Preserving Unique Concepts in Different Modalities for Contrastive Learning: Improvements to Contrastive Audio-Visual Masked Autoencoder

Anonymous authors
Paper under double-blind review

Abstract

Contrastive learning is widely adopted in multimodal model training to narrow the distance between cross-modal feature representations, facilitating multimodal information fusion and unified representation learning. During training, contrastive loss guides the model to capture matched features of the same concept across different modalities. Nevertheless, it also causes the model to discard modality-specific conceptual information that exists solely within a single modality and cannot be shared cross-modally. To address this limitation, we decouple each modality's features into shared conceptual components and modality-exclusive components. Cross-modal contrastive alignment is only imposed on shared components, which preserves modality-specific semantics while sustaining the model's cross-modal reasoning capacity. Our experiments verify that globally aligning full features via contrastive loss degrades downstream task accuracy, whereas retaining modality-unique concepts yields steady performance improvements. Built upon the audio-visual multimodal model CAV-MAE, we redesign its contrastive learning module to preserve exclusive conceptual information from both visual and audio modalities. Experiments across multiple datasets demonstrate that our method outperforms the vanilla baseline on downstream task. The results validate the rationality and effectiveness of our approach for audio-visual multimodal learning. Further analyses confirm that audio and visual modalities contain distinct exclusive concepts, which substantially boost the model's general representation capability and downstream performance. Overall, our work proves the efficacy of preserving modality-specific conceptual information, especially for audio-visual multimodal tasks.

1 Introduction

Multimodal representation learning aims to integrate information from diverse modalities to construct a unified semantic space, where cross-modal contrastive learning has been widely adopted to capture and align multimodal features. Traditional contrastive learning paradigms typically force the global representations of samples from different modalities closer to each other, compelling the model to focus on cross-modal shared semantic concepts. However, this global feature alignment mechanism has an inherent limitation: it drives the encoders of each modality to over-focus on overlapping shared concepts between modalities, while suppressing or even discarding unique concept information that exists solely within a single modality. In practice, these modality-unique features—such as ambient background noise in audio or fine-grained visual textures in video—are equally critical for enhancing the foundational representation capability of the model and facilitating downstream complex task understanding.

In the audio-visual fusion field, this issue of modality information suppression remains prevalent. As a representative paradigm for audio-visual representation learning, self-supervised

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.

models combining contrastive learning and masked autoencoding (such as CAV-MAE Gong et al. (2022)) have achieved strong cross-modal comprehension. Nonetheless, during its contrastive learning stage, CAV-MAE forcibly aligns the global features produced by the audio and visual encoders, leading the model to neglect modality-unique concepts that provide essential supplementary cues for downstream semantic recognition. Overcoming the limitations of global alignment, that is, preserving modality-unique semantics fully while maintaining strong cross-modal associations, remains a major challenge in audio-visual pre-training.

To address the aforementioned issues, we introduce a Concept Dictionary mechanism along with a disentangled contrastive learning framework that reconciles "shared concept alignment" with "unique concept preservation." By mapping and factorizing continuous feature spaces into combinations of discrete concept vectors, the concept dictionary provides an effective avenue for interpretable feature disentanglement. Building upon this, we explicitly disentangle the audio-visual representations extracted by CAV-MAE into cross-modal shared concepts and modality-unique concepts. In our refined contrastive learning pipeline, the model enforces selective alignment exclusively on the shared concept components while preserving the unique concept components. This mechanism significantly enhances the completeness and richness of single-modality representations without compromising cross-modal semantic understanding.

We conduct extensive experiments on the VGGSound and AudioSet-20K datasets. Experimental results demonstrate that our proposed method outperforms the baseline (CAV-MAE) across downstream tasks. Furthermore, through interpretability analysis enabled by the concept dictionary, we empirically confirm the existence of modality-unique concepts in audio-visual modalities and validate the necessity of preserving them to boost foundational model capacity.

