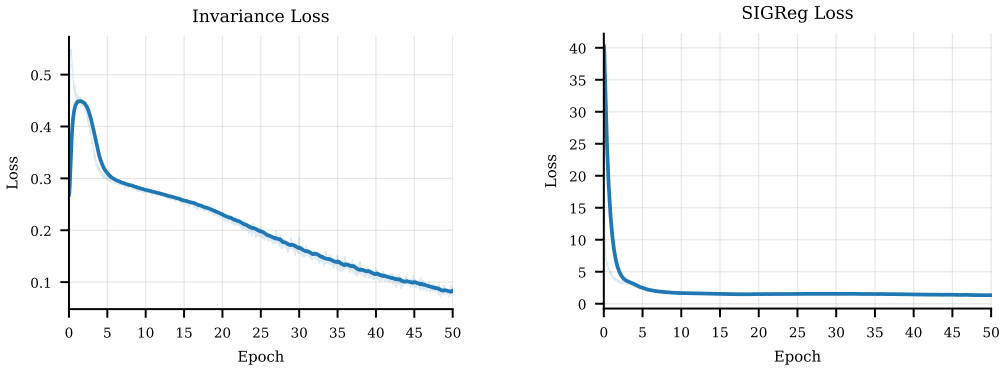


(a) AudioSet-2M (57 ep.) (b) VGGSound (50 ep.)

Figure A2: Embedding standard deviation over training. SIGReg drives the per-dimension std toward 1.0 (the isotropic Gaussian target).

C.1.2 VGGSound Pretraining Curves

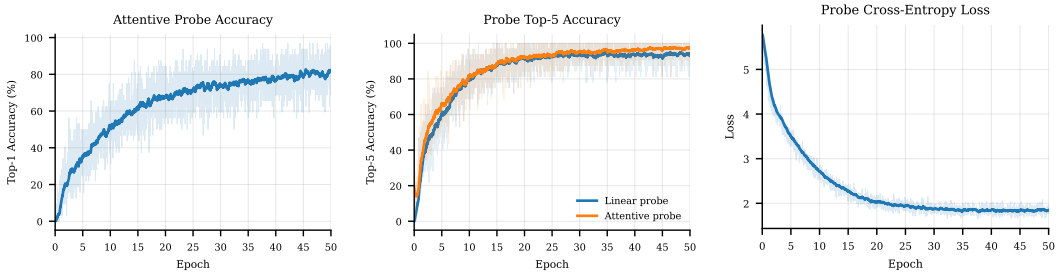
Pretraining loss components Fig. A3 decomposes the LeJEPA objective into its two unweighted terms. The invariance term decreases steadily, indicating that view embeddings converge toward the global center $\bar{z}$ and that cross-modal alignment is successful. The SIGReg term drops sharply during the first epoch and then plateaus near zero, confirming that the embedding distribution matches the target isotropic Gaussian early in training and remains there for the rest of the run.



(a) Invariance loss (b) SIGReg loss

Figure A3: Pretraining loss components (VGGSound-only run, 50 epochs).

Online probe curves Fig. A4 reports the online linear and attentive probes on frozen backbone features during pretraining: attentive probe top-1 accuracy, top-5 accuracy for both probes, and the combined cross-entropy loss. The attentive probe consistently outperforms the linear probe by ∼2–3 points across both top-1 and top-5.

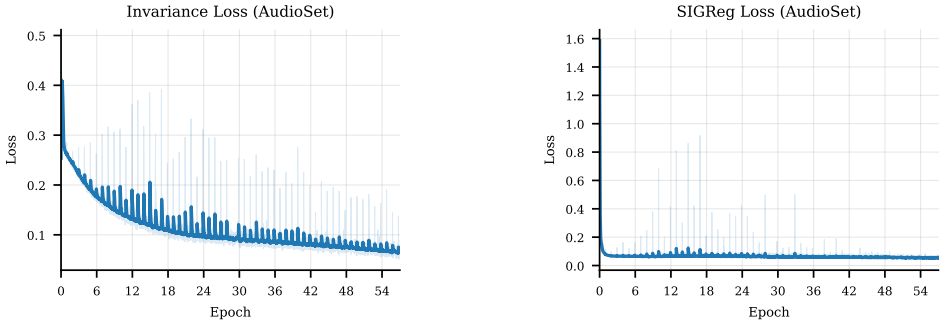


(a) Attentive probe top-1 (b) Linear / attentive top-5 (c) Probe cross-entropy loss

Figure A4: Online probing curves on frozen backbone features during pretraining.

