

# MultiMem: Measuring and Mitigating Memorization in Multi-Modal Contrastive Learning

Wenhao Wang, Franziska Boenisch, Michael Backes, Adam Dziedzic  
{wenhao.wang, boenisch, backes, adam.dziedzic}@cispa.de  
CISPA Helmholtz Center for Information Security

## Abstract

*Memorization in machine learning models enables high performance on rare in-distribution samples by capturing their atypical patterns. However, it also causes harmful retention of noise and outliers, degrading generalization. While memorization has been extensively studied in both supervised and self-supervised learning in the vision domain, it remains unexplored in multi-modal contrastive learning. We address this gap by introducing MultiMem, the first metric designed to quantify memorization in multi-modal contrastive learning. Through our systematic analysis, we demonstrate that cross-modal semantic misalignment has the strongest influence on memorization, with text being the dominant modality driving memorization, followed by video, image, and audio. We show that targeted augmentations applied across all modalities effectively reduce memorization as measured by our MultiMem metric and improve model performance. Overall, this work establishes the first framework for measuring and mitigating memorization in multi-modal contrastive learning, preventing harmful data retention and contributing to higher-performing models.*

## 1. Introduction

Multi-modal contrastive learning aims to jointly process and align data from diverse modalities such as images, text, audio, and video. This paradigm has demonstrated high performance across a wide range of tasks, including image captioning [11, 17, 37], visual question answering [25, 29, 33], zero-shot classification [19, 22, 36], and cross-modal retrieval [1, 2, 30]. These successes underscore the benefits of integrating heterogeneous modalities, which lead to improved generalization and more robust semantic representations. However, it remains unclear to what extent memorization contributes to the observed improvements, motivating the need to better understand the role of memorization

in multi-modal contrastive learning. Previous studies have shown that in both supervised learning (SL) [5, 6] and self-supervised learning (SSL) [26], model’s memorization of training data points is essential for generalization. In the vision domain, it has been observed that models tend to memorize outliers in the training set, which correspond to mislabeled samples in SL [3, 5] and atypical examples in SSL [27]. A similar study of memorization in multi-modal contrastive learning is lacking.

Multi-modal learning introduces unique challenges not seen in uni-modal settings [13–15], such as inconsistencies between modalities, varying noise levels, and the lack of explicit modality-alignment. These factors limit the direct transfer of insights from uni-modal studies to multi-modal learning. Importantly, most existing definitions of memorization are designed for specific modalities and tasks, for example, based on label prediction in SL [5] or augmentation matching in SSL [27] for the vision domain. Such definitions do not generalize well to the multi-modal setting: Measuring memorization within individual modalities or among a limited subset of all the modalities used to train a given multi-modal model, fails to capture multi-modal memorization *faithfully*. Therefore, a new definition tailored to multi-modal contrastive learning is needed.

There are only two studies on measuring memorization for the bi-modal image-text models like CLIP [22], namely *déjà vu* memorization [12] and *CLIPMem* [28]. The *déjà vu* method measures memorization by masking one modality and testing whether the model can recover the other, while *CLIPMem* measures memorization by comparing alignment between model pairs trained with or without a given sample (*i.e.*, an image-text pair). However, these methods are not directly applicable to multi-modal contrastive learning which involves additional modalities, such as video and audio.

To address these limitations, we propose **MultiMem: a general-purpose metric for measuring memorization in multi-modal contrastive learning**, which is designed for any number and type of modalities involved. Specifically, MultiMem builds on the leave-one-out framework [5, 27, 34] for measuring memorization and compares the outputs of a

---

Accepted at the 19th European Conference on Computer Vision (ECCV), 2026.

pair of multi-modal models trained with and without a given multi-modal sample on this particular sample.

Through our extensive empirical study on multiple contrastively trained multi-modal models such as **Audio-CLIP** [10] (including Audio, Image, and Text modalities), **AVT-CLIP** (a custom-built tri-modal model with Audio, Video, and Text modalities, introduced in Section 4.1), and **AVIT-CLIP** (a custom-built quad-modal model with Audio, Video, Image, and Text modalities, introduced in Section 4.1), we observe that: (1) Global memorization, measured across all modalities, behaves differently from memorization observed within any subset of modalities and to capture the full extent of a model’s memorization, we need to assess memorization jointly across all its modalities. (2) The most memorized samples are not simply mis-captioned, as reported for the most memorized samples in text-image models [28], but semantically misaligned across all modalities: the information provided by different modalities not only contradict each other but is mostly semantically *not related*. (3) Multi-modal models increasingly memorize cross-modal patterns, rather than primarily text as observed in bi-modal models [28]. This suggests that multi-modal models’ memorization behavior is closer aligned with the one of SSL models [27] which were shown to memorize pattern, rather than SL models which memorize labels [5].

Based on the above observations, we explore *different strategies to mitigate memorization and improve generalization* in multi-modal models. **In-training:** we *actively measure memorization during multi-modal training* to identify the top-memorized samples. At a given stage of training, we re-group the top-memorized samples into new batches and continuously apply noise-based augmentations (only) to them in subsequent training steps. **Post-training:** we *identify highly memorized samples post-training* and then fine-tune the model on the remaining training samples. Our experiments show that both approaches **substantially reduce memorization** by up to **20%**, while increasing model performance up to **8%** for retrieval, **10%** for zero-shot, and **4%** for downstream classification tasks. In summary:

- We propose *MultiMem*, a metric that measures memorization by comparing the outputs of a pair of multi-modal models trained with and without a given multi-modal sample on this particular sample
- Our analysis with MultiMem *demonstrates the key differences between multi-modal models with more than two modalities and previously studied bi-modal or uni-modal models*, including distinct memorization patterns, a shift from label-driven to pattern-driven memorization behavior, and a stronger influence of cross-modal semantic inconsistency on memorization.
- Based on our findings, we propose *two approaches to mitigate memorization and improve generalization in multi-*

*modal models:* either during training or after training. Our extensive experiments show that both methods effectively reduce the model’s memorization level and lead to substantial improvements in performance.

## 2. Background and Related Work

**Multi-Modal Contrastive Learning.** Contrastive Language-Image Pretraining (CLIP) [22] proposed a joint training framework for image and text modalities, where text and image representations are aligned using the InfoNCE loss [20]. CLIP enables strong zero-shot image classification by using language prompts as class representations, and supports accurate image-text retrieval through a shared representation space. **VideoCLIP** [31] extends CLIP to the video domain by substituting the original image encoder with a transformer-based video encoder. The video is first uniformly sampled into a fixed number of frames, which are individually processed by a frozen convolutional neural network (CNN) with a trainable multi-layer perceptron (MLP) to obtain frame-level representations. These frame-level representations are then encoded as tokens by a transformer and aggregated by applying average pooling to obtain the representation for the entire video. This modification enables the model to process temporal-visual information while retaining a contrastive training framework. VideoCLIP shows strong performance in zero-shot video understanding tasks, such as action recognition and video-text retrieval, without task-specific supervision. **AudioCLIP** [10] extends CLIP to the audio domain by introducing an additional audio encoder based on an ESResNeXt convolutional network [9]. The audio input is first converted into a log-mel spectrogram and then projected into a shared representation space, where it is aligned with image and text representations. The model is trained using the same contrastive objective as CLIP, encouraging alignment between semantically related audio, image, and text samples, while pushing apart mismatched pairs. As the model incorporates one more modality than CLIP, the training loss is computed as the sum of the pairwise InfoNCE losses (*i.e.*, Audio-Image, Audio-Text, and Image-Text). AudioCLIP performs well in zero-shot audio classification and cross-modal retrieval tasks, outperforming previous methods that rely on task-specific supervision.

**Memorization in SL & SSL.** Although memorization has been associated with potential risks of sensitive information leakage, multiple studies [5, 23, 27] in supervised learning (SL) and self-supervised learning (SSL) suggest that memorization also plays an important role in the generalization of models. They also show that both SL and SSL models tend to memorize atypical samples during training, while the nature of these atypical samples varies. For example, in

SL, memorization often correlates with mislabeled or noisy samples, while in SSL, it tends to occur on images with rare or distinctive visual patterns [18, 26].

A common definition of label memorization in SL is the leave-one-out style definition by [6]:

$$\text{Mem}(\mathcal{A}, S, i) = \Pr_{f \leftarrow \mathcal{A}(S)} [f(x_i) = y_i] - \Pr_{g \leftarrow \mathcal{A}(S \setminus i)} [g(x_i) = y_i], \quad (1)$$

where  $\mathcal{A}$  is a training algorithm (here for models  $f$  and  $g$ ) and  $S \setminus i$  represents train set  $S$  without the data point  $(x_i, y_i)$ . In this definition, a data point is considered memorized if the model’s prediction of the point’s ground truth training label changes significantly based on whether the point was or was not used to train the model.

For SSL, where no labels are available, [27] proposed a new metric:

$$\text{SSLMem}(S, \text{Aug}, i) = \mathbb{E}_{\substack{g \sim \mathcal{A}(S \setminus i) \\ x_i', x_i'' \sim \text{Aug}(x_i)}} [d(g(x_i'), g(x_i''))] - \mathbb{E}_{\substack{f \sim \mathcal{A}(S) \\ x_i', x_i'' \sim \text{Aug}(x_i)}} [d(f(x_i'), f(x_i''))], \quad (2)$$

where  $d(\cdot, \cdot)$  represents a distance function, commonly the  $\ell_2$  distance, and  $\text{Aug}(\cdot)$  is the augmentation sets used during training. Here, the expectation captures the *alignment* of the respective models as the expected Euclidean distance between their representations of two augmentations of the same sample. Then, SSLMem is computed as the *alignment difference* between the model pairs trained with and without that sample. *Both metrics are limited to uni-modal models, making them unsuitable for evaluating memorization effects in multi-modal settings.*

**Memorization in Bi-modal Contrastive Models.** [12] proposed a retrieval-based evaluation protocol to detect *déjà vu* memorization in vision-language models. Their method assesses whether a model trained on one data subset can retrieve objects that appeared only in a different, held-out subset, thereby indicating unintended memorization across disjoint training sets. However, these approaches only offer qualitative evaluations of whether memorization occurs, without providing a quantitative measurement of how strongly a model memorizes specific data points across all modalities.