The main contributions of this paper are summarized as follows:

- A New Perspective on Modality Information Preservation: We reveal the inherent shortcoming of "single-modality concept suppression caused by global alignment" in existing audio-visual contrastive learning frameworks, and validate the importance of preserving modality-unique concepts for enhancing representation capacity via a concept dictionary mechanism.
- A Concept-Disentangled Contrastive Learning Framework: We refine the traditional contrastive learning pipeline by explicitly factorizing audio-visual features into shared and unique concepts via a concept dictionary, achieving selective alignment of cross-modal shared semantics alongside effective preservation of modality-unique semantics.
- Superior Performance and Interpretability Validation: Built upon the CAV-MAE architecture, our method outperforms the baseline CAV-MAE across downstream tasks on both VGGSound and AudioSet-20K datasets, while conferring stronger contextual interpretability to the model.

2 Related Work

2.1 Audio-Visual Multimodal Pre-training

In the audio-visual multimodal domain, due to the high cost of acquiring annotated data, self-supervised learning has become the primary avenue for constructing universal multimodal representations. Existing approaches typically utilize pretext tasks to mine supervisory signals from unlabeled data. Among them, CAV-MAE Gong et al. (2022) serves as a representative foundation model in the audio-visual field, successfully combining contrastive learning with masked autoencoders (MAE) He et al. (2022). It aligns cross-modal features by drawing audio and visual representations closer via contrastive learning, while simultaneously learning rich multimodal representations by reconstructing masked patches in audio and visual sequences.

Building upon CAV-MAE, numerous studies have extended and optimized its framework. Several works focus on architectural and modular improvements, such as Araujo et al. (2025); Huang et al. (2022); Georgescu et al. (2023), which enhance foundational representations' performance by introducing different balancing mechanisms or asymmetric decoders. Other works, such as Wang et al. (2024) , emphasize parameter-efficient fine-tuning methods, enabling CAV-MAE to be transferred to various downstream tasks rapidly and at a low cost. Furthermore, extensive research has successfully applied CAV-MAE to diverse downstream applications, including dynamic facial social analysis Bohy et al. (2024), audio effect editing Guo et al. (2026), temporal action localization Ishikawa et al. (2025), and integration with large language models (LLMs). These studies fully demonstrate the central role of CAV-MAE in audio-visual multimodal learning. Unlike the existing works, we aim to improve the contrastive learning mechanism within CAV-MAE to enhance its downstream transfer and generalization capabilities without introducing extra parameters during testing.

2.2 Cross-Modal Contrastive Learning

Contrastive learning is one of the core paradigms of self-supervised representation learning, aimed at constructing highly discriminative representation spaces by pulling positive sample pairs together and pushing negative sample pairs apart. In the multimodal domain, works such as Radford et al. (2021); Li et al. (2021); Elizalde et al. (2023) widely adopt contrastive learning to achieve alignment across modalities, including image-text, video-text, and audio-text. Similarly, CAV-MAE relies on global contrastive learning to accomplish semantic alignment between audio and visual modalities. While Kalantidis et al. (2020); Dong et al. (2024) focus on sample-level positive and negative sample mining and synthesis, the existing contrastive learning strategies Wang et al. (2026); Kusupati et al. (2022); Zhang et al. (2024) mostly concentrate on pulling global or local shared information closer across modalities, often overlooking the unique concept semantics within individual modalities.

2.3 Concept Dictionaries and Sparse Autoencoders

In recent years, extracting interpretable features from internally neural activations via dictionary learning has emerged as a frontier in mechanistic interpretability research. Representative works, such as Templeton et al. (2026); Hao et al. (2022), and related SAE studies Gao et al. (2025), utilize Sparse Autoencoders (SAE) to project and decompose compact, polysemantic neural activations into an extremely high-dimensional sparse space. This automatically disentangles millions of single-semantic "concept features," constructing implicit concept dictionaries internal to the models.