C.1.3 AudioSet Pretraining Curves

We complement the AudioSet pretraining loss in the main text with the per-component invariance and SIGReg curves (Fig. A5). The qualitative behaviour mirrors the VGGSound-only run: the invariance term decreases steadily and the SIGReg term drops sharply during the first epoch and then plateaus near zero, confirming that SIGReg behaves as expected at AudioSet scale.

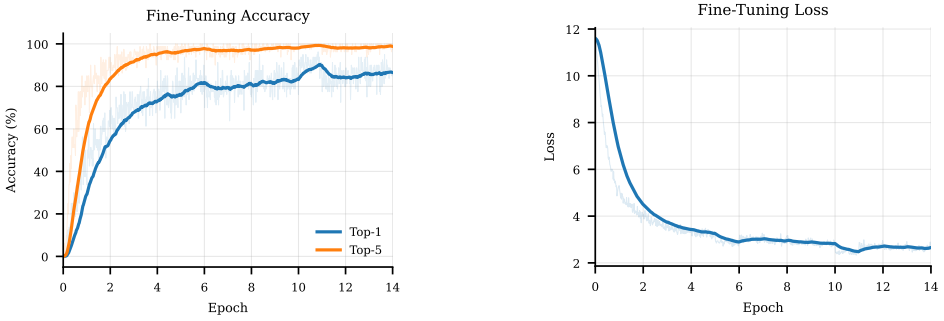


(a) Invariance loss (b) SIGReg loss

Figure A5: AudioSet pretraining loss components over 57 epochs of AudioSet-2M.

C.1.4 Fine-Tuning Curves

Fig. A6 shows the end-to-end fine-tuning curves (top-1/top-5 accuracy and cross-entropy loss) when the pretrained LeAVJEPA backbone is fine-tuned with a linear classifier on the `[CLS]` token.



(a) Top-1 / Top-5 accuracy (b) Cross-entropy loss

Figure A6: Fine-tuning curves for end-to-end fine-tuning of the pretrained LeAVJEPA backbone.

C.2 Qualitative Analysis of Learned Representations

C.2.1 Semantic Structure of the Embedding Space

Fig. A7 asks whether the embedding space is organised by semantic content. We encode N=11,143 VGGSound training clips spanning 60 classes, grouped into six coarse semantic families (instruments, animals, vehicles, water/weather, human voice, and sports), take the `[CLS]` embedding of each clip from the fine-tuned LeAVJEPA ViT-B backbone, and project the L2-normalised embeddings to two dimensions with t-SNE; points are coloured by family. Clips organise into compact, well-separated clusters that respect the family grouping: each family occupies coherent regions of the projected space, and the finer per-class structure within a family (for example the individual instruments) is visible as distinct same-colour sub-clusters. The residual overlap is concentrated between families that are acoustically and visually related, most notably animal vocalisations and human voice, which indicates that the learned space is structured by semantics rather than by low-level modality cues.

