NOT ALL PAIRS ALIGN: CAPTION-GUIDED AUDIO-VISUAL CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We address the problem of learning audio-visual representations from co-occurring modalities. Multimodal contrastive learning typically assumes that paired observations provide shared views of the same underlying information. While this may be true for datasets collected in constrained environments, it does not always hold. For instance, co-occurring audio and visual signals can differ substantially in their content due to off-screen sounds. Nevertheless, standard contrastive objectives treat all paired samples as equally aligned. In this work, we propose **C**aption **G**uided **A**lignment **A**ware **L**oss (**CGAAL**), a model-agnostic objective that leverages modality-specific captions to estimate semantic agreement between audio and video. CGAAL uses these agreement scores to weight the contribution of each positive pair. A warm-up period and gradual thresholding strategy progressively introduces this alignment-aware training signal. We demonstrate the generality of CGAAL by applying it to three contrastive baselines, consistently improving their performance on audio-visual retrieval.

1 INTRODUCTION

Multimodal contrastive learning has emerged as an effective approach for learning transferable representations from naturally paired data, e.g., images–text and audio–video pairs (Radford et al., 2021; Wu et al., 2023; Alayrac et al., 2020; Chen et al., 2021). By treating paired observations as positives and unpaired ones as negatives, these methods learn a shared embedding space without labels. Their success is often explained by the multiview redundancy assumption: each modality shares enough task-relevant information to learn useful representations (Federici et al., 2020).

This assumption is particularly problematic for audio and video. Although audio-visual signals may co-occur in the same video clip, they need not be semantically aligned. For example, a video showing someone playing golf while the audio contains speech unrelated to the game. Under the multiview redundancy assumption, standard contrastive learning nevertheless encourages strong alignment of all pairs, introducing a potentially misleading training signal (see Figure 1).

Recent work challenges the assumption that multimodal observations carry only redundant information, distinguishing shared information from information that is unique to one modality or that emerges only through their interaction (Dufumier et al., 2025; Wen et al., 2026). These methods address this by learning a single representation that captures shared, unique, and synergistic components, and have shown clear gains over purely redundancy-based objectives. We approach the same

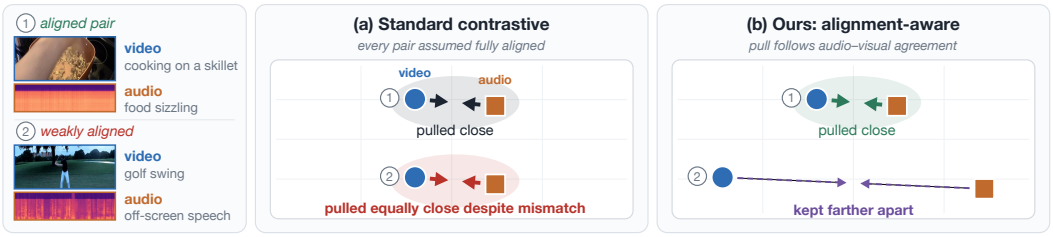

Figure 1: **Alignment in audio-video contrastive learning**. Standard contrastive methods assume strong semantic correspondence between modalities of a given sample, encouraging alignment even when the modalities share little semantic content. This can encourage the model to associate audio and visual content that does not correspond. We thus propose CGAAL which determines how strongly to pull the modalities' embeddings of a sample according to a pair's semantic agreement.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

problem from a different angle. Rather than explicitly capturing unique and shared information, we act on the alignment objective itself addressing how strongly modalities in a pair should be aligned.

We introduce **CGAAL**[1], a **C**aption **G**uided **A**lignment **A**ware **L**oss, which incorporates an estimate of each pair's semantic correspondence into contrastive training. From modality-specific captions describing the audio and visual content, we use an LLM to estimate how closely the two modalities semantically align. These scores weight the positive pairs during contrastive training, placing less emphasis on pairs that share little semantic content (see Figure 1). We combine this weighting with a curriculum learning inspired approach to incorporate the alignment scores progressively. Importantly, CGAAL requires neither captions nor an LLM at inference time, and can be integrated into different contrastive learning frameworks. Our contributions are:

(1) We identify variable semantic correspondence as a fundamental limitation of audio-visual contrastive learning and show that methods designed to distinguish shared from unique, and synergistic information do not resolve this issue for cross-modal retrieval.
(2) We propose to use modality-specific captions as a guide for semantic alignment between modalities in each sample by incorporating an alignment score into the training loss via a weighting scheme, adjusting how strongly paired embeddings are pulled together.
(3) We demonstrate our method's ability to improve audio visual retrieval performance across a range of contrastive learning models and show zero-shot capabilities on three unseen datasets.