To build more accurate concept dictionaries, subsequent research has introduced architectural variants such as Rajamanoharan et al. (2024a;b), which significantly improve concept extraction accuracy and sparsity. Inspired by these mechanistic interpretability studies, this paper incorporates a concept dictionary mechanism into the contrastive learning framework of audio-visual multimodal models. We introduce BatchTopK-SAE Bussmann et al. (2024) to construct concept dictionaries for both audio and visual modalities, explicitly factorizing multimodal features into "cross-modal shared concepts" and "modality-unique concepts." While maintaining alignment for cross-modal shared semantics, this effectively protects single-modality specific concepts that are suppressed or eliminated in traditional global contrastive learning, thereby significantly boosting the model's performance on downstream tasks.

3 Methodology

In traditional audio-visual multimodal learning, contrastive learning is often used to bring global features of audio and visual modalities closer together within a unified space. This enables the model to understand and integrate cross-modal expressions of the same concept by extracting shared concept features across modalities. However, applying contrastive learning to holistic multimodal features easily leads to the suppression or neglect of unique single-modality concepts.

To mitigate this issue, we introduce a Concept Dictionary Module prior to contrastive learning. This module explicitly decomposes features from each modality into cross-modal shared concept representations and single-modality unique concept representations, enabling selective contrastive alignment. By preserving the model's understanding of shared multimodal concepts while retaining modality-specific concepts, our approach improves fine-tuning performance on downstream tasks.

3.1 CAV-MAE Architecture and Contrastive Learning

Our methodology is built upon CAV-MAE, a widely utilized foundational self-supervised audio-visual pre-training architecture. Given an input audio spectrogram A and video frames V, modality-specific encoders E_A, E_V and a multimodal shared encoder E_{joint} extract the corresponding multimodal representations:

$$\mathbf{F}_A = E_{joint}(E_a(X_a)) \in \mathbb{R}^{N_A \times D} \tag{1}$$

$$\mathbf{F}_V = E_{joint}(E_v(X_v)) \in \mathbb{R}^{N_V \times D} \tag{2}$$

where N_A and N_V denote the number of spatio-temporal patches in the audio and visual sequences, respectively, and D represents the hidden feature dimension.

In the standard CAV-MAE training paradigm, the model is jointly optimized using masked autoencoding (MAE) and global cross-modal contrastive learning. The contrastive loss $\mathcal{L}_{\text{cl}}$ forces globally pooled features of audio and visual to be close in distance:

$$\mathcal{L}_{\text{cl}} = \text{InfoNCE}(\bar{F}_A, \bar{F}_V) \tag{3}$$

where $\bar{F}_m$ ($m \in \{A, V\}$) denotes the global feature obtained by averaging over the sequence dimension. However, forcing global representations to align assumes that all feature dimensions serve cross-modal shared semantics. This implicitly suppresses and ignores modality-unique conceptual semantics, such as background noise in audio or fine-grained textures in the visual modality.

3.2 Direct Feature Interval Partitioning

A feature vector can generally be viewed as a linear combination of various semantic concept information. To retain modality-specific concepts without introducing extra trainable parameters, an intuitive, naive strategy is to hard-split the feature vector along dimension intervals into a cross-modal shared concept part F_m^s and a single-modality unique concept part F_m^u:

$$F_m = [F_m^s \parallel F_m^u] \tag{4}$$

where $[\,\|\,]$ denotes feature concatenation along the channel dimension, $F_m^s \in \mathbb{R}^{N_m \times D_s}$, and $F_m^u \in \mathbb{R}^{N_m \times D_u}$.

Under this interval partitioning scheme, cross-modal contrastive learning operates selectively only on the shared feature components:

$$\mathcal{L}_{\text{cl_split}} = \text{InfoNCE}(\bar{F}_A^s, \bar{F}_V^s) \tag{5}$$

Meanwhile, the unique feature component F_m^u receives no contrastive alignment constraints and relies entirely on CAV-MAE's original masked data reconstruction task for implicit gradient updates. This design forces the model to focus on cross-modal alignment within F_m^s while naturally retaining modality-specific information in F_m^u.