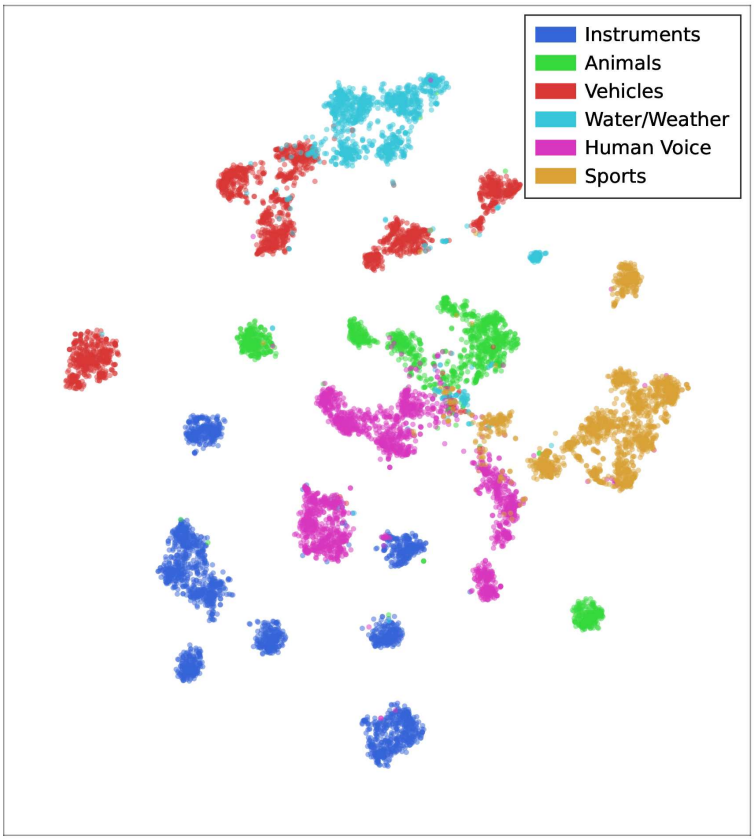
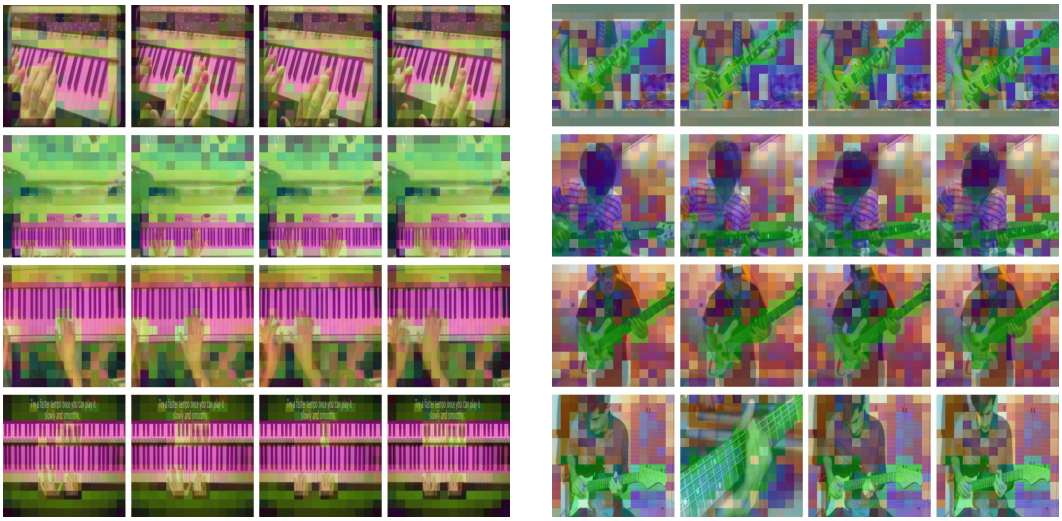


Figure A7: Semantic structure of the embedding space. t-SNE of the `[CLS]` embeddings of the fine-tuned LeAVJEPA ViT-B backbone for N=11,143 VGGSound training clips drawn from 60 classes grouped into six semantic families (colours). Clips cluster by family with finer per-class sub-structure of the same colour; the residual overlap falls mainly between acoustically related families (animal calls vs. human voice).

C.2.2 Feature PCA of Video Patch Tokens

Fig. A8 visualises the video patch-token features of the fine-tuned LeAVJEPA ViT-B encoder in the style of DINO/DINOv2 feature PCA. For each instrument class we fit a *single* PCA jointly over the pooled last-layer video patch tokens of four VGGSound clips, map the top three principal components to RGB, and overlay the result on the video frames (rows are clips, columns are frames; audio tokens are left uncoloured); because the basis is shared across the four clips, a given colour denotes the same direction in feature space. The sounding object (the piano keyboard, the body and neck of the bass guitar) then takes a consistent colour across clips and frames, clearly separated from the player and the background, indicating that the encoder represents the visually salient sound source as a coherent, instance-corresponding region of feature space, even though it is trained only with clip-level objectives and never receives pixel- or region-level supervision, providing a qualitative view of spatial feature structure; the audio-conditioning control is discussed in the main text.



(a) Playing piano (b) Playing bass guitar

Figure A8: Feature PCA of video patch tokens for two VGGSound instrument classes. A single PCA is fit jointly over the last-layer video patch tokens of four clips per class (rows); its top three components are mapped to RGB and overlaid on four frames per clip (columns). The instrument takes a consistent colour across clips and frames, distinct from the player and the background.

C.3 Additional Benchmark Comparisons

This subsection collects the fully fine-tuned results that the main text summarizes in one sentence each, together with two VGGSound evaluations that vary the probe and the training budget.

Fully fine-tuned results. Table A3 lists the end-to-end fine-tuned AudioSet-20K and ESC-50 results of the baselines in Table 2, as reported in the respective papers, next to our fine-tuned ViT-B. On AudioSet-20K the ViT-B reaches 29.6 mAP with joint input, 23.7 with audio only, and 10.3 with video only; joint input outperforms either modality alone, which establishes complementary visual information without establishing that the remaining asymmetry is caused by training duration. On ESC-50 the ViT-B is fine-tuned for 60 epochs per fold with a linear head, head learning rate 10^{-3}, backbone multiplier 0.1, SpecAugment with 48 frequency and 150 time-frame masks, random time shift, and no video tokens (audio-only input); we report final-epoch accuracy averaged over the official five folds. Fine-tuned and frozen rows are not equivalent evaluations, and the fine-tuned baselines remain well ahead in both benchmarks.