2 Related Work

Multimodal contrastive learning uses naturally paired data to align representations from different modalities, typically through an InfoNCE loss (Oord et al., 2018). A prominent example is CLIP (Radford et al., 2021), which contrasts image-text pairs to learn a shared embedding space. Similar contrastive objectives have since been applied to other modality pairings, such as audio-text (Wu et al., 2023), audio-visual (Parida et al., 2026), video-text (Li et al., 2026) and point cloud-text (Paul et al., 2026), as well as extended to jointly align three or more modalities (Chen et al., 2023; Liu et al., 2024; Shvetsova et al., 2022; Guzhov et al., 2022; Wang et al., 2024; Girdhar et al., 2023). These approaches apply contrastive alignment to global representations of each input. Other work instead considers alignment at finer granularities. For instance, video-text retrieval approach X-CLIP (Ma et al., 2022) computes similarities between videos, frames, sentences and words. More recently, FDCA (Yue Wu, 2025) first separates the text at sentence and token level and then aligns them to the video. Audio-video retrieval approaches have also considered finer-grained alignment: CAV-MAE (Araujo et al., 2025) aligns visual frames with temporally localised audio segments, while Fan et al. (2026) improve spatial alignment through multiscale spatial attention that aligns audio with particular objects and events. Despite differing in the number and granularity of the modalities they align, these approaches all treat naturally co-occurring modalities as positives pairs regardless of how much semantic content the modalities share. In contrast, we estimate a modality alignment score and use this to control how strongly each pair should be aligned.

Learning with partial cross-modal correspondence. Several approaches have attempted to relax the assumption that paired samples are equally aligned. Similar to our work, soft-weighted objectives have been used by Srinivasa et al. (2023); Lu et al. (2025); Fradkin et al. (2024). However, to obtain the soft-weighting, these approaches first train a model under the multiview-redundancy assumption, assuming all pairs are strongly aligned. The similarities in this embedding space then act as the soft-weights to guide the partial cross-modal correspondence. We argue that this approach is fundamentally problematic as the model trained with the multi-view redundancy assumption will have already lost much of the nuance in a pair's alignment. Other related works assume partial alignment in image-text paired data where only a part of the image should be aligned with certain words in the caption (Zhou et al., 2024; Rao et al., 2025; Huang et al., 2021a). All these works assume that some alignment between part of the image and words in the caption exist. This is not the case in audio-visual data where at times the audio and the vision do not align at all (random music or speech that is not relevant to the visuals). Closer to our work is Sounding Actions (Chen et al., 2024) which uses text to establish a consensus signal for audio-visual alignment. Their focus is on non-audible actions, i.e. actions for which audio is silent. They learn to align audio and vision when both correspond to the action described in the visual caption, suppressing other audio–visual associations even when the sound and visuals themselves agree. In contrast, we consider any semantic

[1]We pronounce 'CGAAL' as 'Seagaal' for simplicity.

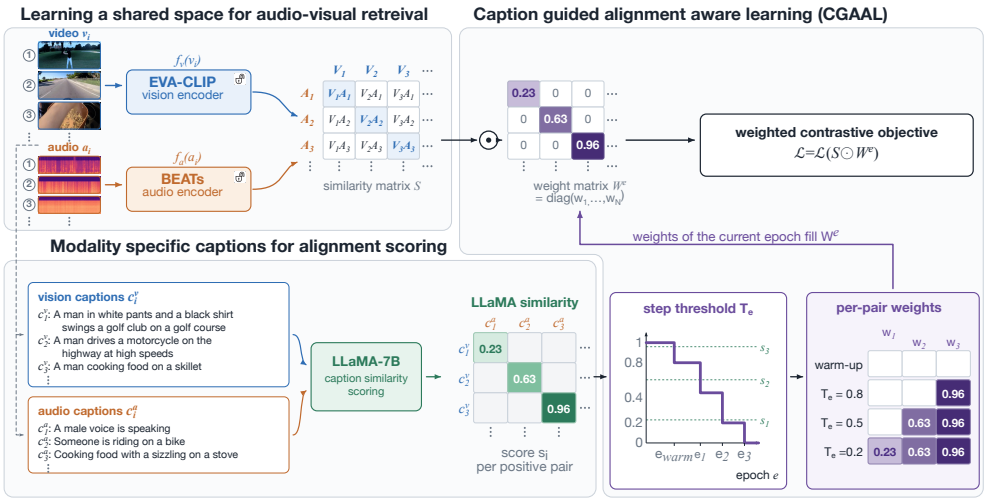

Figure 2: **Overview of our caption-guided alignment-aware learning (CGAAL)**. As in standard contrastive learning, we (a) learn a shared space for audio-visual retrieval where audio and video encoders project each modality into a shared embedding space where we measure their similarity. Rather than treating all co-occuring pairs as equally positive, we instead use (b) modality specific captions to estimate the alignment of each positive pair. After first warming up the representation with all positive pairs, CGAAL uses the estimated alignment scores to weight positive pairs in a contrastive loss (c). A gradually decreasing threshold introduces the alignment scores gradually so that the strongest correspondences influence the loss first. This produces a shared embedding space in which proximity better reflects audio–visual correspondence, rather than co-occurrence alone.

correspondence between the audio and visual streams, and use its estimated strength to determine how much each pair contributes to learning an audio–visual representation.