3.3 Concept Dictionary and Feature Factorization

Although direct feature splitting provides a straightforward disentanglement intuition, the complex non-linear coupling across dimensions in continuous representation spaces means hard partitioning can easily damage underlying semantic integrity. In contrast, a concept dictionary views features as a collection of distinct concepts and, via a learnable dictionary matrix, decomposes input features into a combination of multiple concept vectors. To

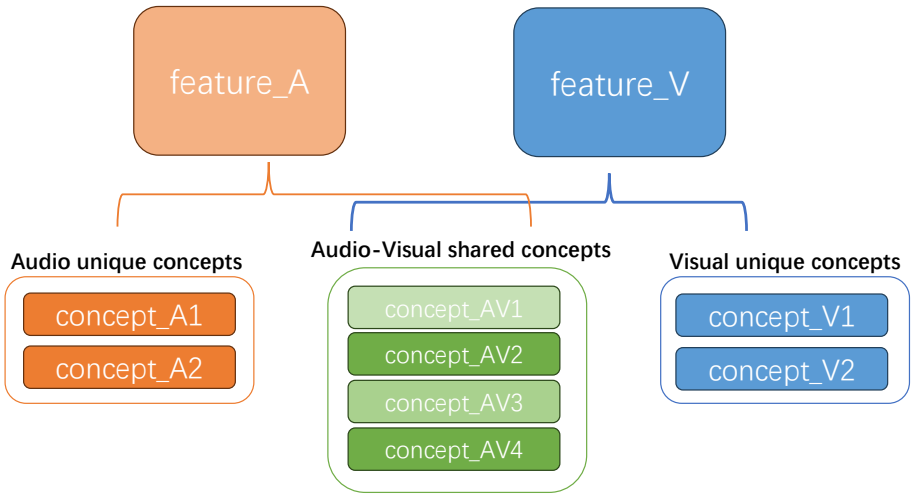


Figure 1: Overview of Concept Decomposition Across Modalities

address this, we further introduce a concept dictionary module based on BatchTopK Sparse Autoencoders (BatchTopK-SAE), which projects and decomposes the continuous feature space into a sparse combination of discrete concept vectors.

We define a learnable concept dictionary matrix $D_{\text{dict}} \in \mathbb{R}^{K \times D}$, where K is the number of dictionary basis concept vectors ($K \gg D$, representing an overcomplete dictionary). For any modality feature matrix $F \in \mathbb{R}^{N \times D}$, its factorization and reconstruction through the concept dictionary is defined as:

$$\boldsymbol{Z} = \text{BatchTopK}\big(\text{F}D_{\text{dict}}^{\top}\big), \quad \hat{F} = D_{\text{dict}}\boldsymbol{Z} \tag{6}$$

where $Z \in \mathbb{R}^{N \times K}$ represents the sparse activation coefficient matrix over the discrete concept bases. The BatchTopK activation function ensures that only the top-k most relevant core concepts across the batch are retained, achieving high-coherence, interpretable concept decomposition.

Combining this with the interval partitioning concept from the above section, we partition the disentangled concept activation matrix Z into two irrelevant concept subspaces: Cross-modal Shared Concept Group $Z^s \in \mathbb{R}^{N \times K_s}$ and Single-modality Unique Concept Group $Z^u \in \mathbb{R}^{N \times K_u}$, where $K_s + K_u = K$.

Accordingly, the cross-modal shared semantic features reconstructed from the concept dictionary are expressed as:

$$\hat{F}^s = Z^s D^s_{\text{dict}} \tag{7}$$

Following concept dictionary feature disentanglement, we constrain the contrastive learning objective to the reconstructed shared concept feature space. The new selective contrastive loss $\mathcal{L}_{\text{cl_sc}}$ is formulated as:

$$\mathcal{L}_{\text{cl_sc}} = \text{InfoNCE}(\bar{\hat{F}}^s_A, \bar{\hat{F}}^s_V) \tag{8}$$

By selectively aligning only the shared concept reconstruction features $\hat{F}^s_A$ and $\hat{F}^s_V$, the model avoids disrupting the information conveyed by the unique concept group Z^u. Consequently, during self-supervised pre-training, the model achieves strong alignment of shared semantics while keeping unique semantics intact.