Table A3: Full-model fine-tuning results. Baseline metrics are as reported in the respective papers and baseline parameter counts follow Teotia et al. (2026); LeAVJEPA rows are our AudioSet-pretrained ViT-B. AudioSet-20K entries are mAP for audio-only (A), video-only (V), and audio-video (A-V) inputs; ESC-50 is five-fold accuracy. Best baseline result per column in bold.

			AudioSet-20K (mAP)			ESC-50
Method	Params	Pre-train data	A	V	A-V	Acc.
CAV-MAE (Gong et al., 2023)	170M	IN+AS	37.70	19.80	42.00	–
MAViL (Huang et al., 2023)	170M	IN+AS	41.80	24.80	44.90	94.4
XKD (Sarkar & Etemad, 2024)	170M	AS	–	–	–	**96.5**
EquiAV (Kim et al., 2024)	170M	IN+AS	**42.40**	**25.70**	**46.60**	96.0
LeAVJEPA ViT-B (ours)	87M	AS	23.7	10.3	29.6	91.4

C.3.1 Evaluation and Training-Budget Comparisons

Matching the frozen probe of CAV-MAE Sync. We compare a ViT-B backbone pretrained on AudioSet for 25 epochs with CAV-MAE Sync (Araujo et al., 2025) using a probe with two transformer layers followed by a `[CLS]` classifier. The backbone remains frozen while the probe is fitted for 10 VGGSound epochs. LeAVJEPA reaches 52.96% top-1 and 79.71% top-5, compared with the published 52.7% top-1. A linear probe on the same backbone reaches 47.43%. This is comparable performance under matched probe structure and epoch budgets, rather than evidence of an improvement: our evaluation averages six temporal crops and uses 8-second audio, whereas the baseline uses an extended forward pass and 4-second audio. Using a single crop reduces our result by 0.72 percentage points.

Training duration on VGGSound. Extending ViT-B training from 15 pretraining and 3 fine-tuning epochs to 57 and 20, together with a fine-tuning configuration using a backbone learning-rate multiplier of 0.1 and modality dropout 0.25, increases attentive-head top-1 from 53.9% to 60.4% (58.3% with the linear head), the values reported in Table 4. Since the training budget and the fine-tuning configuration both change, the improvement cannot be attributed to pretraining duration alone.

AudioSet-20K probe fits. The ViT-L audio-video result of 36.03 mAP in Sec. 1 and Table 2 follows the evaluation protocol of MJEPA (Teotia et al., 2026). It uses the ViT-L backbone pretrained for 31 epochs and a frozen depth-4 attentive probe fitted for 40 epochs with learning rate 3×10^{-4}, weight decay 10^{-4}, batch size 128, and four-clip logit averaging on 19,446 evaluation videos. A refit reaches 35.54 mAP. These are distinct fits, not multi-seed confidence intervals.

C.4 Ablations and Hyperparameter Sensitivity

We complement the main results with one architectural ablation and two hyperparameter sensitivity studies, all run as VGGSound-only pretraining with the same ViT-B early-fusion training configuration as the controlled run reported in the main text. The dual-encoder ablation is run on 8× NVIDIA H200 to match the controlled main-text run; the SIGReg-weight λ and local-view K sweeps are run on 1× NVIDIA H200 (batch size 32 per GPU) under a tighter compute budget. All curves are clipped to the shortest variant's final epoch, ensuring a common training budget.

Partial-view ablation (full table). Table A4 lists the local-view variants of Fig. 4 with top-5 accuracy, invariance loss, and embedding standard deviation. All variants use the same ViT-B configuration (effective batch size 512, learning rate 5×10^{-4}, λ=0.05, two global and two local views) and are compared at epoch 30, with probe accuracies on the held-out test set and losses averaged over the preceding half epoch. Both loss-term variants of Fig. 4 sit at the floor throughout.

Table A4: View ablation on VGGSound at epoch 30. Probe entries are online top-1/top-5 test accuracy; results are single runs.

