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

Wenhao Wang, Franziska Boenisch, Michael Backes, and Adam Dziedzic

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. Full version with appendix available at `http://arxiv.org/abs/2606.22220`

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 [28]. 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 [28] 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* [27]. 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, 28, 34] for measuring memorization and compares the outputs of a 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 **AudioCLIP** [10] (including **A**udio, **I**mage, and **T**ext modalities), **AVT-CLIP** (a custom-built tri-modal model with **A**udio, **V**ideo, and **T**ext modalities, introduced in Section 4.1), and **AVIT-CLIP** (a custom-built quad-modal model with **A**udio, **V**ideo, **I**mage, and **T**ext 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 [27], 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 [27]. This suggests that multi-modal models' memorization behavior is closer aligned with the one of SSL models [28] 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.,* **A**udio-**I**mage, **A**udio-**T**ext, and **I**mage-**T**ext). 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, 28] 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]:

$$\begin{aligned}\mathbf{Mem}\left(\mathcal{A}, S, i\right) = &\underset{f \leftarrow \mathcal{A}(S)}{\mathbf{Pr}}\left[f\left(x_i\right) = y_i\right] \\ &- \underset{g \leftarrow \mathcal{A}(S^{\setminus i})}{\mathbf{Pr}}\left[g\left(x_i\right) = y_i\right],\end{aligned} \tag{1}$$

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, [28] proposed a new metric:

$$\begin{aligned}\mathbf{SSLMem}(S, Aug, i) = &\underset{\substack{g \sim \mathcal{A}(S^{\setminus i}) \\ {x_i}', {x_i}'' \sim \mathrm{Aug}(x_i)}}{\mathbb{E}}\left[d\left(g\left({x_i}'\right), g\left({x_i}''\right)\right)\right] \\ &- \underset{\substack{f \sim \mathcal{A}(S) \\ {x_i}', {x_i}'' \sim \mathrm{Aug}(x_i)}}{\mathbb{E}}\left[d\left(f\left({x_i}'\right), f\left({x_i}''\right)\right)\right],\end{aligned} \tag{2}$$

where $d(\cdot, \cdot)$ represents a distance function, commonly the ℓ_2 distance, and $\mathrm{Aug}(\cdot)$ 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 [27]. 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:

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

where f and g are CLIP models trained on datasets with and without a data point (I,T), consisting of an **I**mage and **T**ext (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.

$$\begin{aligned}\mathcal{A}_{\text{align}}(f,I,T) = &\mathop{\mathbb{E}}_{(I',T')\sim\text{Aug}(I,T)}\left[\text{sim}(f_{\text{img}}(I'), f_{\text{txt}}(T'))\right] \\ &-\mathop{\mathbb{E}}_{t}\left[\text{sim}(f_{\text{img}}(I), f_{\text{txt}}(t))\right] - \mathop{\mathbb{E}}_{i}\left[\text{sim}(f_{\text{img}}(i), f_{\text{txt}}(T))\right],\end{aligned} \tag{4}$$

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 **I**mage and **T**ext 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 Φ as:

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

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

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

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:

$$\textbf{MultiMem}(i, H, f) = \underset{f \sim S}{CMC}(i, H) - \underset{g \sim S^{\setminus i}}{CMC}(i, H). \tag{7}$$

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

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 (**A**udio + **V**ideo + **T**ext) and AVIT-CLIP (**A**udio + **V**ideo + **I**mage +

Text). The detailed encoder architectures and training datasets used are shown in Table 1 while other training details and hyperparameters are presented in Table 6 in Appendix A.2.