Learning beyond shared information. Several recent works aim to go beyond redundant information shared across modalities and also capture unique and/or synergistic information. FactorCL (Liang et al., 2023) factorizes task-relevant multimodal information into shared and modality-specific components. It uses multimodal augmentations to approximate which information is relevant to the task, and contrastive mutual-information objectives to preserve this information while suppressing task-irrelevant content. CoMM (Dufumier et al., 2025) contrasts augmented fused representations rather than aligning unimodal embeddings pairwise, allowing it to capture synergistic interactions alongside shared and unique information. InfMasking (Wen et al., 2026) extends this approach by masking different subsets of modality features during fusion to encourage a broader range of synergistic interactions. All three methods were intended for classification tasks where their assumptions behind augmentations, such as minimal label preserving assumptions, hold since the modalities of each sample share a single label. We experimentally show that this does not hold in audio-video retrieval, where fine-grained multimodal alignment is important. Instead, we incorporate a prior signal of the degree of agreement between modalities using their captions as a proxy to inform the model on how semantically aligned they are.

3 Method

Our aim is to learn audio-visual representations that account for varying semantic correspondence between naturally paired modalities, rather than treating all pairs as equally valid positives. Figure 2 provides an overview of our approach: given the audio-visual retrieval setting introduced in Section 3.1, we use modality specific captions to estimate the semantic alignment of each pair (Section 3.2) and incorporate these scores into our alignment-aware loss in (Section 3.3).

3.1 Learning a Shared Space for Audio-Visual Retrieval

Given a dataset of N paired audio-visual samples $\{(v_i, a_i)\}_{i=1}^{N}$, where v_i is a visual frame sampled from a video clip and a_i is the accompanying audio from the same clip, audio-visual retrieval aims to learn a shared embedding space in which related samples can be found across modalities. To learn this space, modality specific encoders, F_a and F_v, are trained to project the audio and vision

modalities to features $F_a(a_i)$ and $F_v(v_i)$ in a shared space where similarity can be measured directly. Standard contrastive learning approaches, such as CLIP, do this via an InfoNCE loss. For each audio sample a_i in a batch of size B, the visual sample from the same video clip v_i is treated as a positive, while visual samples from other clips v_j serve as negatives. The audio-to-visual loss encourages the positive pair to have higher similarity than these alternatives:

$$\mathcal{L}^{\text{InfoNCE}}_{\mathcal{A}\to\mathcal{V}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\big(\langle F_a(a_i), F_v(v_i)\rangle/\tau\big)}{\sum_{j=1}^{B}\exp\big(\langle F_a(a_i), F_v(v_j)\rangle/\tau\big)} \tag{1}$$

Here, τ is the temperature and $\langle\cdot,\cdot\rangle$ denotes the dot product. The visual-to-audio loss ($\mathcal{L}^{\text{InfoNCE}}_{\mathcal{V}\to\mathcal{A}}$) is defined analogously and the overall contrastive loss averages these two directions. At inference, this allows for retrieval based on the dot product between representations, ranking candidates according to their similarity to (or distance from) the query.

By using co-occurrence as supervision, InfoNCE learns a shared retrieval space without requiring manual labels. However, the co-occurrence assumption is only appropriate when modalities describe the same semantics, for example, footage of a dog barking corresponds closely to the sounds of its bark. However, a paired sample may only be partially relevant semantically, such as a video of someone walking through a park which contains the sound of footsteps alongside a dog barking off-screen. While it's likely to hear a dog barking in a park, the alignment is only partial as the dog is not visible in view. Both pairs nevertheless enter InfoNCE as positives when using the co-occurrence supervision, creating pressure to align them to the same degree regardless of how tightly they actually correspond. In this paper, we aim to learn the shared audio-visual embedding space while accounting for this variation in the semantic alignment. This is beneficial for retrieval as candidates are ranked relative to one another. Pulling weakly corresponding pairs together can blur the differences in similarity needed to place a closely matching candidate above a less relevant one.