Pretraining views	Linear	Attentive	Invariance	Embed. std
Audio-only and video-only local (LeAVJEPA)	43.9 / 72.7	47.1 / 74.9	0.042	0.989
Random modality dropout (p=0.5)	38.1 / 67.7	41.3 / 71.2	0.060	0.966
Both modalities + masking	23.1 / 49.2	34.1 / 63.1	0.018	0.982
Both modalities in local views	13.7 / 33.5	24.7 / 51.2	0.004	0.984
No local views (K=0)	10.3 / 26.9	23.2 / 48.2	0.003	0.985
Audio-only pretraining	14.1 / 35.1	22.5 / 48.5	0.005	0.984

Shared vs. dual encoder (ablation). Fig. A9 compares the shared early-fusion encoder used in LeAVJEPA against a dual-encoder variant that processes audio and video through separate ViT-B encoders and averages the two `[CLS]` tokens. The shared encoder was comparable to the dual-encoder variant in this single-run comparison on both linear- and attentive-probe accuracy throughout training; we adopt early fusion for its simpler, single-encoder design.

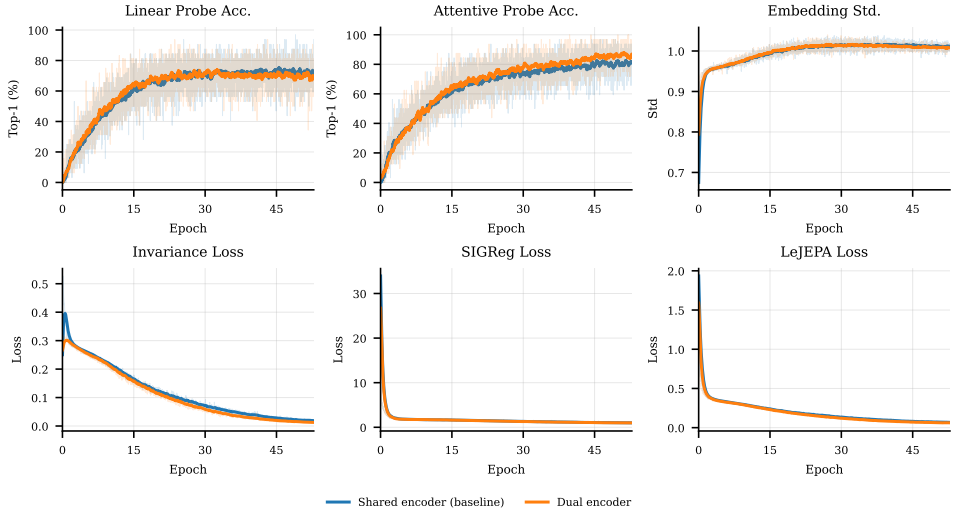


Figure A9: Architectural ablation: shared vs. dual encoder on VGGSound pretraining. Linear-/attentive-probe top-1 accuracy, embedding standard deviation, invariance loss, SIGReg loss, and total LeJEPA loss are plotted against pretraining epoch. Curves are clipped to the shorter of the two runs (53 epochs).

C.4.1 HYPERPARAMETER SENSITIVITY

SIGReg weight λ (LeJEPA hyperparameter). Fig. A10 sweeps the SIGReg weight λ in the LeJEPA loss $\mathcal{L} = (1-\lambda)\mathcal{L}_{\text{inv}} + \lambda\mathcal{L}_{\text{SIGReg}}$ across $\{0.03, 0.05, 0.10\}$. λ=0.03 and λ=0.05 produce nearly identical probe-accuracy curves, both reaching roughly 65–70% training-split linear and attentive top-1 by epoch 21. Increasing the weight to λ=0.10 degrades the representation sharply: the linear probe stays near 13% and the attentive probe near 24%, and the invariance loss plateaus around 0.33 instead of decreasing towards 0.1. The embedding standard deviation remains close across settings, so the failure at λ=0.10 is a loss of view alignment rather than collapse. LeAVJEPA is therefore insensitive to λ at or below the default but sensitive to larger SIGReg weights.