The only existing method for *quantifying* memorization across the text and vision modalities is CLIPMem [28]. It computes the difference in alignment scores between a pair of CLIP-style models, trained with and without this data point. Formally, CLIPMem is defined as:

$$\text{CLIPMem}(I, T) = \mathcal{A}_{\text{align}}(f, I, T) - \mathcal{A}_{\text{align}}(g, I, T), \quad (3)$$

where  $f$  and  $g$  are CLIP models trained on datasets with and without a data point  $(I, T)$ , consisting of an **Image** and **Text** (caption). The alignment score  $\mathcal{A}_{\text{align}}$  is computed as the cosine similarity between image and text representations, corrected by subtracting similarities to unrelated text and image samples  $(i, t)$  that were not used in training  $f$  or  $g$ .

$$\mathcal{A}_{\text{align}}(f, I, T) = \mathbb{E}_{(I', T') \sim \text{Aug}(I, T)} [\text{sim}(f_{\text{img}}(I'), f_{\text{txt}}(T'))] - \mathbb{E}_t [\text{sim}(f_{\text{img}}(I), f_{\text{txt}}(t))] - \mathbb{E}_i [\text{sim}(f_{\text{img}}(i), f_{\text{txt}}(T))], \quad (4)$$

where  $(i, t)$  is a set of randomly chosen image and text testing samples that were not used in training  $f$  or  $g$ . CLIPMem indicates that samples where caption and corresponding image do not align well (so-called **mis-captioned** samples) are most memorized. While CLIPMem is limited to measuring bi-modal memorization between **Image** and **Text** modalities, our MultiMem metric generalizes it to capture memorization among *many more diverse modalities*.

### 3. Our MultiMem Metric

The memorization metrics discussed in the previous section are tailored to uni-modal or bi-modal models and do not account for global interactions among all modalities in a multi-modal setting. We demonstrate empirically in Figure 1 that these approaches are insufficient for faithfully capturing memorization in multi-modal models. Therefore, we introduce our MultiMem metric, which is specifically designed to quantify *global* memorization across all modalities.

We build MultiMem on the leave-one-out framework [5, 27, 28] to measure memorization with respect to a pair of models,  $f$  and  $g$ . Here, model  $f$  is trained on the full training set  $S$ , while model  $g$  is trained on the subset  $S \setminus i$  where the multi-modal sample  $x_i$  was removed.

Given the complex interactions among modalities in multi-modal contrastive learning, we first require a proxy that quantifies the quality of the model’s representations of an input data point  $x_i$  which can be compared between models. Since the learning objective in multi-modal contrastive learning is to make the representations returned by the model on the different modalities of the same sample as consistent as possible, and the representations of unrelated examples as far as possible, we consider cross-modal consistency (*CMC*) on the sample itself vs. on unrelated samples as a proxy. We calculate the cross-modal consistency as follows: Given a model with  $n$  different modalities, we define the representation space  $\Phi$  as:

$$\Phi_{x_i} = \begin{bmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \\ \vdots \\ \hat{\phi}_n \end{bmatrix} \in \mathbb{R}^{n \times d}, \quad (5)$$

where  $\phi_j \in \mathbb{R}^d$  is the representation of  $x_i$  in the  $j$ -th modality, and  $\hat{\phi}_j$  denotes its  $\ell_2$ -normalized version.

Then we define cross-modal consistency  $CMC(i, H)$  on data point  $x_i$  (with respect to a held-out set  $H$ ) as:

$$CMC(i, H) = \frac{1}{2} \mathbb{E}_{(\cdot)' \sim \text{Aug}} \mathbf{1}_n^\top \left( \Phi_{x_i'} \Phi_{x_i}^\top \right) \mathbf{1}_n - \frac{1}{2} \mathbb{E}_{\substack{h \in H \\ (\cdot)' \sim \text{Aug}}} \mathbf{1}_n^\top \left( \Phi_{x_i'} \Phi_{h'}^\top \right) \mathbf{1}_n, \quad (6)$$

where  $\mathbf{1}_n$  denotes an all-ones vector of length  $n$  and  $(\cdot)' \sim \text{Aug}$  indicates that we compute an expectation over representations computed on different random augmentations of the samples, increasing stability of the metric. The held-out set are randomly picked from non-trained samples in the validation set of training datasets.

The first term of the score captures the similarities across all modality pairs within the representation space of  $x_i$ . The second term measures the distribution differences across all modality pairs between the representation space of  $x_i$  and held-out samples  $h \in H$ . Subtracting these terms yields a score that is high when the modalities of  $x_i$  are strongly aligned with each other, and weakly aligned with unrelated examples, modeling the model’s intended objective.

Finally, we define the MultiMem of data point  $x_i$  as the  $CMC$  difference between model  $f$  and model  $g$ :

$$\text{MultiMem}(i, H, f) = CMC(i, H)_{f \sim S} - CMC(i, H)_{g \sim S^i}. \quad (7)$$

Unlike prior approaches that focus on pairwise similarities between modalities, this resulting metric leverages the entire distribution of representations across all modalities. This provides a principled way to evaluate *global* memorization in multi-modal models, capturing interactions beyond single modality pairs.

## 4. Evaluating Multi-Modal Memorization

We first describe the setup for our experiments and then analyze memorization in multi-modal contrastive learning using our MultiMem.

### 4.1. Experimental Setup

**Models and Datasets.** We run our experiments on the following models: OpenCLIP [4], AudioCLIP [10], VideoCLIP [31], and our custom-built AVT-CLIP (Audio + Video + Text) and AVIT-CLIP (Audio + Video + Image + Text). The detailed encoder architectures and training datasets used are shown in Table 1 while other training details and hyperparameters are presented in Table 8 in Section A.2.

**Dataset Splitting.** Following [27], we divide each training dataset into three subsets: (1) a candidate set  $S_C$ , which is

used only for training model  $f$  and whose memorization we want to measure; (2) an independent set  $S_I$ , which is used only for training model  $g$ ; and (3) a shared set  $S_S$ , which is used for training both  $f$  and  $g$ .  $S_C$  and  $S_I$  have an equal number of training samples to ensure that  $f$  and  $g$  are trained with the same number of data points. Detailed splits are provided in Table 7 in Section A.2. We report the average MultiMem score on  $S_C$  as the memorization for model  $f$ . Next, we discuss evaluation metrics.

**Evaluation Metrics.** In prior work, retrieval tasks typically refer to settings where one modality is used to retrieve another, which we refer to as the *uni-to-uni* retrieval setting. To enable the measurement of models’ performance on more than two modalities, we introduce a **multi-to-uni** retrieval task for evaluation. In this task, representations from multiple source modalities are combined to retrieve data points from a target modality. Specifically, we compute a retrieval score by summing the pairwise cosine similarities between the target modality representation and *each* source modality representation. A higher retrieval score shows better semantic alignment between target and input modalities. A retrieval is considered successful if the ground-truth target sample appears among the top  $N$  retrieved samples with the highest retrieval scores; otherwise, it is considered a failure. The overall retrieval success rate on the test set is used to evaluate the model’s performance. In the following experiments, we set  $N = 5$  for both uni-to-uni and multi-to-uni retrieval tasks. The final retrieval accuracy is reported using TOP@5 ( $T@5$ ). Note that retrieval tasks are deterministic with no randomness in inference, so we only report the results once instead of an average with standard deviation. We further rely on **linear probing** and **zero-shot** classification tasks to assess model performance under our mitigations. For the zero-shot classification, we follow the settings introduced in the CLIP paper [22], where classification is performed by computing the similarity between label embeddings and the representations of the target modality.

### 4.2. Measuring Multi-modal Memorization

**Memorization Distribution.** We study the memorization distribution of training samples on three multi-modal models: AudioCLIP, AVT-CLIP, and AVIT-CLIP. For AudioCLIP, we quantify the memorization level with three metrics: (1) tri-modal MultiMem with all three modalities, (2) the CLIPMem (based on Image-Text pair), and (3) bi-modal MultiMem restricted to the remaining modality pairs (Audio-Image and Audio-Text). We present our results in Figure 1. Overall, MultiMem in Figure 1a is able to measure the highest level of memorization on  $S_C$  when compared with the baselines on the same AudioCLIP model. Additionally, it separates the scores for  $S_C$  and  $S_S$  significantly better, indicating a more sensitive measurement of memorization.

Model	Modality	Encoder	Training Set
CLIP	Image + Text	ViT + Transformer	COCO [16]
VideoCLIP	Video + Text	(CNN + Transformer) + Transformer	MSR-VTT [32]
AudioCLIP	Audio + Image + Text	ESResNeXt + ViT + Transformer	UrbanSound8K [24]+ Spectrogram
AVT-CLIP	Audio + Video + Text	ESResNeXt + (CNN + Transformer) + Transformer	MSR-VTT
AVIT-CLIP	Audio + Video + Image + Text	ESResNeXt + (CNN + Transformer) + ViT + Transformer	MSR-VTT + Frame-image
ImageBind-AVIT	Audio + Video + Image + Text	ESResNeXt + (CNN + Transformer) + ViT + Transformer	MSR-VTT + Frame-image

Table 1. The encoder architecture and datasets used by the models in this paper.