3.2 Modality Specific Captions for Alignment Scoring

To account for varying semantic alignment between paired modalities, we first estimate the degree of semantic correspondence for each audio-visual pair. This can then be used to control how strongly the pair is pulled together during training. One option is to annotate correspondence manually, however this is expensive. Alternatively, we could use pretrained embeddings, though this estimate could be misleading as unimodal models do not consider cross-modal correspondence, and typical contrastive multimodal models may have already supressed information unique to each modality. Instead, we propose to use modality specific captions for the video, c_i^v, and audio, c_i^a, as proxies of their semantic content. These captions may be human annotated or generated using modality-specific captioning models. We hypothesise that unimodal captioning methods have been sufficiently developed to well describe the semantics in a single modality, making these separate captions a practical basis for comparing the semantics of two modalities. Large-language models (LLMs) can then be used to compare these captions and estimate their potential alignment.

For each audio-visual sample, the LLM is instructed to predict an alignment score, s_i, in the range $[0, 1]$, where 0 indicates no semantic correspondence and 1 indicates complete semantic correspondence. We leverage in-context learning, providing three highly aligned, three moderately aligned and three weakly aligned caption pairs, together with our proposed alignment scores. Figure 3 illustrates the LLM-based alignment scoring procedure which we formalise as F_L:

$$s_i = F_L(c_i^v, c_i^a) \in [0, 1]. \tag{2}$$

We show sample prediction scores in Figure 4 which illustrates how the modality-specific captions act as proxies for the semantic content of the underlying vision and audio modalities, and how differences in their semantic correspondence are reflected in the resulting alignment scores. Highly aligned pairs contain substantial shared semantic information across the captions, while partially aligned pairs contain both shared and modality-specific information. For example, in the moderately aligned sample in the second column, the vision caption uniquely provides the highway context, which cannot be heard in the audio, while the remaining information is shared across modalities. In weakly aligned pairs, the captions contain less shared semantic information, indicating lower estimated semantic correspondence between the underlying modalities.

Prompt for LLM-based audio–visual alignment scoring

Step 1: Task definition

Your task is to compare a vision caption and an audio caption and determine how semantically aligned they are.

Scoring rules:

- Return a decimal number between 0.0 and 1.0.
- Consider actions, objects, environments, events, and overall scene meaning.

Step 2: In-context examples of varying alignment levels

c_i^v: A dog runs across a grassy park chasing a ball.
c_i^a: A dog barking and running outdoors while people play.
Explanation: Both captions describe a dog outdoors engaged in active play in a park.
Score: 0.91

Step 3: Inference on a vision--audio caption pair from the dataset

c_i^v: A man in white pants and a black shirt swings a golf club on a golf course.
c_i^a: A male voice is speaking.
Score: —

LLM output: 0.23

Figure 3: **Alignment score estimation.** We prompt the LLM to predict the alignment score by: (1) setting the criteria for measuring the semantic alignment with clear inputs (modality captions) and the output (alignment score); (2) providing examples with different alignment strength alongside an explanation to calibrate the scores across various examples; and (3) providing the vision and audio captions pair for which a score needs to be predicted.

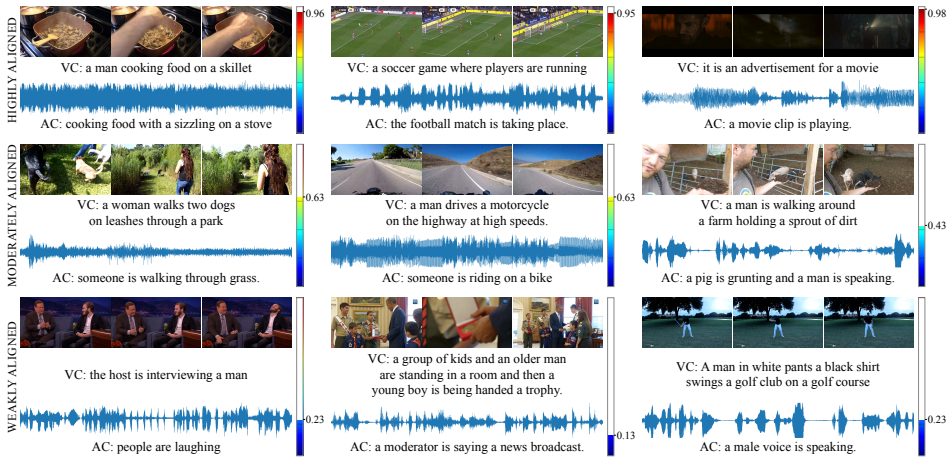

Figure 4: **Examples of modality-specific captions with varying semantic alignment.** We show high (top), moderate (middle) and weak (bottom) alignment examples. High (row1,col1): vision shows cooking on stove while audio has distinct sizzling sound. Moderate (row2,col3): walking is not evident from the audio. Weak (Row3 col 3): no evidence of golfing in the audio.