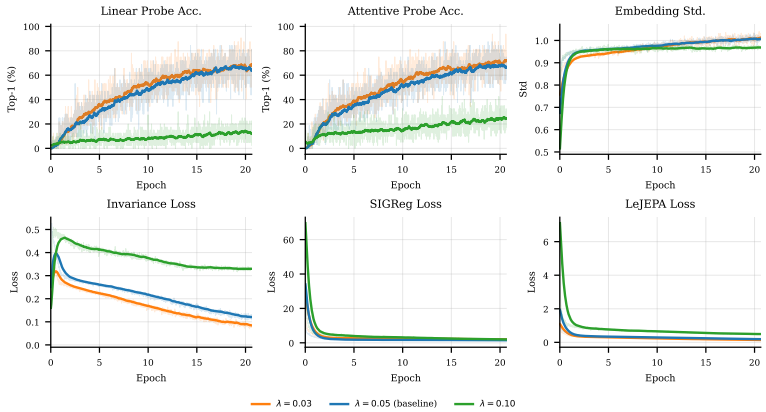


Figure A10: LeJEPA hyperparameter sensitivity: SIGReg weight λ on VGGSound pretraining, sweeping $\lambda \in \{0.03, 0.05, 0.10\}$ with λ=0.05 as the baseline used elsewhere in the paper. Probe accuracies are from the online probes on the training split. λ=0.03 and 0.05 behave almost identically, whereas λ=0.10 leaves both probes far lower and the invariance loss far higher. Curves are clipped to the shortest run (λ=0.10, 21 epochs).

Number of local views K (LeJEPA hyperparameter). Fig. A11 sweeps the number of local (modality-dropout) views K across $\{2, 4, 6, 8\}$ while keeping the two global views fixed. The linear and attentive probes are largely insensitive to K, suggesting that the default of two local views already provides sufficient cross-modal regularisation; we did not see a clear benefit from increasing K.

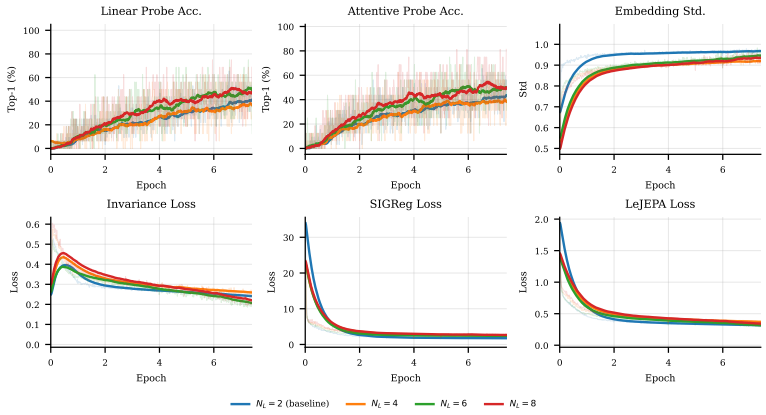


Figure A11: LeJEPA hyperparameter sensitivity: number of local views K on VGGSound pretraining, sweeping $K \in \{2, 4, 6, 8\}$ with K=2 as the paper's default. Curves are clipped to the shortest run (K=8, 7 epochs).

C.5 Loss-Component and Pipeline Ablations

We additionally run three hyperparameter ablations of the VGGSound training procedure, each with one change from the 1-GPU ablation baseline (ViT-B, batch size 32, λ=0.05, modality dropout and video tubelet/frequency-time masking, attentive probe). All runs share the same wall-clock budget and reach 7 epochs of VGGSound pretraining; figures are clipped to the common range. The probe accuracies of the two loss-term ablations are reported in Fig. 4 (right), from longer reruns; here we plot the LeJEPA loss components and the embedding standard deviation, which expose the mechanism: with λ=0 the embedding collapses to a point and with λ=1 views never align.

C.5.1 Removing the SIGReg term (λ=0)

Fig. A12 sets λ=0, training only on the invariance loss without the SIGReg regulariser. The invariance loss drops to $\sim 10^{-7}$ almost immediately and the embedding standard deviation collapses from the baseline trajectory towards $\sim 10^{-4}$, both consistent with the trivial constant-output solution. The SIGReg loss itself, although not optimised, rises far above the baseline range, confirming that the resulting embedding distribution is arbitrarily far from the target isotropic Gaussian. This isolates SIGReg as the component that prevents representation collapse in the audio-visual setting, mirroring the role it plays in LeJEPA on images.