Table 1: Performance comparison (Acc↑) of direct feature splitting and concept splitting under different shared ratios (r) on the VGGSound dataset.

Method / Strategy	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Mean ± Std
CAV-MAE (Baseline)	59.44	59.54	59.86	59.34	59.44	59.52 ± 0.20
Direct Feature Splitting						
Shared Ratio $r = 0.75$	59.32	59.64	59.98	59.62	59.83	59.68 ± 0.25
Shared Ratio $r = 0.50$	59.68	59.69	59.60	59.24	59.56	59.55 ± 0.18
Concept Splitting						
Shared Ratio $r = 0.75$	60.76	60.59	60.64	60.54	60.54	60.61 ± 0.09
Shared Ratio $r = 0.50$	60.55	60.47	60.54	60.23	60.32	60.42 ± 0.14

Table 2: Performance comparison (mAP↑) of direct feature splitting and concept splitting under different shared ratios (r) on the AS20K dataset.

Method / Strategy	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Mean ± Std
CAV-MAE (Baseline)	0.3286	0.3325	0.3321	0.3316	0.3285	0.3307 ± 0.0020
Direct Feature Splitting						
Shared Ratio $r = 0.75$	0.3293	0.3344	0.3320	0.3221	0.3352	0.3306 ± 0.0053
Shared Ratio $r = 0.50$	0.3335	0.3266	0.3266	0.3254	0.3279	0.3280 ± 0.0032
Concept Splitting						
Shared Ratio $r = 0.75$	0.3373	0.3379	0.3359	0.3350	0.3358	0.3364 ± 0.0012
Shared Ratio $r = 0.50$	0.3368	0.3350	0.3350	0.3338	0.3338	0.3349 ± 0.0016

3.4 Joint Optimization Objective

Alongside the selective contrastive loss $\mathcal{L}_{\text{cl_sc}}$, our model retains CAV-MAE's original reconstruction objective to include the audio and visual masked autoencoding reconstruction loss $\mathcal{L}_{\text{mae}}$ along with the sparse reconstruction loss $\mathcal{L}_{\text{sae}}$ of the BatchTopK-SAE concept dictionary itself.

The final end-to-end joint optimization loss function $\mathcal{L}_{\text{total}}$ for pretraining is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mae}} + \lambda_1 \mathcal{L}_{\text{cl_sc}} + \lambda_2 \mathcal{L}_{\text{sae}} \tag{9}$$

where λ_1 and λ_2 are hyperparameters balancing the contributions of individual loss terms. For downstream fine-tuning, the concept dictionary and modality-specific encoders are kept frozen, while extra classification heads are trained to handle different downstream tasks.

4 Experiments

4.1 Experimental Setup

To eliminate performance variations caused by different settings in training environments, optimizer configurations, and initialization, while ensuring fairness and reliability across comparative experiments, all evaluated models strictly adhere to identical experimental settings. These include key hyperparameters such as epoch count, batch size, initial learning rate, and masking ratio, as well as maintaining identical backbone network architectures and

Table 3: Performance comparison under different shared feature ratios (r) on the VGGsound dataset.

Method / Strategy	Acc ↑
CAV-MAE	59.44
Direct Feature Splitting	
Shared Ratio $r = 0.25$	59.52
Shared Ratio $r = 0.50$	59.56
Shared Ratio $r = 0.75$	59.83

Table 4: Performance comparison with different shared concepts ratios (r) on the AS20K dataset.