3.3 Caption Guided Alignment Aware Learning (CGAAL)

We use our caption-guided alignment scores to vary the contribution of positive pairs to the contrastive objective: pairs with less shared content provide weaker evidence for learning cross-modal correspondence. We first define how the scores are used to weight the loss function, then describe how an initial warm-up and gradual thresholding introduce these weights during training.

Alignment weighting. Each positive pair provides a different amount of evidence for audio–visual correspondence. We therefore weight each pair's contribution to the contrastive loss according to its caption-derived alignment score. Let w_i denote the weight of the modality pairs, (v_i, a_i), of the positive sample in the batch B. Then the audio-to-video loss becomes:

$$\mathcal{L}_{\text{CGAAL},\,\mathcal{A}\to\mathcal{V}} = -\frac{1}{\sum_{i=1}^{B} w_i} \sum_{i=1}^{B} w_i \log \frac{\exp\big(\langle F_a(a_i), F_v(v_i)\rangle/\tau\big)}{\sum_{j=1}^{B} \exp\big(\langle F_a(a_i), F_v(v_j)\rangle/\tau\big)} \tag{3}$$

Again τ is the temperature, the video-to-audio loss is defined analogously and the overall loss is the mean of audio-to-video and video-to-audio losses. Due to the simplicity of our reweighting approach, it can be easily integrated with different contrastive learning objectives. Next we describe how this varies under gradual thresholding and warm-up.

Warm-up period. At the start of training, the encoders have not yet learned a shared audio–visual space. We thus train a model for e_{warmup} epochs with standard InfoNCE, setting $w_i = 1$ for all positive pairs during this short period. This warms up the encoders with the training data, removing any modality shifts between pretraining and training data.

Gradual thresholding. After the warm-up period, we gradually introduce the alignment scores as sample weights progressively. Motivated by curriculum learning, this allows the model to learn easier samples with more obvious alignment first. We initialise an epoch-dependent threshold, $T_{e_{warmup}}$, which determines the samples for which the caption-derived alignment score s_i is activated as the effective weight, i.e., $w_i = s_i$, immediately after warm-up. A sample uses its alignment score as its weight only when the score exceeds the current threshold. Every Δe epochs, the threshold is reduced according to $T_e = \max(0, T_{e-\Delta e} - \Delta T)$, where ΔT is a hyperparameter to control the rate of threshold reduction. As the threshold decreases, progressively lower-alignment samples receive their corresponding alignment scores as weights, until T_e reaches zero and all samples are alignment-weighted. This means that w_i is not fixed every epoch, therefore, we introduce epoch-dependent weights w_i^e into our loss:

$$\mathcal{L}_{\text{CGAAL},\,\mathcal{A}\to\mathcal{V}} = -\frac{1}{\sum_{i=1}^{B} w_i^e} \sum_{i=1}^{B} w_i^e \, \log \frac{\exp\big(\langle F_a(a_i), F_v(v_i)\rangle/\tau\big)}{\sum_{j=1}^{B} \exp\big(\langle F_a(a_i), F_v(v_j)\rangle/\tau\big)} \tag{4}$$

Overall, this method encourages the model to focus on bringing the audio and vision representations of aligned samples closer together in the embedding space since they have a higher weight in the loss. When the modalities contain more unique information, the alignment score is lower, and the loss places less emphasis on pulling that pair together. In this way, CGAAL reduces the pressure to align all co-occurring pairs as though they are only multiple views of the same shared information.

4 Experiments

We first describe the dataset used and implementation details before evaluating CGAAL against contrastive learning baselines and assessing its zero-shot transfer ability. We then ablate the contributions of alignment weighting, gradual thresholding, and warm-up.

4.1 Dataset Details

Pretraining. We use a subset of the VAST-27M dataset (Chen et al., 2023) constituting a randomly selected 100K training samples and 1927 test samples. The dataset provides 5 vision captions and 5 audio captions per sample which were automatically generated using modality-specific captioning models. We compute an alignment score for each of the 25 possible audio-visual caption pairs and average these scores to obtain a single alignment score for each sample.

Synthetic Misalignment. We create a second pretraining set to stress test our method's ability to learn a robust embedding space when challenged with additional misaligned samples. For 15% of the training samples, we randomly shuffle the audio among samples. We chose 15% to resemble the expected noise level in data collected from the internet (Huang et al., 2021b; Sharma et al., 2018). This gives us two training conditions: (1) Without Synthetic Misalignment, which only includes the natural misalignment in the VAST-27M subset and (2) With Synthetic Misalignment, which includes an additional synthetic source of misalignment for 15% of the training data. In both settings, evaluation is performed on the original, uncorrupted test set. Figure 5 compares the distribution of alignment scores for the two pretraining sets. Without synthetic misalignment the pretraining set already contains many pairs with low estimated alignment, motivating a method that accounts for varying correspondence rather than treating every pair identically. Shuffling the audio shifts the selected pairs further towards lower scores, creating a more challenging scenario.