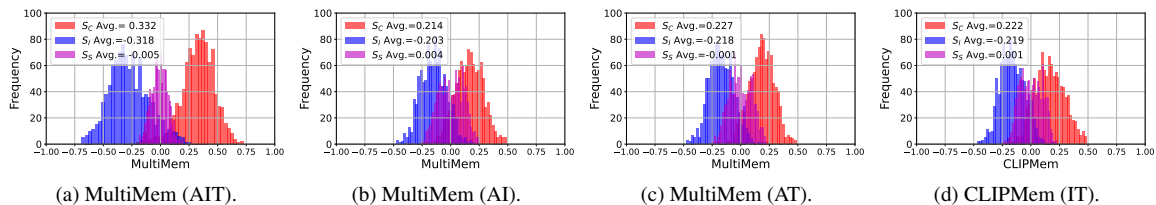


Figure 1. **Memorization should be measured on all modalities instead of only on modality pairs.** (a) Our MultiMem scores across all three modalities (AIT: Audio, Image, and Text) for AudioCLIP. We quantify pairwise memorization on all modality pairs: (b) Audio-Image and (c) Audio-Text (with MultiMem), and (d) Image-Text (with CLIPMem).

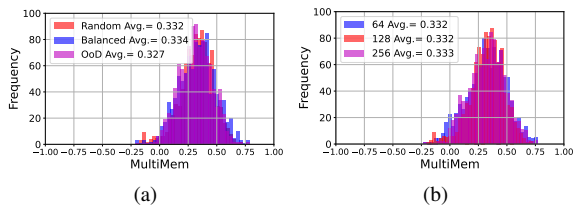


Figure 2. **Our MultiMem is robust to hyperparameters  $H$ .** (a) We varied the composition of  $H$  from 128 randomly chosen samples (used in metric) to 128 samples with class balance and 128 OoD samples from the AudioSet dataset [7]. (b) We varied the number of samples in  $H$ .

We observe the same trends for AVT-CLIP and AVIT-CLIP, as we show in Figure 7, Figure 8, and Figure 10 in Section A.3. These findings suggest that in multi-modal models with more than two modalities, a memorization metric that jointly considers all modalities is necessary for a more accurate assessment of memorization.

**Robustness to Held-out Set.** To examine the sensitivity of our MultiMem metric to the choice of held-out set  $H$ , we evaluate the metric using several selection strategies for  $H$ : (1) randomly sampled data points, (2) data points chosen to achieve balanced class representation, and (3) samples drawn from an entirely different dataset within the same domain. Furthermore, we investigate the robustness of the metric to the size of  $H$  by varying the number of data points included in the set. The results in Figure 2 show that our MultiMem is robust to the choice of  $H$  in both size and composition.

Size of $ S_C $	500	1000	1500	2000
Avg. MultiMem	0.405	0.408	0.409	0.408

Table 2. Our MultiMem is robust to dataset splitting ratio.

**Robustness to Dataset Splitting Ratio** To validate whether our MultiMem is robust to dataset splitting ratio (i.e.  $S_S/S_C$  ratio), we added additional experiments on AVT-CLIP with different dataset splitting ratios on the MSR-VTT dataset. We use  $S_C = S_I \in \{500, 1500, 2000\}$  in addition to the baseline setting of  $S_C = S_I = 1000$  reported in the main paper, while keeping  $S_S$  fixed at 6000 throughout. Our results in Table 2 highlight that reported memorization remains the same over all setups.

**Highly Memorized Samples.** We examine the highly memorized samples in VideoCLIP and AVT-CLIP to gain deeper insights into the underlying causes of strong memorization. For VideoCLIP, we find that, similar to the results reported in the work of [28], videos with misaligned captions and visual content (i.e., mis-captioned videos) tend to experience a higher level of memorization. However, in AVT-CLIP, we find that highly memorized samples are typically those with semantic misalignment across *multiple* modalities. These findings highlight the importance of evaluating cross-modal consistency rather than focusing only on text quality. Examples of the most memorized samples for VideoCLIP and AVT-CLIP are shown in Figure 11 in Section A.3 and a complete list of the top 10 most memorized samples (including their audio and video content and full captions) is

provided in the supplementary material.

**Memorization and Generalization.** To investigate the relationship between generalization and memorization in multi-modal models, we introduce Gaussian noise with different strengths into the representations of all modalities during training as augmentations to control the model’s memorization level, as proven successful by previous works [21, 28, 35]. The models’ generalization is evaluated by the  $T@5$  retrieval task introduced in Section 4.1. More specifically, we implement the experiments on VideoCLIP (bi-modal), AudioCLIP (tri-modal) and AVIT-CLIP (quad-modal) by injecting Gaussian noise with mean  $\mu = 0$  and standard deviations  $\sigma = [0.025, 0.050, 0.075, 0.100, 0.125, 0.150, 0.175, 0.200]$ .

In Figure 3a, when the noise strength is lower than 0.1, we observe that a higher Video-Text MultiMem results in a worse downstream performance, and vice versa. This aligns with the results reported by [28] for bi-modal memorization in CLIP models. Moreover, our results in Figure 3b and Figure 3c show that AudioCLIP’s generalization (trained on three modalities) is negatively correlated with the *tri-modal* MultiMem. In contrast, there is no consistent correlation between generalization and memorization when applying CLIPMem or the bi-modal version of MultiMem, highlighting again the need to take all modalities into account when assessing multi-modal memorization. We also observe similar trends in AVIT-CLIP for Figure 3d. Together, these results suggest that partial-modal memorization is insufficient to explain how memorization contributes to generalization in multi-modal models and highlight again the benefits of a truly multi-modal metric.

Finally, when the noise strength exceeds 0.1, the performance of both AudioCLIP and VideoCLIP begins to decline, while global memorization increases. This occurs because excessive noise destroy the semantic consistency between modalities. As a result, learning across modalities becomes more difficult, leading to increased global memorization and weakening the model’s generalization.

**Impact of Augmentations.** We further study how applying augmentations (injecting noise) to different modality configurations (*e.g.*, single modality, a subset of all modalities, and all modalities) during training influences the generalization of multi-modal models. We adopt SSLMem from [27] to measure memorization for single modality and use MultiMem for multi-modal memorization. The results in Table 3 on AudioCLIP suggest that applying augmentations to a single modality slightly increases the SSLMem of all modalities and enhances the model’s global generalization. However, it also leads to increased memorization in modality pairs that do not involve the augmented modality, which in turn degrades the performance of retrieval tasks based on that

modality pair (as indicated by the underlined values in the tables). Moreover, applying augmentations to a subset of the modalities reduces both global and pairwise memorization, resulting in improved performance over all retrieval tasks. Notably, applying augmentations to all modalities always yields the best performance and the lowest memorization (as reflected by the bolded values in the tables), which provides a foundation for the MultiMem-based memorization mitigation strategies that we introduce in Section 5.

**Memorization with ImageBind.** We train an additional AVIT model using the ImageBind-loss [8]. Unlike our setup, which performs pairwise alignment of *all* modalities, the ImageBind-loss takes the image representation from a pre-trained image encoder as a reference and aligns all other modalities exclusively to this representation. During training, the image encoder remains frozen. Our results show that training with the ImageBind loss leads to increased global memorization (0.571 versus 0.488 achieved with our original AVIT) and reduced performance in retrieval task (36.5% versus 48.2% our AVIT model). These results suggest that training with a pairwise alignment across *all* modalities is beneficial as it reduces memorization and improves generalization. In Section A.4, we present further details and an additional results on how the integration of more modalities during training improves generalization.

**Balancing Per-Modality Memorization.** Given the differences in how each modality contributes to the overall memorization (see for example Table 3 for AudioCLIP), we further show that enforcing a balanced memorization distribution among modalities during training leads to a lower overall memorization and improved model performance. We apply  $\ell_1$ -normalized weights of [0.332, 0.324, 0.344], which are negatively correlated to their memorization contribution [0.222, 0.227, 0.214], to the three modality pairs (I-T, A-T, A-I) for AudioCLIP training. The new trained model achieves a lower MultiMem of 0.290 (compare to 0.332 of baseline model), a 3.2% improvement in retrieval task performance and a more balanced bi-modal MultiMem level of [0.192, 0.195, 0.193] for three modality pairs (I-T, A-T, A-I). We present the full setup in Section A.4.

**Memorization Behavior in Multi-modal, SSL and SL.** We apply UnitMem [26] to analyze where inside the models memorization happens. We compare AudioCLIP (tri-modal), and compare it with CLIP (bi-modal), SL, and SSL models (uni-modal). All these models are trained with UrbanSound8K dataset + spectrogram images. A higher average UnitMem at a given layer indicates a greater contribution of that layer to the model’s global memorization.