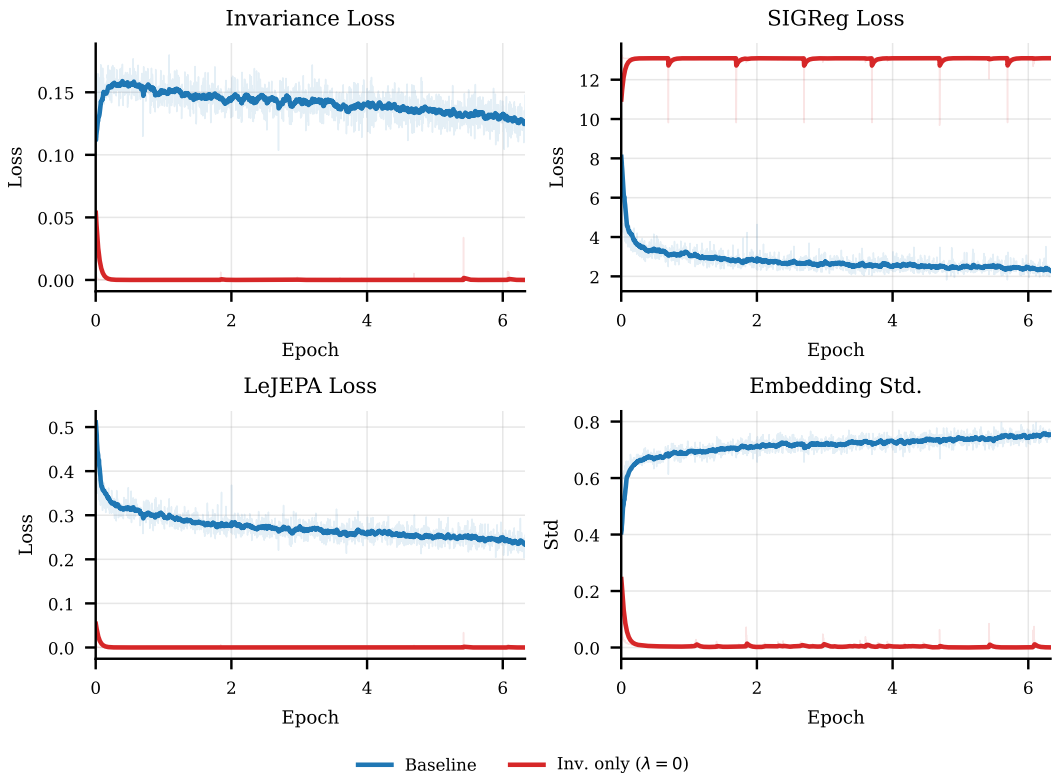


Figure A12: Removing SIGReg (λ=0) vs. the baseline (λ=0.05) on VGGSound pretraining. Without the SIGReg term, the invariance loss collapses to zero and the embedding standard deviation drops several orders of magnitude below the baseline, indicating representation collapse.

C.5.2 REMOVING THE INVARIANCE TERM (λ=1)

Fig. A13 sets λ=1, training only on SIGReg. The embedding standard deviation reaches the target $\sim$1, confirming that SIGReg alone can shape the distribution, but the invariance loss stays roughly an order of magnitude above the baseline throughout training: there is no signal pulling different views of the same clip together. This isolates the invariance loss as the component responsible for cross-modal alignment.