Zero-shot. We test whether CGAAL improves audio-video retrieval on datasets unseen during training with AVE (Tian et al., 2018), DiDeMo (Hendricks et al., 2018) and MSR-VTT (Xu et al., 2016). AVE focuses on 28 types of audio events which are also clearly visible. DiDeMo contains unedited personal videos from Flickr while MSR-VTT contains short web videos spanning diverse

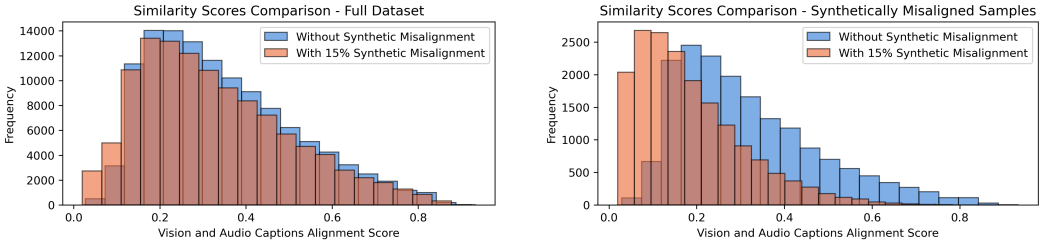

Figure 5: **Impact of Synthetic Misalignment on Training Set**. We show the distributions of caption-based alignment scores for the full pretraining set (left) and the 15% of the samples selected for audio shuffling (right). The original pretraining set (blue) already contains many low-scoring pairs, suggesting substantial natural variation in audio–visual correspondence. Adding synthetic misalignment (orange) shows a clear shift towards lower scores.

categories. After excluding clips with no audible audio, the evaluation sets contain 402, 997 and 2617 videos for AVE, DiDeMo and MSR-VTT respectively.

4.2 Experimental Setup

Implementation Details. Following VAST (Chen et al., 2023), we use BEATs as the audio encoder F_a and the EVA-Clip vision encoder as F_v with their original pretrained weights. We fine-tune both encoders end-to-end during pretraining for 48 epochs using Adam with a batch size of 32, a learning rate of 1e-5 and weight decay of 1e-2. For each model we integrate our method with, we tune the hyperparameters: e_{warmup}, Δe and ΔT. Unless otherwise specified, we use our model in combination with a CLIP-like contrastive objective where we use e_{warmup}=2, Δe=1 and set ΔT=0.3 for the first two gradual thresholding epochs and ΔT=0.4 for the third gradual thresholding epoch. For SLIP (+Ours) and CMC (+Ours), we use e_{warmup}=2, Δe=2 and ΔT=0.2. We choose LLaMa-7B as the LLM for the alignment scores prediction.

Evaluation Metrics. We report recall R@K for K=1,5,10 alongside median rank for both audio-to-video and video-to-audio retrieval. To summarize recall, we use RSum: the sum of all six recall values across both directions.

4.3 Comparative Results

Baselines. We compare CGAAL with two groups of methods. CoMM (Dufumier et al., 2025) and InfMasking (Wen et al., 2026) aim to learn beyond what modalities share, additionally aiming to capture unique and synergistic information in the learned embeddings. They provide a comparison with approaches that also question whether shared information alone is sufficient for multimodal learning, but pursue this through different objectives. Our standard contrastive baselines are Cross+Self (Yuan et al., 2021), SLIP (Mu et al., 2022), CMC (Tian et al., 2019) and CLIP (Radford et al., 2021) following Dufumier et al. (2025) and Wen et al. (2026). These methods differ in which views they contrast, how the negatives are formed and whether cross-modal learning is combined with within-modal objectives. CLIP uses one cross-modal term with a learned temperature and negatives only from the opposing modality. CMC averages a fixed-temperature term over all view pairs, with negatives from both views. SLIP adds a separately-scaled SimCLR term to CLIP. Cross+Self uses one contrastive term throughout, pairing a cross-modal term with an averaged intra-modal term for each modality. We apply CGAAL to CLIP, SLIP, and CMC to examine whether accounting for pairwise correspondence benefits different contrastive objectives. All methods use the same pretrained EVA-CLIP vision encoder and BEATs audio encoder for a fair comparison.

Results. Table 1 reports results for audio-to-video (A2V) and video-to-audio (V2A) retrieval on the VAST-27M subset after training with and without synthetic misaligment. CoMM and InfMasking obtain lower results than standard contrastive baselines both with and without synthetic misalignment. This suggests that their approaches to capturing unique and synergistic information do not transfer well to retrieval. Their objectives focus on improving the combined representation of multiple modalities which does not necessarily help the audio rank its matching video above others.