We observe that, compared to CLIP, the vision encoder in AudioCLIP aligns more with SSL, rather than falling

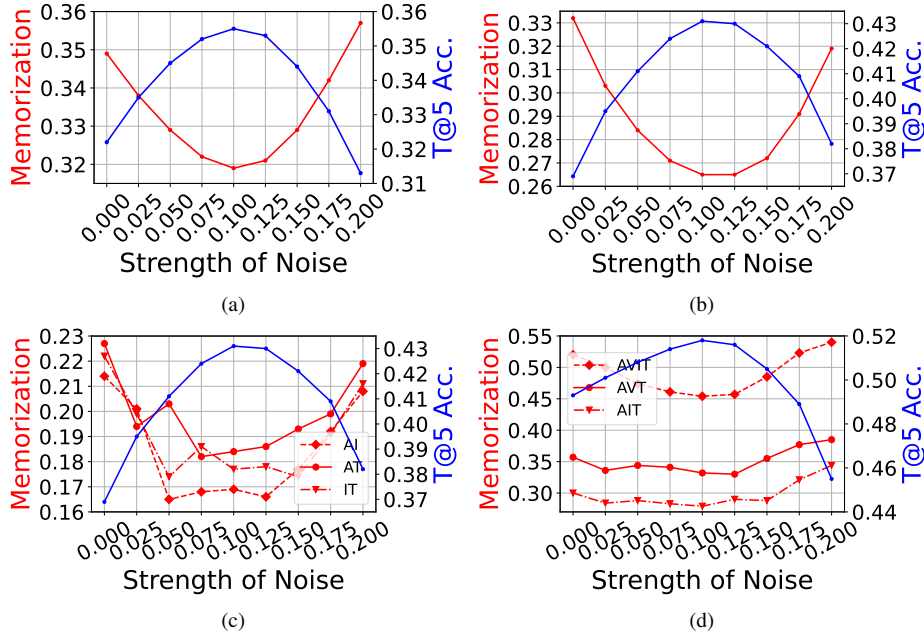


Figure 3. **Generalization of multi-modal models is negatively correlated with the models’ global memorization.** (a) MultiMem (VT) vs. generalization in VideoCLIP. (b) tri-modal MultiMem (AIT) vs. generalization in AudioCLIP. (c) bi-modal MultiMem (AI, AT, IT) vs. generalization in AudioCLIP. (d) Quad-modal MultiMem (AVIT) and tri-modal MultiMem (AVT, AIT) vs. generalization in AVIT-CLIP.

Augmented modality	IT MultiMem	AT MultiMem	AI MultiMem	AIT MultiMem	A SSLMem	I SSLMem	T SSLMem	T@5(%) I-T	T@5(%) A-T	T@5(%) A-I	T@5(%) AI-T
None	0.222	0.227	0.214	0.332	0.188	0.191	0.210	33.1	30.8	26.5	36.9
Audio	<u>0.235</u>	0.201	0.204	0.320	0.199	0.196	0.214	<u>32.2</u>	31.3	27.1	38.2
Image	0.201	<u>0.235</u>	0.188	0.312	0.194	0.210	0.217	<u>33.8</u>	<u>30.0</u>	28.3	39.0
Text	0.181	0.191	<u>0.216</u>	0.294	0.196	0.201	0.235	34.4	32.1	<u>26.3</u>	39.8
Audio + Image	0.200	0.199	0.199	0.299	0.207	0.219	0.233	33.9	31.8	27.5	39.4
Audio + Text	0.187	0.187	0.192	0.283	0.228	0.211	0.244	34.9	32.6	28.0	40.7
Image + Text	0.180	0.189	0.174	0.271	0.211	0.240	0.251	35.4	32.3	29.1	42.1
Audio + Image + Text	<b>0.177</b>	<b>0.184</b>	<b>0.169</b>	<b>0.264</b>	0.236	0.262	0.269	<b>35.7</b>	<b>33.0</b>	<b>29.5</b>	<b>43.1</b>

Table 3. **Adding augmentations to different modalities in AudioCLIP.** We add random Gaussian noise with  $\mu = 0$ ,  $\sigma = 0.1$  to the representations as augmentations. A: Audio, I: Image, and T: Text.

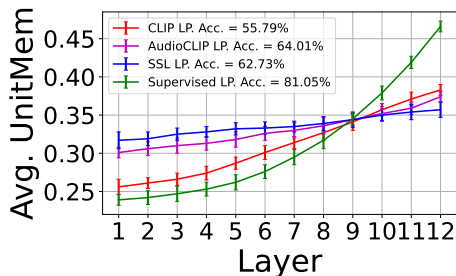


Figure 4. **UnitMem: AudioCLIP is more aligned with SSL.**

between SSL and SL as CLIP. This suggests that the introduction of the audio modality partially reduces the dominance of the text modality (*e.g.*, with labels or captions)

during training. Thereby, the memorization behaviors of the vision encoder shifts from being label-driven (in SL) or caption-driven (in CLIP), to pattern-driven as in SSL [27]. We hypothesize that this is because the learning signals in multi-modal models no longer come from a single supervision source (*e.g.*, labels or captions). Instead, the model focuses more on the shared or similar semantic patterns *across modalities* such as audio, vision, and text. Therefore, the pattern underlying this shift is the *strong semantic interdependence* across modalities in multi-modal models.

## 5. Our Mitigation Strategies

Our findings from the previous section highlight that mitigating cross-modal memorization over all modalities can improve the generalization of multi-modal models. Based on

this motivation, we design two methods built on MultiMem to mitigate high memorization, either during or post training.

**In-Training Mitigation.** Instead of applying noise to the entire dataset, which harms overall performance and incurs additional computational overhead, we selectively add Gaussian noise ( $\mu = 0$ ,  $\sigma = 0.1$ ) as augmentation only to the representations of the most memorized samples during training to mitigate memorization. Specifically, at every 10-epoch interval (*i.e.*, epochs 10, 20, ..., 90), we use MultiMem to measure memorization for all training samples and select the top 5% most memorized samples. The 5% ratio is selected based on our experiments in Figure 13 in Section A.4, which show that this ratio yields the highest model performance and the lowest memorization level. We then aggregate these samples into new mini-batches and apply noise-based augmentations to their representations during the following training steps while keeping training unaltered for all other mini-batches.

In our experiment, we train AudioCLIP on the UrbanSound8K + Spectrogram dataset for 100 epochs with our mitigation in place and report performance on the retrieval task in Figure 5a, the classification task on UrbanSound8K in Figure 5b, and the zero-shot classification task on AudioSet [7] in Figure 5c.

The results yield two main findings. First, applying our mitigation strategy at any training stage effectively reduces global memorization and improves overall performance across all three tasks. Second, model performance in all three tasks follows a trend of initial improvement, followed by a plateau, then further improvement, with a slight decline observed at epoch 90. We attribute this pattern to the timing of memorization mitigation. Applying mitigation too early in training may fail to accurately capture the most memorized samples, as the model is far from convergence. As a result, highly memorized examples that are not identified in the early stages may continue to increase the model’s memorization in later epochs, thereby harming generalization. Applying memorization mitigation too close to the end of training may lead to insufficient decoupling of modality correlation for highly memorized samples. This is the reason why the performance drop is observed at epoch 90 compared to epoch 80. These results indicate that applying noise-based augmentation near the end of training (*e.g.*, at 80%) to a selected ratio of most memorized samples according to MultiMem can effectively reduce memorization and enhance generalization. In Section A.4, we provide further insights into this strategy. (1) We compare the mitigation effects between using noise-based augmentation and gradient clipping (which is widely used in Differential Privacy area) for most memorized samples. (2) We compare it with an alternative approach that directly removes a fixed proportion of the most memorized samples during training. The results show that

our strategy yields better performance improvements. (1) We also examine the effect of repeatedly applying this strategy at multiple stages of training. The findings indicate that repeated application leads to additional gains compared to single use. However, the improvement becomes marginal when applied more than twice.

**Post-Training Mitigation.** Post-training, we first train the model and then use MultiMem to identify the most memorized samples. Then, we remove these samples from training and fine-tune the model for some additional steps on the remaining data. In our experiments, we train an AudioCLIP model on the UrbanSound8K dataset for 100 epochs, then, we use *MultiMem* to identify the most memorized samples and remove the top [50, 100, 150, 200, 300, 500, 700] of them. Finally, we fine-tune the model with the remaining dataset for an additional 25 epochs. For comparison, we additionally employ three alternative strategies for selecting samples to remove as baselines: using *CLIPMem*, computing the sum of *cosine similarities* across modality pairs, and *randomly* selecting samples.

We show the results in Figure 5 (lower row) for retrieval, classification on UrbanSound8K, and zero-shot tasks accuracy on AudioSet. We observe that regardless of the method used to mitigate the model’s memorization (apart from *random*), the model’s performance improves to varying degrees when fewer than 500 samples are removed on all three tasks. Among them, *MultiMem* yields the best performance improvement, followed by the *cosine similarity* and *CLIPMem*. In contrast, the baseline of removing *random* samples does not show a clear trend of performance improvement. This highlights the effectiveness of MultiMem in post-training memorization mitigation and improving generalization.

**Regularization vs Memorization Mitigation.** Prior work has shown that Gaussian noise may act as a regularizer that independently improves performance rather than genuinely mitigating memorization [5, 6, 35]. To verify that the observed performance improvements and memorization mitigation are actually from our proposed mitigation strategy rather than the generic regularization effect of noise, we conducted new experiments where we add *the same amount of noise* to different types of samples. 1) most memorized samples and least memorized samples identified by MultiMem, 2) samples with highest gradients or loss, 3) randomly selected samples. We report the performance gains on retrieval, linear probing, and zero-shot tasks, and MultiMem. If the improvements were only due to generic regularization rather than memorization mitigation, we would observe similar results over all setups. However, the results in the Table 4 show discrepancies: 1) adding noise to random samples does not change performance significantly. 2) Adding noise to least memorized samples slightly degrades performance. 3)