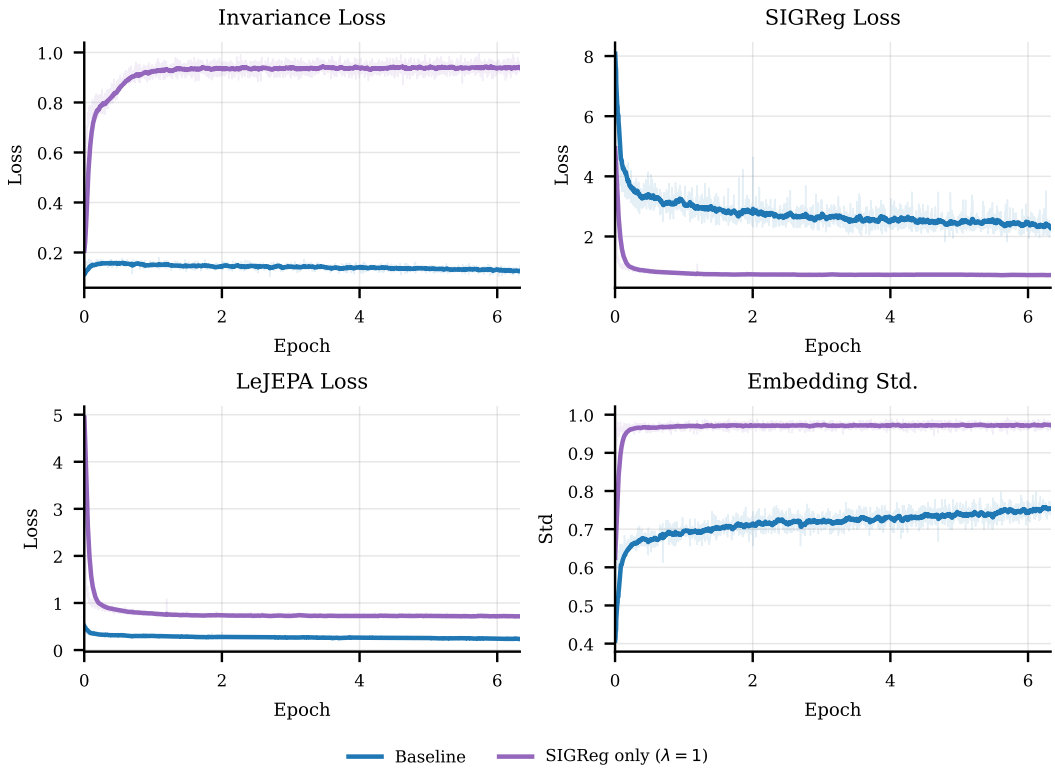


Figure A13: Removing the invariance loss (λ=1) vs. the baseline (λ=0.05) on VGGSound pretraining. SIGReg alone drives the embedding standard deviation to its target, but the invariance loss never decreases, indicating that the model never learns to align global and local views without the invariance term.

C.5.3 Removing modality dropout and masking

Fig. A14 disables modality dropout, tube masking, and frequency/time masking (p_{drop}=0, r_{vid}=0, r_{aud}=0). Local views then differ from the global views only through random resized crops and horizontal flips, while always retaining both modalities. The invariance loss drops one to two orders of magnitude below the baseline trajectory because the alignment task is trivialised: there is no partial-view perturbation forcing the model to predict the missing modality from the surviving one. SIGReg behaves similarly to the baseline, so the embedding distribution is still well shaped, but the LeJEPA loss as a whole loses its information bottleneck. This suggests that removing these partial-view perturbations makes alignment easier; the single-factor modality-dropout controls appear in Sec. 4.1.

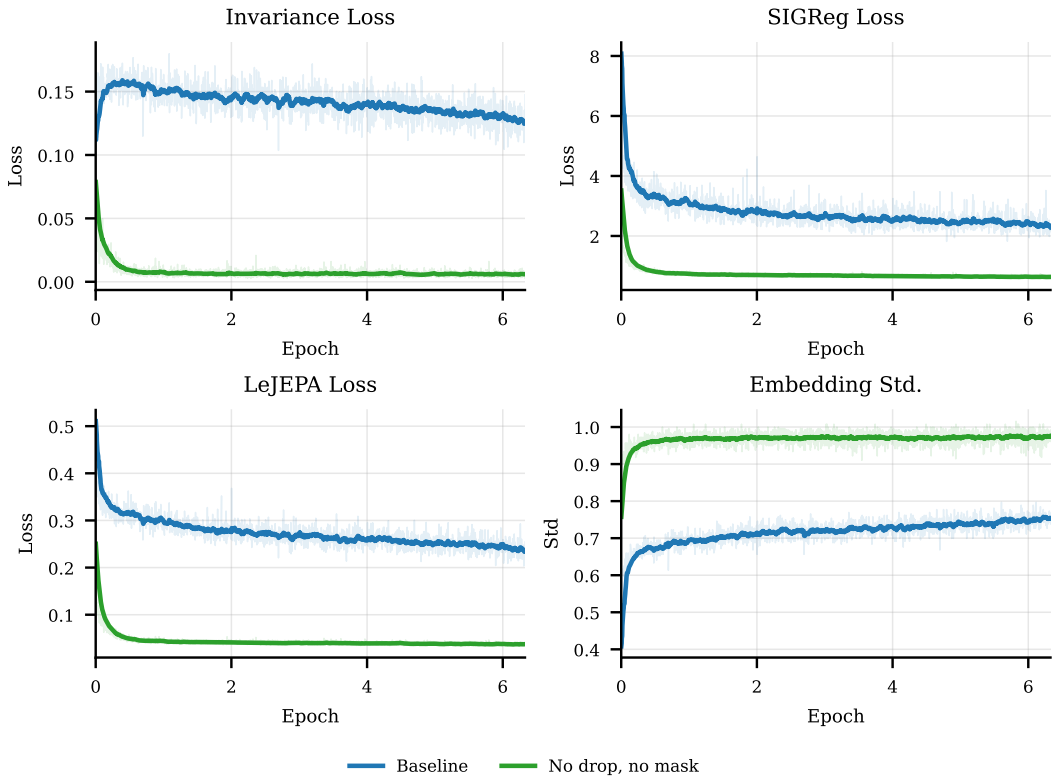


Figure A14: Removing modality dropout and tube/freq-time masking vs. the baseline on VGGSound pretraining. Without partial-view perturbations the invariance loss collapses to near zero, indicating that alignment between global and local views becomes trivial when both modalities are always present and unmasked.