Adding CGAAL consistently improves the retrieval performance of SLIP, CMC and CLIP. These methods differ in how they construct their contrastive objectives, so the gains indicate that accounting for alignment is useful across several ways of learning audio-visual representations. CLIP is the

Table 1: **Comparative Results.** Combining our approach with standard contrastive baseline (SLIP, CMC and CLIP) improves retrieval performance both with and without synthetic misalignment.

	Without Synthetic Misalignment									With Synthetic Misalignment								
	A2V				V2A					A2V				V2A				
	R1 ↑	R5 ↑	R10 ↑	MR ↓	R1 ↑	R5 ↑	R10 ↑	MR ↓	RSum ↑	R1 ↑	R5 ↑	R10 ↑	MR ↓	R1 ↑	R5 ↑	R10 ↑	MR ↓	RSum ↑
InfMasking	1.2	5.7	9.0	159	1.9	6.18	10.4	153	34.4	1.1	4.1	6.6	233	1.3	4.4	7.1	253	24.6
CoMM	5.7	18.3	26.5	41	5.2	17.7	26.3	38	99.7	4.3	14.3	22.6	51	5.1	16.0	24.4	47	86.7
Cross+Self	10.5	28.2	39.2	18	10.3	28.4	39.3	19	155.9	10.0	27.0	38.0	21	9.0	26.2	37.1	21	147.3
SLIP	12.9	34.3	46.0	13	13.0	33.8	45.3	13	185.3	11.4	29.9	41.2	17	11.9	29.9	40.7	18	165.0
SLIP (+Ours)	14.1	35.0	47.3	12	14.1	33.9	46.2	13	190.6	12.4	31.4	42.8	15	12.2	31.8	43.1	15	173.7
ΔSLIP	+1.2	+0.7	+1.3	+1	+1.1	+0.1	+0.9	0	+5.3	+1.0	+1.5	+1.6	+2	+0.3	+1.9	+2.4	+3	+8.7
CMC	14.7	37.6	49.5	11	16.9	38.3	49.8	11	206.8	13.0	34.4	45.4	14	13.9	34.4	**46.2**	**13**	187.3
CMC (+Ours)	15.5	38.6	50.3	10	17.4	39.8	50.9	11	212.5	13.4	34.8	45.7	13	14.2	34.8	46.1	**13**	189.0
ΔCMC	+0.8	+1.0	+0.8	+1	+0.5	+1.5	+1.1	0	+5.7	+0.4	+0.4	+0.3	+1	+0.3	+0.4	-0.1	0	+1.7
CLIP	15.9	38.3	50.5	10	17.2	38.8	50.8	10	211.5	13.0	33.5	44.4	14	13.3	33.0	43.5	16	180.7
CLIP (+Ours)	**17.9**	**41.1**	**52.6**	**9**	**17.9**	**40.6**	**52.0**	**9**	**222.1**	**14.2**	**36.1**	**47.3**	**12**	**14.5**	**35.9**	46.0	**13**	**194.0**
ΔCLIP	+2.0	+2.8	+2.6	+1	+0.7	+1.8	+1.2	+1	+10.6	+1.2	+2.6	+2.9	+2	+1.2	+2.9	+2.5	+3	+13.3

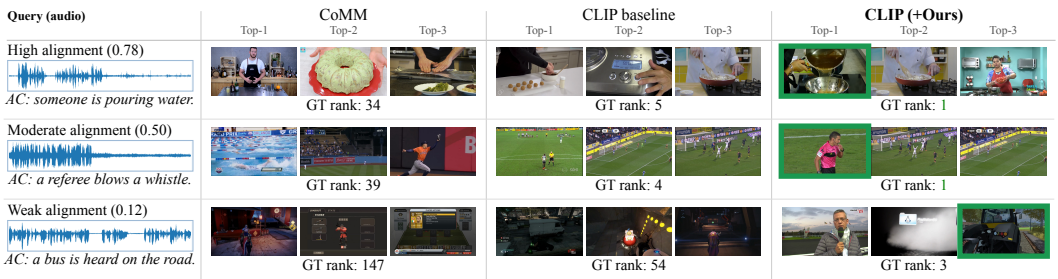

Figure 7: **Qualitative Results** for audio-to-video retrieval over different levels of alignment. Our method shows better fine-grained semantic correspondence between the audio and retrieved videos such as in the top row where ours retrieves pouring water and others retrieve general cooking videos.

strongest contrastive baseline, leading to CLIP (+Ours) achieving the best overall performance for both V2A and A2V retrieval for all but one metric.