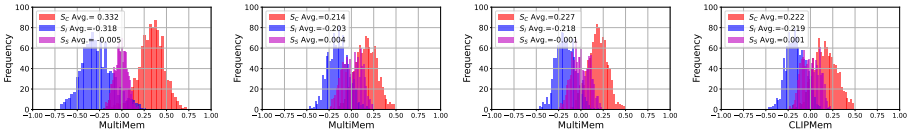
Model	Modality	Encoder	Training Set
CLIP	Image + Text	ViT + Transformer	COCO [16]
VideoCLIP	Video + Text	(CNN + Transformer) + Transformer	MSR-VTT [32]
AudioCLIP	Audio + Image + Text	ESResNeXt + ViT + Transformer	UrbanSound8K [24]+ Spectrogram
AVT-CLIP	Audio + Video + Text	ESResNeXt + (CNN + Transformer)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

Dataset Splitting. Following [28], 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 5 in Appendix A.2. We report the average MultiMem score on S_C as the memorization for model f. We further prove our MultiMem is robust to dataset splitting in Appendix A.4.

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 $\mathbf{N} = 5$ 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



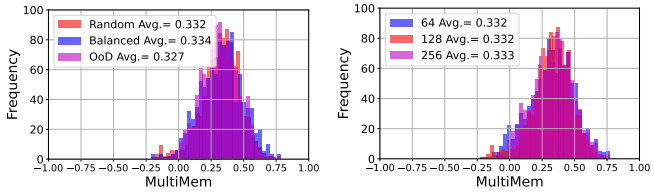
(a) MultiMem (AIT). **(b)** MultiMem (AI). **(c)** MultiMem (AT). **(d)** CLIPMem (IT).

Fig. 1: Memorization should be measured on all modalities instead of only on *modality pairs*. (a) Our MultiMem scores across all three modalities (AIT: **A**udio, **I**mage, and **T**ext) for AudioCLIP. We quantify pairwise memorization on all modality pairs: (b) **A**udio-**I**mage and (c) **A**udio-**T**ext (with MultiMem), and (d) **I**mage-**T**ext (with CLIPMem).

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. We observe the same trends for AVT-CLIP and AVIT-CLIP, as we show in Figure 7, Figure 8, and Figure 10 in Appendix 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.

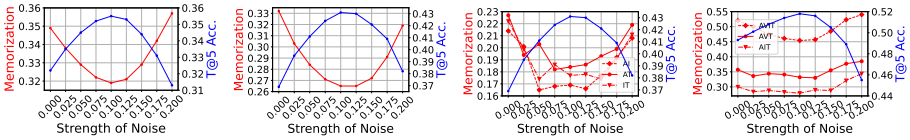
Robustness to Held-out Set. To examine the sensitivity of our MultiMem metric to the choice of held-out set H, we evaluate the metric using several selection strategies for H: (1) randomly sampled data points, (2) data points chosen to achieve balanced class representation, and (3) samples drawn from 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.

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 [27], 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 Appendix A.3 and a complete list of the top



(a) (b)

Fig. 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.



(a) (b) (c) (d)

Fig. 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.

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 [21,27,35]. 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 [27] 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

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	33.8	<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	**0.177**	**0.184**	**0.169**	**0.264**	0.236	0.262	0.269	**35.7**	**33.0**	**29.5**	**43.1**

Table 2: 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.

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 [28] to measure memorization for single modality and use MultiMem for multi-modal memorization. The results in Table 2 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 pretrained 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 Appendix 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 2 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 ℓ_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 Appendix 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 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 [28]. 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.

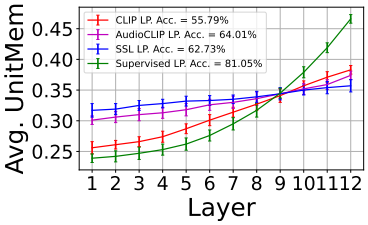


Fig. 4: UnitMem: AudioCLIP is more aligned with SSL.