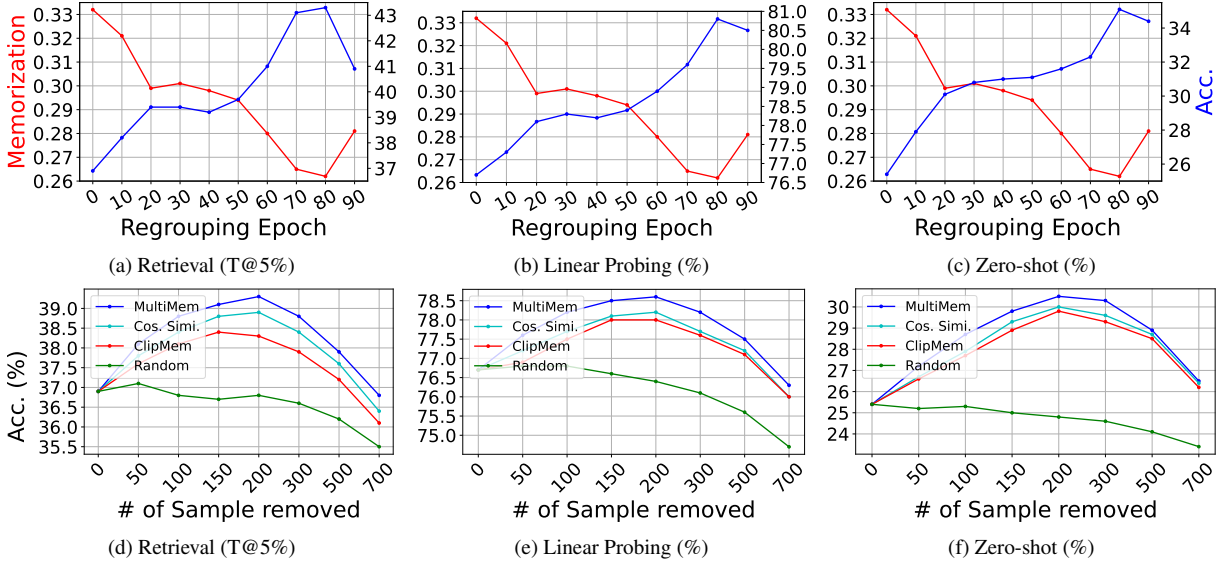


Figure 5. **Mitigation Strategies.** The first row shows the impact of in-training memorization mitigation at different training epochs, evaluated on (a) retrieval, (b) linear probing, and (c) zero-shot classification tasks. The second row presents the effect of post-training mitigation with increasing number of most memorized samples removed, analogously for (d) retrieval, (e) linear probing, and (f) zero-shot classification tasks.

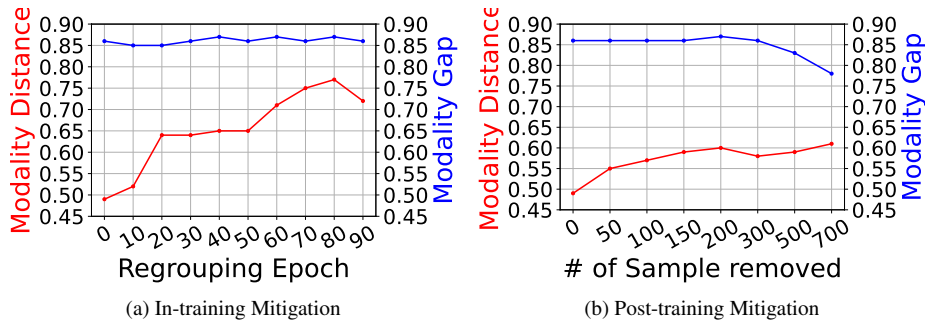


Figure 6. **Average modality representation distance for 1% most memorized samples versus average modality gap for all training samples before and after mitigation**

Our **MultiMem** approach has much higher performance gains compared to loss-based and gradient-based methods and also achieves the best memorization mitigation. This shows that our reported improvements indeed stem from memorization removal and not general regularization.

**Insights into Mitigation.** We further compared the average modality distance between the representations of each modality for the most memorized samples, as well as the average modality gap [15] across all samples in the training set, before and after mitigation. Results in Figure 6 show that: (1) The model struggles to align semantic misaligned samples through generalization, thereby memorizes those datapoints so that results in an extremely low average sample modality distance. This shows that semantic misalignment is the key driver of memorization in multi-modal models.

(2) Both of our mitigation methods increase the average modality distance of the most memorized samples thereby reducing the memorization level of the model. However, in-training mitigation performs better than post-training mitigation, and therefore yields a greater improvement in the generalization of the model. At the same time, in-training mitigation preserves the original average modality gap better. This is consistent with the conclusion in [15], that preserving the original modality gap tends to yield better model performance. These results show that our mitigation methods address the fundamental cause of memorization in multi-modal models, *i.e.*, controlling the average modality distance for semantic misaligned samples.

**Impact of Noise Injection vs. Augmentation.** Prior work [21, 28, 35] has shown that Gaussian noise can indeed

Mitigation	Post-training				In-training			
	Retrieval↑	Linear Prob.↑	Zero-Shot↑	Mem.↓	Retrieval↑	Linear Prob.↑	Zero-Shot↑	Mem.↓
None	36.9%	76.7%	25.4%	0.332	36.9%	76.7%	25.4%	0.332
Random	37.1%	76.8%	25.5%	0.330	37.2%	76.9%	25.5%	0.329
MultiMem (least)	36.6%	76.6%	25.2%	0.336	36.1%	76.1%	25.0%	0.338
Gradient	38.1%	77.8%	29.4%	0.307	40.4%	78.8%	31.1%	0.281
Loss	38.4%	77.9%	29.6%	0.308	39.9%	78.9%	30.9%	0.280
<b>MultiMem (most)</b>	<b>39.3%</b>	<b>78.6%</b>	<b>30.5%</b>	<b>0.282</b>	<b>43.3%</b>	<b>80.8%</b>	<b>35.1%</b>	<b>0.262</b>

Table 4. Regulation verification by comparing MultiMem with other mitigation strategy.

Mitigation	Post-training				In-training			
	Retrieval↑	Linear Prob.↑	Zero-Shot↑	Mem.↓	Retrieval↑	Linear Prob.↑	Zero-Shot↑	Mem.↓
None	36.9%	76.7%	25.4%	0.332	36.9%	76.7%	25.4%	0.332
<b>Noise</b>	<b>39.3%</b>	<b>78.6%</b>	<b>30.5%</b>	<b>0.282</b>	<b>43.3%</b>	<b>80.8%</b>	<b>35.1%</b>	<b>0.262</b>
Augmentation	38.8%	78.4%	30.0%	0.288	42.5%	80.1%	34.0%	0.269

Table 5. Modality-specific augmentation vs. noise injection.

serve as a generic augmentation strategy across modalities. This, however, raises the concern of whether such noise genuinely achieves an effect comparable to that of modality-specific augmentations, and whether it yields a similar level of memorization mitigation. To address this concern, we conduct an additional control experiment on AudioCLIP. Specifically, we replace noise injection with modality-specific augmentations: Gaussian blur ( $\sigma=(0.1, 2.0)$ ) combined with random color jitter ( $\text{transforms.ColorJitter}(0.4, 0.4, 0.4, 0.1)$ ) for the image modality, synonym substitution ( $p=0.2$ ) for captions, and time shifting (+10%) for audio. As shown in Table 5, mitigation guided by these modality-specific augmentations remains effective, achieving performance comparable to that of noise injection. This indicates that the mitigation effect arises from the underlying mechanism itself rather than from the generic regularization properties of noise.

## 6. Conclusions

We propose MultiMem, a novel metric for measuring and characterizing memorization in multi-modal contrastive learning with arbitrary modality configurations. Our results show clear differences compared to commonly studied bi-modal models like CLIP. We find that training with multiple modalities not only mitigates the single-modality dominance in global memorization observed in bi-modal models, but also makes the contrastive models’ memorization behavior more similar to that of self-supervised learning. Specifically, text modality, which has been viewed in previous work [28] as similar to labels, is no longer the only reason for high memorization when it does not align with other

modalities. Inconsistency in semantic information between multiple modality pairs is instead the leading cause of high memorization. Finally, we show that our proposed MultiMem metric can be used to inform mitigations against memorization, either during training or after training, that improve generalization more efficiently than other metrics.

## Acknowledgments

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project number 550224287. Franziska Boenisch received funding from the European Research Council (ERC) under the European Union’s Horizon Europe research and innovation programme (grant agreement No 101220235). We would like to acknowledge our sponsors, who support our research with financial and in-kind contributions: OpenAI and G-Research. We also thank members of the SprintML group for their feedback.

## References

- [1] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4959–4968, 2022. 1
- [2] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21466–21474, 2022. 1
- [3] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings*

- of the National Academy of Sciences, 117(48):30063–30070, 2020. 1
- [4] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023. 4
- [5] Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959, 2020. 1, 2, 3, 8
- [6] Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020. 1, 3, 8
- [7] 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 *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017. 5, 8
- [8] 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*, pages 15180–15190, 2023. 6
- [9] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Esresnet: Environmental sound classification based on visual domain models. In *2020 25th international conference on pattern recognition (ICPR)*, pages 4933–4940. IEEE, 2021. 2
- [10] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022. 2, 4
- [11] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17980–17989, 2022. 1
- [12] Bargav Jayaraman, Chuan Guo, and Kamalika Chaudhuri. Déjà vu memorization in vision-language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 50722–50749, 2024. 1, 3
- [13] Songtao Li and Hao Tang. Multimodal alignment and fusion: A survey. *arXiv preprint arXiv:2411.17040*, 2024. 1
- [14] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10):1–42, 2024.
- [15] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. 1, 9
- [16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 5
- [17] Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. Cptr: Full transformer network for image captioning. *arXiv preprint arXiv:2101.10804*, 2021. 1
- [18] Casey Meehan, Florian Bordes, Pascal Vincent, Kamalika Chaudhuri, and Chuan Guo. Do ssl models have déjà vu? a case of unintended memorization in self-supervised learning. *Advances in Neural Information Processing Systems*, 36:42775–42798, 2023. 3
- [19] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*, pages 529–544. Springer, 2022. 1
- [20] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2
- [21] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8238–8247, 2022. 6, 9
- [22] 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*, pages 8748–8763. PmLR, 2021. 1, 2, 4
- [23] Ildus Sadrtdinov, Nadezhda Chirkova, and Ekaterina Lobacheva. On the memorization properties of contrastive learning. *arXiv preprint arXiv:2107.10143*, 2021. 2
- [24] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044, 2014. 5
- [25] Haoyu Song, Li Dong, Weinan Zhang, Ting Liu, and Furu Wei. Clip models are few-shot learners: Empirical studies on vqa and visual entailment. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6088–6100, 2022. 1
- [26] Wenhao Wang, Adam Dziedzic, Michael Backes, and Franziska Boenisch. Localizing memorization in ssl vision encoders. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 60475–60516, 2024. 1, 3, 6
- [27] Wenhao Wang, Muhammad Ahmad Kaleem, Adam Dziedzic, Michael Backes, Nicolas Papernot, and Franziska Boenisch. Memorization in self-supervised learning improves downstream generalization. In *ICLR*, 2024. 1, 2, 3, 4, 6, 7
- [28] Wenhao Wang, Adam Dziedzic, Grace C Kim, Michael Backes, and Franziska Boenisch. Captured by captions: On memorization and its mitigation in clip models. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2, 3, 5, 6, 9, 10