Method / Strategy	mAP ↑
CAV-MAE	0.3286
Concept Splitting	
Shared Ratio $r = 0.25$	0.3359
Shared Ratio $r = 0.50$	0.3368
Shared Ratio $r = 0.75$	0.3373

pre-trained weight initializations. Furthermore, to eliminate the impact of random factors (e.g., data shuffling and mask sampling), we conduct multiple independent runs across different random seeds, reporting the mean performance to guarantee rigorous conclusions and high reproducibility.The dictionary size is set to $768 \times 32 = 24576$. The sparsity of the main branch and auxiliary branch are controlled by top_k = 32 and top_k_aux = 16, respectively.

During the downstream event classification fine-tuning stage, to address class imbalance in multi-label and multi-class data distributions, we adopt a balanced sampling policy consistent with the baseline configuration to adjust class weights.

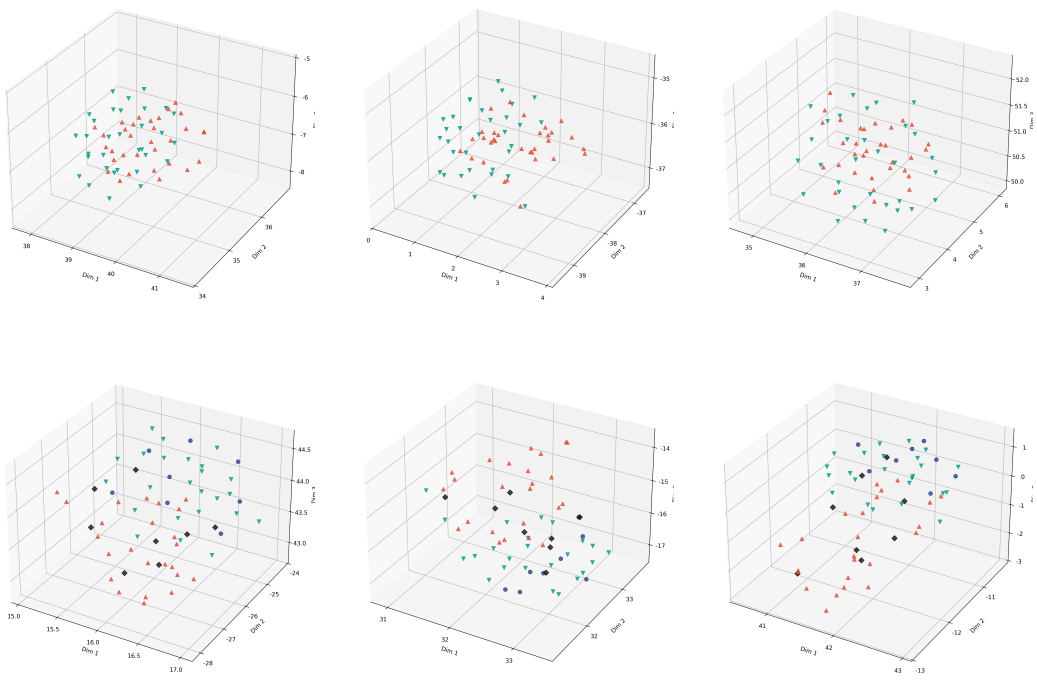
4.1.1 Datasets & Evaluation Metrics

For fair and intuitive comparison, we adopt the experimental protocol of CAV-MAE and perform evaluations on VGGSound and AudioSet-20K to directly benchmark our method against the original model.

- VGGSound Chen et al. (2020): We perform self-supervised pre-training on the VGGSound training set and evaluate video event classification performance on its fine-tuning set. VGGSound is a large-scale audio-visual dataset containing approximately 200,000 10-second video clips sourced from YouTube. We follow the standard CAV-MAE data processing pipeline to uniformly sample and preprocess input video frames and audio spectrograms.

- AudioSet-20K (AS20K) Gemmeke et al. (2017): AS20K is a standard benchmark in the audio-visual multimodal field. Following the setting of CAV-MAE, we utilize a balanced subset of approximately 20,000 video clips for fine-tuning and evaluation.