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

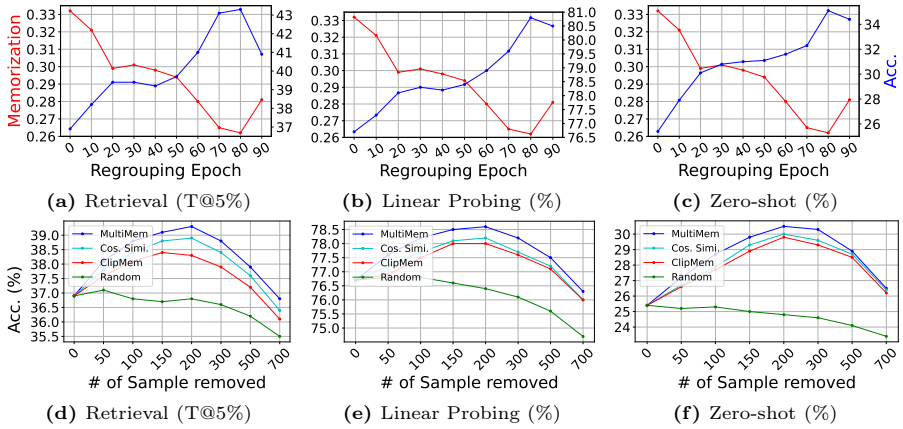


Fig. 5: Mitigation Strategies. The first row shows the impact of in-training memorization mitigation at different training epochs, evaluated on (a) retrieval, (b) linear probing, and (c) zero-shot classification tasks. The second row presents the effect of post-training mitigation with increasing number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

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 Appendix 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.

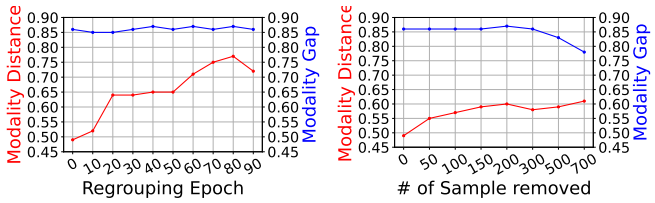
The results yield two main findings. First, applying our mitigation strategy at any training stage effectively reduces global memorization and improves overall performance across all three tasks. Second, model performance in all three tasks follows a trend of initial improvement, followed by a plateau, 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 Appendix 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



(a) In-training Mitigation **(b)** Post-training Mitigation

Fig. 6: Average modality representation distance for 1% most memorized samples versus averagae modality gap for all training samples before and after mitigation

	Post-training				In-training			
Mitigation	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
MultiMem (most)	**39.3%**	**78.6%**	**30.5%**	**0.282**	**43.3%**	**80.8%**	**35.1%**	**0.262**
Augmentation	38.8%	78.4%	30.0%	0.288	42.5%	80.1%	34.0%	0.269

Table 3: Regulation verification by comparing MultiMem with other mitigation strategy.

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 3 show discrepancies: 1) adding noise to random samples does not change performance significantly. 2) Adding noise to least memorized samples slightly degrades performance. 3) 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, 27, 35] has shown that Gaussian noise can indeed 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 (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 last two columns of Table 3, 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 [27] 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. Baldrati, A., Bertini, M., Uricchio, T., Del Bimbo, A.: 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. pp. 4959–4968 (2022)
2. Baldrati, A., Bertini, M., Uricchio, T., Del Bimbo, A.: Effective conditioned and composed image retrieval combining clip-based features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21466–21474 (2022)
3. Bartlett, P.L., Long, P.M., Lugosi, G., Tsigler, A.: Benign overfitting in linear regression. Proceedings of the National Academy of Sciences **117**(48), 30063–30070 (2020)
4. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2818–2829 (2023)
5. Feldman, V.: Does learning require memorization? a short tale about a long tail. In: Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing. pp. 954–959 (2020)
6. Feldman, V., Zhang, C.: What neural networks memorize and why: Discovering the long tail via influence estimation. Advances in Neural Information Processing Systems **33**, 2881–2891 (2020)
7. Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: Proc. IEEE ICASSP 2017. New Orleans, LA (2017)
8. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15180–15190 (2023)
9. Guzhov, A., Raue, F., Hees, J., Dengel, A.: Esresnet: Environmental sound classification based on visual domain models. In: 2020 25th international conference on pattern recognition (ICPR). pp. 4933–4940. IEEE (2021)
10. Guzhov, A., Raue, F., Hees, J., Dengel, A.: Audioclip: Extending clip to image, text and audio. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 976–980. IEEE (2022)
11. Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., Wang, L.: Scaling up vision-language pre-training for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17980–17989 (2022)
12. Jayaraman, B., Guo, C., Chaudhuri, K.: Déjà vu memorization in vision-language models. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. pp. 50722–50749 (2024)
13. Li, S., Tang, H.: Multimodal alignment and fusion: A survey. arXiv preprint arXiv:2411.17040 (2024)
14. Liang, P.P., Zadeh, A., Morency, L.P.: Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. ACM Computing Surveys **56**(10), 1–42 (2024)
15. Liang, V.W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.Y.: Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. Advances in Neural Information Processing Systems **35**, 17612–17625 (2022)

16. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13. pp. 740–755. Springer (2014)
17. Liu, W., Chen, S., Guo, L., Zhu, X., Liu, J.: Cptr: Full transformer network for image captioning. arXiv preprint arXiv:2101.10804 (2021)
18. Meehan, C., Bordes, F., Vincent, P., Chaudhuri, K., Guo, C.: 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)
19. Mu, N., Kirillov, A., Wagner, D., Xie, S.: Slip: Self-supervision meets language-image pre-training. In: European conference on computer vision. pp. 529–544. Springer (2022)
20. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
21. Peng, X., Wei, Y., Deng, A., Wang, D., Hu, D.: Balanced multimodal learning via on-the-fly gradient modulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8238–8247 (2022)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
23. Sadrtdinov, I., Chirkova, N., Lobacheva, E.: On the memorization properties of contrastive learning. arXiv preprint arXiv:2107.10143 (2021)
24. Salamon, J., Jacoby, C., Bello, J.P.: A dataset and taxonomy for urban sound research. In: Proceedings of the 22nd ACM international conference on Multimedia. pp. 1041–1044 (2014)
25. Song, H., Dong, L., Zhang, W., Liu, T., Wei, F.: 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). pp. 6088–6100 (2022)
26. Wang, W., Dziedzic, A., Backes, M., Boenisch, F.: Localizing memorization in ssl vision encoders. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. pp. 60475–60516 (2024)
27. Wang, W., Dziedzic, A., Kim, G.C., Backes, M., Boenisch, F.: Captured by captions: On memorization and its mitigation in clip models. In: The Thirteenth International Conference on Learning Representations (2025)
28. Wang, W., Kaleem, M.A., Dziedzic, A., Backes, M., Papernot, N., Boenisch, F.: Memorization in self-supervised learning improves downstream generalization. In: ICLR (2024)
29. Wang, Y., Yasunaga, M., Ren, H., Wada, S., Leskovec, J.: Vqa-gnn: Reasoning with multimodal knowledge via graph neural networks for visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21582–21592 (2023)
30. Wen, K., Xia, J., Huang, Y., Li, L., Xu, J., Shao, J.: Cookie: Contrastive cross-modal knowledge sharing pre-training for vision-language representation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2208–2217 (2021)
31. Xu, H., Ghosh, G., Huang, P.Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., Feichtenhofer, C.: Videoclip: Contrastive pre-training for zero-shot video-text understanding. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 6787–6800 (2021)

32. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5288–5296 (2016)
33. Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Zero-shot video question answering via frozen bidirectional language models. Advances in Neural Information Processing Systems **35**, 124–141 (2022)
34. Ye, J., Borovykh, A., Hayou, S., Shokri, R.: Leave-one-out distinguishability in machine learning. In: The Twelfth International Conference on Learning Representations (2023)
35. Yuan, Z., Zhang, B., Xu, H., Liang, Z., Gao, K.: 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). pp. 9–18 (2024)
36. Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free adaption of clip for few-shot classification. In: European conference on computer vision. pp. 493–510. Springer (2022)
37. Zhou, Y., Long, G.: 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)