- [29] Yanan Wang, Michihiro Yasunaga, Hongyu Ren, Shinya Wada, and Jure Leskovec. Vqa-gnn: Reasoning with multi-modal knowledge via graph neural networks for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21582–21592, 2023. [1](#)
- [30] Keyu Wen, Jin Xia, Yuanyuan Huang, Linyang Li, Jiayan Xu, and Jie Shao. Cookie: Contrastive cross-modal knowledge sharing pre-training for vision-language representation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2208–2217, 2021. [1](#)
- [31] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, 2021. [2](#), [4](#)
- [32] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. [5](#)
- [33] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *Advances in Neural Information Processing Systems*, 35:124–141, 2022. [1](#)
- [34] Jiayuan Ye, Anastasia Borovykh, Soufiane Hayou, and Reza Shokri. Leave-one-out distinguishability in machine learning. In *The Twelfth International Conference on Learning Representations*, 2023. [1](#)
- [35] Ziqi Yuan, Baozheng Zhang, Hua Xu, Zhiyun Liang, and Kai Gao. Openvna: A framework for analyzing the behavior of multimodal language understanding system under noisy scenarios. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 9–18, 2024. [6](#), [8](#), [9](#)
- [36] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pages 493–510. Springer, 2022. [1](#)
- [37] Y Zhou and G Long. Style-aware contrastive learning for multi-style image captioning. In *EACL 2023-17th Conference of the European Chapter of the Association for Computational Linguistics, Findings of EACL 2023*, 2023. [1](#)

## A. Appendix

### A.1. Hardware Usage & Calculation Overhead

Two devices are used for all experiments mentioned in this paper: a cloud server with four A100 GPUs and a local workstation with Intel 13700k CPU, Nvidia 4090 graphics card, a total of 10Tb storage space and 64GB of RAM.

Applying In-training mitigation strategy once during training will bring 0.70% (2.6 min out of 372 min) overhead in total time usage, without requiring any additional memory. Applying Post-training mitigation strategy will bring 0.55% (2.1 min out of 372 min) overhead in total time usage for memorization measurement, without requiring any additional memory.

Measuring MultiMem for all samples in candidate set  $S_C$  averagely brings 0.48% overhead in total time usage, without requiring any additional memory. Since the computational complexities of our MultiMem on both the number of modalities and the number of training samples are  $O(n^2)$ , which matching the complexity of multi-modal contrastive learning. Therefore, the relative overhead remains essentially unchanged as modalities or data scale up.

### A.2. Extended Experiment Setup

**Glossary.** For the reader’s convenience, we provide a glossary with all important notation used in the main paper in Table 6.

**Dataset Split.** Table 7 provides a detailed summary of the splitting  $S_C$ ,  $S_I$ , and  $S_S$  discussed in main paper.

**Training Setup.** Table 8 provide a detailed summary of the training configurations for all models used in our experiments. As stated in the main paper, all settings for existing models follow the default configurations of their respective original implementations.

**AVT-CLIP and AVIT-CLIP.** For AVT-CLIP, the video and text modalities are first pre-trained for 20 epochs, after which the audio modality is introduced. The model is then trained jointly on all three modalities for an additional 100 epochs. For AVIT-CLIP, the image and text modalities are first pre-trained for 20 epochs, after which the video and audio modalities are introduced and trained jointly on all four modalities for an additional 200 epochs.

**Experimental Setup for UnitMem.** In this experiment, we use the original UrbanSound8K dataset as the source of both audio inputs and text labels. Following the evaluation setup of AudioCLIP, we also use spectrograms of the audio signals as image inputs. In the SL setting, we employ a ViT encoder followed by a two-layer MLP classifier, trained with

spectrogram images as inputs and text labels as supervision. For SSL settings, we employ the ViT-based DINO framework and train the ViT encoder on only spectrogram images. Both the image augmentation sets and hyperparameters follow the default configurations of their respective original implementations.

### A.3. Extended Experiment Results

#### Measuring Memorization in AVT-CLIP and AVIT-CLIP.

For AVT-CLIP, we measure the memorization with two metrics: 1) tri-modal MultiMem with all three modalities and 2) bi-modal MultiMem of all modality pairs (A-V, A-T, and V-T). The results are shown in Figure 7, which aligns with the results for AudioCLIP in the main paper. For AVIT-CLIP, we also measure the memorization with two metrics: 1) quad-modal MultiMem with all four modalities and 2) tri-modal MultiMem for VideoCLIP and AVT-CLIP (*i.e.*, AVT MultiMem and AIT MultiMem). The results are shown in Figure 8. We can find that when AVT-CLIP or AudioCLIP is extended to a quad-modal setting, the previously used tri-modal MultiMem becomes insufficient in capturing both the distribution of memorized samples and the accurate measurement of memorization levels. This again highlights the necessity of using all modalities to measure memorization.

#### Measuring Memorization AVIT-CLIP with synthesized images.

We use Stable Diffusion v1.5 to generate images from the captions of the MSRVT dataset to construct a dataset with true four modalities: (1) audio, (2) video, (3) image, and (4) text. We adopt this setup based on the assumption that using newly generated images with styles different from the video frames can mitigate the potential unnecessary memorization caused by overly similar visual styles between images and video frames. Note that we synthesize the images for the MSRVT dataset because there are no natural 4-modality datasets. A few examples of generated images are presented in Figure 9.

The results are shown in Figure 10, we find the memorization distribution fully follow the same trend observed in Figure 8.

#### Top 10 most memorized samples for VideoCLIP and AVT-CLIP.

Unlike images, videos contain dynamic visual and audio signals, which cannot be effectively displayed in the paper. To address this, We provide two examples here and others are shown in supplementary materials. As shown in Figure 11a, video3405 is a completely dark-screen video, providing almost no visual semantic information relevant to its caption: “a man talking about hydroponic fluid pressure.” As a result, it ranks first in VideoCLIP with a high MultiMem of 0.529. However, its audio clearly mentions “hydroponic fluid pressure”, which aligns well with the caption. Therefore, its MultiMem score in AVT-CLIP is only

Symbol	Explanation
A	Audio modality
V	Video modality
I	Image modality
T	Text modality
AVT-CLIP	Variant of CLIP with Audio + Video + Text modalities
AVIT-CLIP	Variant of CLIP with Audio + Video + Image + Text modalities
H	Held-out test set used for MultiMem
Aug	Augmentation set for data points

Table 6. We provide an overview on the important notation in main paper.

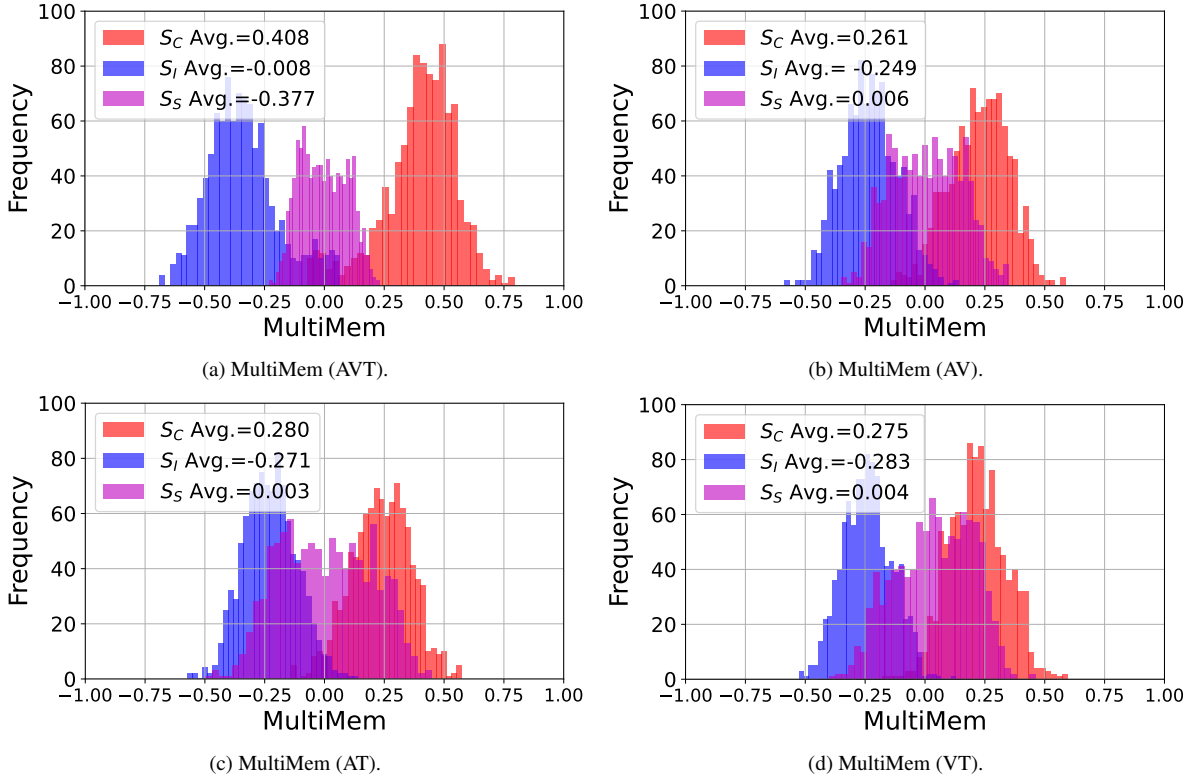


Figure 7. **Measuring memorization only on modality pair is insufficient for AVT-CLIP trained on MSRVT.** (a) Measure memorization across all three modalities (Audio, Video, and Text) of AudioCLIP. (b)-(d) Measure pairwise memorization on all modality pairs (Audio-Video, Audio-Text, and Video-Text) of AVT-CLIP.