- Evaluation Metrics: Following the standard evaluation protocol of CAV-MAE, we employ Top-1 Accuracy (Acc) to evaluate classification performance on VGGSound, and mean Average Precision (mAP) to evaluate multi-label classification performance on AS20K.



(a) video-1 (b) video-2 (c) video-3

Figure 2: Spatial distribution of multimodal concepts across samples. Top row: conceptual feature spaces extracted by the original model. Bottom row: corresponding feature spaces after concept disentanglement training.
Symbols: ▲ audio concepts, ▼ visual concepts, ♦ audio unique concepts, and ● visual unique concepts.

4.2 Main Comparison Results

As shown in Table 1 and Table 2, our proposed direct feature splitting method (Direct Splitting) outperforms the original CAV-MAE on VGGSound, while the latent feature disentanglement method based on concept dictionaries (Concept Splitting) significantly achieves superior performance across all datasets over the original CAV-MAE under identical experimental settings and multi-seed evaluations.

Based on the experimental results, we highlight two key findings regarding the effectiveness of explicit feature disentanglement: First, direct feature split can improve the performance of CAE-MAE to some extent, not considerable but consistently existing, while concept-driven feature split leads to obvious performance improvement in contrast, validating the existence of the critical limitation that is, uniformly enforcing global feature alignment degrades single-modality's unique semantics. This confirms the technical soundness of separating cross-modal shared concepts from modality-unique concepts. Second, comparing the two strategies demonstrates the superiority of concept dictionary disentanglement. The feature factorization based on BatchTopK-SAE consistently surpasses the hard splitting approach (Direct Splitting). This indicates that complex non-linear couplings across continuous feature dimensions cannot be cleanly disentangled via naive hard slicing, whereas sparse projection via concept dictionaries retains semantic concepts more naturally and completely.

It is particularly worth emphasizing that our proposed concept factorization and selective contrastive learning operate solely during the self-supervised pre-training stage. During downstream task fine-tuning and inference, the concept dictionary module is completely excluded from forward computation. Consequently, our approach enhances the model's

Table 5: Accuracy comparison with different shared feature ratios (r) without MAE loss on the VGGSound dataset.

Method / Strategy	Acc ↑
CAV	37.31
Shared Feature Ratio (r)	
$r = 0.50$	39.85
$r = 0.75$	38.71

representation capabilities for downstream tasks without introducing any extra model parameters or additional computational overhead.

4.3 Ablation Study

4.3.1 Sensitivity Analysis of Unique Concept Ratios

Table 3 and 4 compare model performance across different unique concept partitioning ratios (r). The results show that excessively aggressive partitioning of unique concepts leads to insufficient shared concept representations, thereby weakening cross-modal alignment. Conversely, an overly small unique concept ratio fails to preserve modality-specific semantics effectively. A well-balanced factorization ratio retains modality independence while reinforcing shared concept alignment.

4.3.2 Independent Effectiveness of Selective Contrastive Learning

To verify the independent impact of selective contrastive learning in the absence of reconstruction constraints, we remove the masked autoencoding (MAE) task from CAV-MAE and pre-train the model solely with the contrastive learning objective. As shown in Table 5, when the MAE reconstruction task is removed and only the contrastive learning setup is retained, model performance drops significantly compared to the original version employing both self-supervised tasks. Nevertheless, our concept-driven contrastive learning method still outperforms the baseline CAV model that uses a global feature alignment strategy. This demonstrates that our method can independently enhance the quality of semantic representations in contrastive learning, validating the generalizability and feasibility of the selective alignment strategy across multimodal learning paradigms.

5 Conclusion

In this paper, we presented a novel selective contrastive learning framework for audio-visual multimodal representation learning based on concept factorization. Addressing the limitation of conventional paradigms where global feature alignment suppresses modality-unique semantics, we introduced a concept dictionary mechanism to explicitly disentangle multimodal features into modality-shared and private concept components. This selectively aligns cross-modal shared semantics while preserving modality-unique information intact.