Dataset	Total	$S_S$	$S_I$	$S_C$
COCO	123287	65000	5000	5000
MSR-VTT	10000	7000	1000	1000
UrbanSound8K	8732	6000	1000	1000

Table 7. Detailed dataset split used in this paper.

0.561, ranking 36th. Another example is video6659 shown in Figure 11b, where the visual content features two ani-

mated characters, while the audio contains a conversation between two male-like voices. The title, “a character hunting for love”, is poorly related to both the video and audio. All three modalities are semantically misaligned. As a result, this sample receives a high MultiMem score of 0.712 in AVT-CLIP, ranking second among all samples.

#### A.4. Additional Experiments

**ImageBind Experiment Setup** We follow the original design of ImageBind that leverage the image embedding as the unified anchor space. All other modalities are aligned to this

	SL-ViT	DINO	CLIP	VideoCLIP	AudioCLIP	AVT-CLIP	AVIT-CLIP
Training Epochs	100	300	100	100	100	100	200
Warm-up Epochs	5	20	10	10	10	10	20
Batch Size	1024	1024	512	256	64	64	64
Optimizer	AdamW	AdamW	AdamW	Adam	Adam	Adam	Adam
Learning rate	1e-3	2e-3	1e-3	1e-4	1e-4	1e-4	1e-4
Learning rate Schedule	Cos. Decay	Cos. Decay	Cos. Decay	Cos. Decay	Cos. Decay	Cos. Decay	Cos. Decay

Table 8. Training Setup and Hyperparameter for all models used in this work.

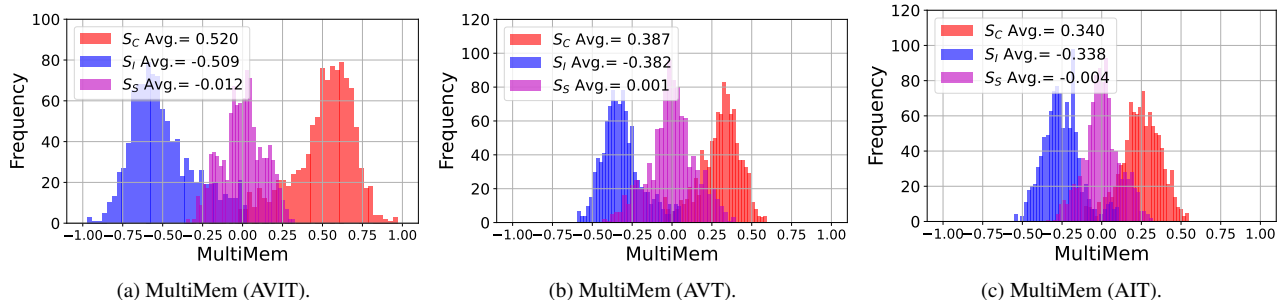


Figure 8. **Measuring memorization only on partial modality is insufficient for AVIT-CLIP.** (a) Measure memorization across all four modalities (Audio, Video, Image, and Text) of AudioCLIP. (b) Measure memorization with AVT MultiMem. (c) Measure memorization with AIT MultiMem.

space through image-to-modality contrastive supervision. We keep the audio, video, and text encoders unchanged as random initialization, but replace the image encoder with our own pretrained image encoder. The encoder produces a normalized image representation that serves as the reference for all contrastive objectives. This design removes all non-image pairwise losses (e.g., audio–video, audio–text, and video–text), ensuring that the image modality forms the central representation space. We use Adam optimizer with a learning rate of 1e-4 and train for 200 epochs. The batch size and dataset split follow the implementation of AVIT-CLIP.

**Enhancing Generalization by Introducing new Modalities.** Our results in Table 9 show that, when extending VideoCLIP to AVT-CLIP, the accuracy of V-T retrieval task increases. This is because adding a new modality to VideoCLIP reduces the text modality’s dominance in memorization, yielding a more balanced memorization distribution across modalities. So that increasing model’s generalization. This is consistent with our claims in Section 4.2. Similar trend could be observed in extending AVT-CLIP or AudioCLIP to AVIT-CLIP.

Note that the AVIT-CLIP always has the best performance in all retrieval tasks (as reflected by the boldface values in the Table 9).

**Effects of Augmenting Different Ratios of Most Memorized Samples.** In Figure 13, we test the effect of applying noise-based augmentation to different proportions of the most memorized samples during training. The results show that when noise-based augmentation is applied to less than 5% of the most memorized samples, global memorization decreases rapidly, and overall performance improves significantly. However, once the proportion exceeds 5%, both memorization and performance begin to stabilize as more samples are augmented. This indicates that decoupling the cross-modal associations of non-highly memorized samples is not effective in mitigating global memorization or improving performance. In contrast, applying augmentation only to the most memorized samples identified by MultiMem achieves the best results with significantly lower computational cost.

**UnitMem results on COCO.** In this experiment, we use the original COCO dataset as the source of both image inputs and text captions. Speeches generated from COCO captions by open-source Tortoise TTS are used as audio modality. For SL, the model is trained using a multi-label classification setting with the cross-entropy loss with COCO dataset. All other settings for SSL, CLIP and AudioCLIP are similar to those in the experiments with UrbanSound8K dataset. We adopt this setup based on the assumption that captions and the audio generated from them share higher cross-modal semantic alignment. This helps mitigate the potential impact



Figure 9. Samples of images generated by Stable Diffusion v1.5.

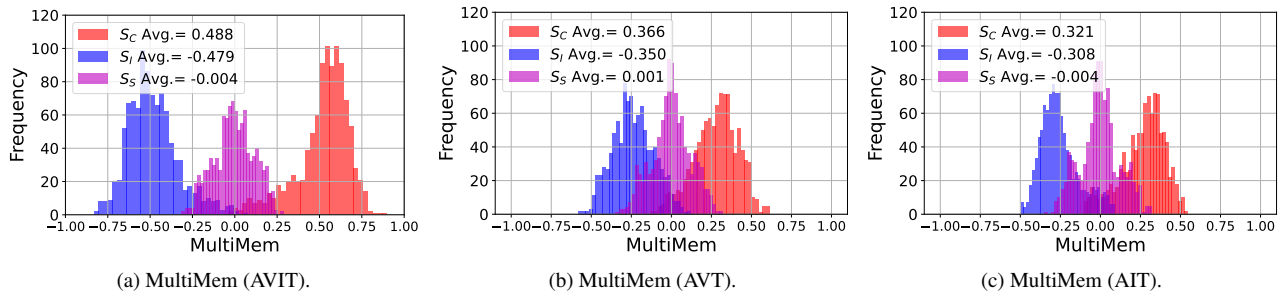


Figure 10. AVIT-CLIP trained with synthesized images follows the same trend with results trained with video-frame images. (a) Measure memorization across all four modalities (Audio, Video, Image, and Text) of AudioCLIP. (b) Measure memorization with AVT MultiMem. (c) Measure memorization with AIT MultiMem.

of weak semantic alignment among spectrograms, audio, and textual labels in the UrbanSound8K dataset. Note that the synthesized audio for COCO only used as extended evidence to verify the trend observed on the UrbanSound8K dataset in main paper.

The results are shown in Figure 12. Despite semantic alignment across modalities being strengthened, we observe that the behavior of AudioCLIP’s vision encoder remains more similar to SSL than to SL. This fully aligns with the result of that implemented on UrbanSound8K dataset in main paper.

**Mitigating Memorization During Training.** In addition to the experiments presented in Section “In-training Mitigation”, we implement three additional experiments to further validate the effectiveness of this approach.

In the first experiment, we compared the different effects of noise-based augmentation and gradient clipping during the in-training mitigation process. The results in Figure 14 shows that gradient clipping needs a much larger learning rate than noise-injection to mitigate the memorization level of the model. Besides, the best performance of gradient

clipping (41.7% +4.8% to 36.9% baseline) is worse than noise-injection(43.3% +6.7% to 36.9% baseline)

In the second experiment, we still use MultiMem to measure the memorization of all training samples at every 10-epoch interval. However, instead of regrouping the top 5% most memorized samples into new batches with noise-based augmentation, we directly **remove** them from the training set. As shown in Figure 15, the performance of *in-training removal* first increases and then drops. This performance drop may be due to similar reasons observed in *in-training regrouping*, i.e., the learning rate becomes too low in the later stages of training, and the remaining number of epochs is insufficient to fully eliminate the negative impact introduced by previously highly memorized samples. Moreover, we observe that the maximum performance gain achieved by *in-training removal* is smaller than that of *in-training regrouping*, further demonstrating the effectiveness of our proposed strategy.

In the third experiment, we further examine that whether applying the *in-training regrouping* strategy multiple times will provide additional memorization mitigation and performance improvement or not. Specifically, we apply the

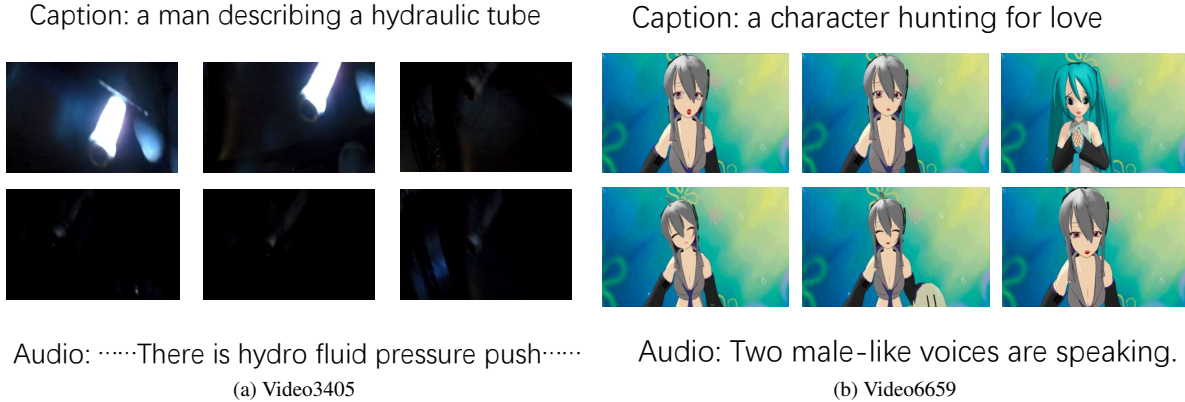


Figure 11. **The most memorized samples for VideoCLIP and AVT-CLIP.**

Model	VT MultiMem	AVT MultiMem	AIT MultiMem	AVIT MultiMem	A SSLMcm	V SSLMcm	I SSLMcm	T SSLMcm	T@5(% V-T)	T@5(% AV-T)	T@5(% AI-T)	T@5(% AVI-T)
VideoCLIP	0.347	-	-	-	-	0.179	-	0.206	32.7	-	-	-
AVT-CLIP	0.275	0.408	-	-	0.238	0.251	-	0.266	33.9	40.4	-	-
AudioCLIP	-	-	0.332	-	0.188	-	0.191	0.210	-	-	36.9	-
Imagebind-AVIT	0.289	0.414	0.339	0.571	0.220	0.231	0.238	0.242	33.0	35.1	32.9	36.5
AVIT-CLIP	0.254	0.366	0.321	0.488	0.244	0.255	0.260	0.267	<b>36.6</b>	<b>41.3</b>	<b>37.0</b>	<b>48.2</b>

Table 9. **Introducing a new modality helps reduce the global memorization, and in turn, improves the model’s generalization.**

*in-training regrouping* strategy twice: once at the end of the warm-up phase (epoch 10) and once near the end of training (epoch 80). The results in Table 10 show that applying the *in-training regrouping* strategy multiple times can further mitigate model memorization and yield better generalization. However, using it three times can hardly achieve a significant performance improvement compared to using it twice, so using it twice at the beginning and end of the training is the best choice considering that it does not bring many extra calculations.

**Model Size and Dataset Size.** We added additional experiments on CLIP models of varying sizes trained on the same data, as well as models of the same size trained on different amount of the training data. The results in Table 11 show that larger models exhibit higher memorization levels when trained on the same data, and also reducing the amount of training data leads to increased memorization for models of the same size. Both trends are consistent with prior work (e.g., [Carlini et al., ICLR 2022], [Tirumala et al., NeurIPS 2022], [27]), as larger model capacity provides more room to retain instance-level details, while sparser data coverage reduces generalization opportunities and increases reliance on memorizing individual samples.

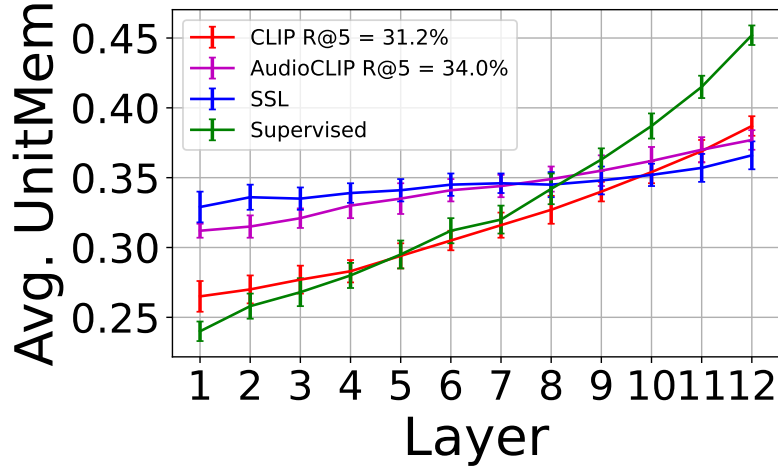


Figure 12. **UnitMem: AudioCLIP is more aligned with SSL.** We implement an extra experiment on COCO with audios generated from the captions.

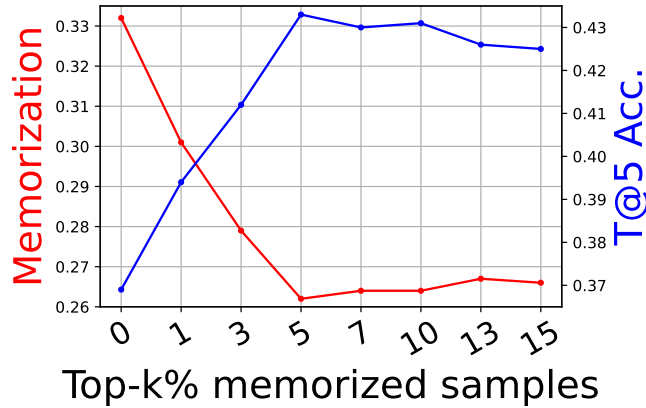


Figure 13. **Effects of noise-based augmentation to the top k% most memorized samples (ranked by MultiMem) during training.**

Regrouping epoch	MultiMem	AI-T retrieval T@5
None	0.332	36.9
10	0.321	38.2
80	0.262	43.3
10 + 80	0.256	43.7
20 + 80	0.252	44.1
10 + 90	0.254	43.8
40 + 50	0.266	42.8
20 + 50 + 80	0.250	44.2
10 + 40 + 70	0.253	43.8

Table 10. **Applying the *in-training regrouping* strategy multiple times further mitigate memorization and improves performance.**

	Baseline	ViT-S + EsResNet-34	ViT-L + EsResNet-101	75% dataset usage	50% dataset usage	25% dataset usage
Avg. MultiMem	0.332	0.308	0.382	0.341	0.360	0.389

Table 11. **Memorization v.s. model size and training datapoint usage.**

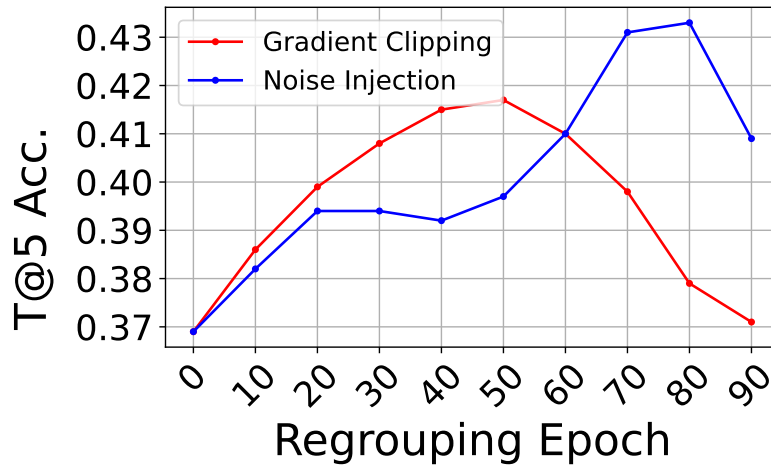


Figure 14. Performance for AudioCLIP under in-training regrouping when using noise-based augmentation and gradient clipping.

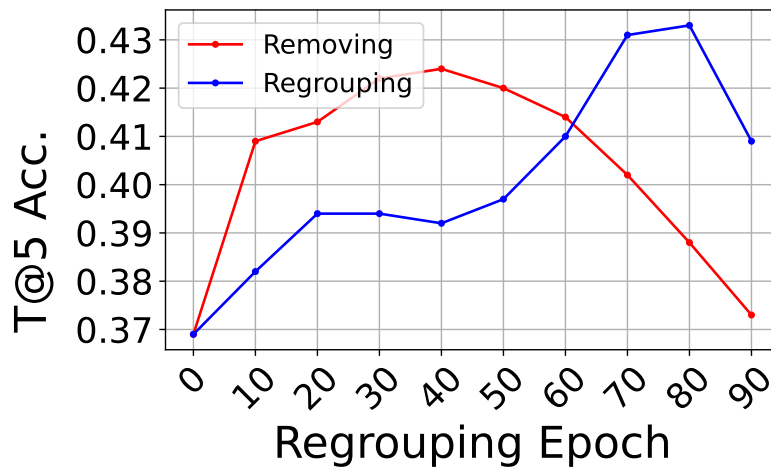


Figure 15. Performance for AudioCLIP under in-training regrouping vs. in-training removal.