Extensive experiments demonstrate that, under strict variable control and without introducing extra parameters or computational overhead during inference, our approach consistently outperforms the baseline CAV-MAE model across downstream tasks on VGGSound and AudioSet-20K. These findings confirm the theoretical correctness and practical utility of maintaining explicit semantic distinctions across modalities to boost foundational representation capacity. In future works, we plan to extend this concept disentanglement paradigm to broader multimodal domains, such as vision-language and audio-language learning.

6 AI disclosure

In this work, we used generative AI tools to assist with translation. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results. Generating synthetic datasets, formulating mathematical claims, providing critical ingredients for proving mathematical claims, and assisting in the writing of proofs are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

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 Computer Vision and Pattern Recognition Conference, pp. 18794–18803, 2025.

Hugo Bohy, Minh Tran, Kevin El Haddad, Thierry Dutoit, and Mohammad Soleymani. Social-mae: A transformer-based multimodal autoencoder for face and voice. In 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG), pp. 1–5. IEEE, 2024.

Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. arXiv preprint arXiv:2412.06410, 2024.

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

Hengkui Dong, Xianzhong Long, and Yun Li. Synthetic hard negative samples for contrastive learning. Neural Processing Letters, 56(1):33, 2024.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE, 2023.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In International Conference on Learning Representations, volume 2025, pp. 26721–26754, 2025.

Jort F Gemmeke, Daniel PW 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 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 776–780. IEEE, 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, pp. 16144–16154, 2023.

Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. Contrastive audio-visual masked autoencoder. arXiv preprint arXiv:2210.07839, 2022.

Xinyue Guo, Xiaoran Yang, Lipan Zhang, Jianxuan Yang, Zhao Wang, and Jian Luan. Av-edit: Multimodal generative sound effect editing via audio-visual semantic joint control. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 40, pp. 21504–21512, 2026.

Yuhan Hao, Tim Stuart, Madeline Kowalski, Saket Choudhary, Paul Hoffman, Austin Hartman, Avi Srivastava, Gesmira Molla, Shaista Madad, Carlos Fernandez-Granda, et al. Dictionary learning for integrative, multimodal, and scalable single-cell analysis. Biorxiv, pp. 2022–02, 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, pp. 16000–16009, 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. arXiv preprint arXiv:2212.08071, 2022.

Yuchi Ishikawa, Shota Nakada, Hokuto Munakata, Kazuhiro Saito, Tatsuya Komatsu, and Yoshimitsu Aoki. Language-guided contrastive audio-visual masked autoencoder with automatically generated audio-visual-text triplets from videos. arXiv preprint arXiv:2507.11967, 2025.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. Advances in neural information processing systems, 33:21798–21809, 2020.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. Advances in Neural Information Processing Systems, 35:30233–30249, 2022.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems, 34: 9694–9705, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pp. 8748–8763. PmLR, 2021.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders. arXiv preprint arXiv:2404.16014, 2024a.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. arXiv preprint arXiv:2407.14435, 2024b.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. arXiv preprint arXiv:2605.29358, 2026.

Kai Wang, Yapeng Tian, and Dimitrios Hatzinakos. Towards efficient audio-visual learners via empowering pre-trained vision transformers with cross-modal adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1837–1846, 2024.

Ziwen Wang, Jiajun Fan, Thao Nguyen, Heng Ji, and Ge Liu. Variational supervised contrastive learning. Advances in neural information processing systems, 38:14184–14223, 2026.

Peiquan Xu, Xinhua Tang, and Shun Yao. Application of circular laser vision sensor (clvs) on welded seam tracking. Journal of Materials Processing Technology, 205(1-3):404–410, 2008.

Shijie Zhang, Bin Zhang, Yuntao Wu, Huabing Zhou, Junjun Jiang, and Jiayi Ma. Segclip: Multimodal visual-language and prompt learning for high-resolution remote sensing semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing, 62:1–16, 2024.