Importantly, CGAAL improves retrieval even when no synthetic misalignment is introduced. Co-occurring audio and vision in the original data do not always describe the same content, so positive pairs already differ in how much evidence they provide for cross-modal alignment. We make this challenge more pronounced by introducing synthetic misalignment. With this, performance falls for the standard contrastive baselines, but adding CGAAL improves both A2V and V2A retrieval. The gains in both settings support the central premise of our method: audio and video should be aligned according to how closely their content matches, rather than simply because they occur together.

4.4 Performance Analysis

To better understand our results, we examine the relation between the performance improvement and alignment score and show qualitative examples.

Performance vs. alignment score. In Figure 6 we compare the retrieval performance with RSum for samples with varying levels of alignment score. CLIP has a small advantage for samples with low alignment (0.0–0.2). These pairs remain correct matches under the retrieval metric despite having little estimated semantic correspondence. Treating every pair equally during training may therefore help CLIP retrieve them. Our approach performs better in every other range, with its largest gain for the most strongly aligned pairs (0.8–1.0). This pattern is consistent with CGAAL placing greater emphasis on pairs whose modalities share more information.

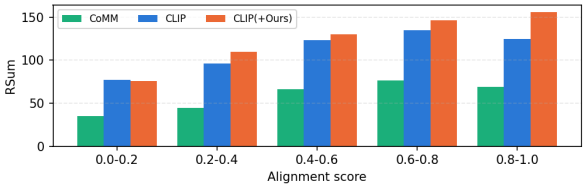

Figure 6: **Performance vs. alignment score**. The relative improvement of our baseline compared to CLIP and CoMM increases with alignment score.

Qualitative results. We show audio-to-video retrieval for three samples with different levels of alignment. Audio captions are displayed only to help readers interpret each query. At inference, the model uses only audio and video modalities, without captions or alignment scores. Across these examples, our approach appears to capture more specific audio–visual correspondence, whereas

Table 2: **Zero-Shot Retrieval.** Adding CGAAL improves a model's ability to generalise to unseen datasets, particularly with increased misalignment in training.

		Without Synthetic Misalignment									With Synthetic Misalignment								
		A2V				V2A					A2V				V2A				
		R1 ↑	R5 ↑	R10 ↑	MR ↓	R1 ↑	R5 ↑	R10 ↑	MR ↓	RSum ↑	R1 ↑	R5 ↑	R10 ↑	MR ↓	R1 ↑	R5 ↑	R10 ↑	MR ↓	RSum ↑
AVE	**CLIP**	6.0	**21.6**	**31.6**	25	8.7	**25.6**	36.6	23	130.1	5.0	19.4	28.4	31	**5.7**	20.9	31.3	27	110.7
	CLIP (+Ours)	**7.7**	21.1	30.6	**23.0**	**10.0**	25.4	**37.1**	**18**	**131.9**	**6.7**	**21.1**	**31.1**	**23.0**	5.5	**26.6**	**35.1**	**21.0**	**126.1**
	ΔCLIP	+1.7	-0.5	-1.0	+2.0	+1.3	-0.2	+0.5	+4	+1.8	+1.7	+1.7	+2.7	+8	-0.2	+5.7	+3.8	+6	+15.4
DiDeMo	**CLIP**	2.9	10.3	15.6	67	**3.1**	10.8	17.4	58	60.1	2.2	7.7	12.7	86	1.2	6.6	12.4	79	42.8
	CLIP (+Ours)	**3.1**	**10.7**	**18.0**	**60**	2.5	**13.3**	**20.1**	**53**	**67.7**	**2.3**	**9.3**	**15.6**	**71**	**3.0**	**10.6**	**17.2**	**55**	**58.0**
	ΔCLIP	+0.2	+0.4	+2.4	+7	-0.6	+2.5	+2.7	+5	+7.6	+0.1	+1.6	+2.9	+15	+1.8	+4.0	+4.8	+24	+15.2
MSR-VTT	**CLIP**	3.9	13.5	20.2	82	4.8	14.8	21.8	71	79.0	3.8	12.1	17.2	114	4.1	12.7	19.1	113	69.0
	CLIP (+Ours)	**4.7**	**15.6**	**21.7**	**70**	**4.9**	**16.8**	**24.1**	**67**	**87.8**	**4.1**	**14.1**	**20.8**	**83**	**4.6**	**14.8**	**21.6**	**74**	**80.0**
	ΔCLIP	+0.8	+2.1	+1.5	+12	+0.1	+2.0	+2.3	+4	+8.8	+0.4	+2.0	+3.6	+31	+0.5	+2.1	+2.5	+39	+11.0

Table 3: **Model Ablation.** Alignment weighting contributes most to our performance. Gradual thresholding gives a further gain with synthetic misalignment, while warm-up improves without it.

	Without Synthetic Misalignment									With Synthetic Misalignment								
	A2V				V2A					A2V				V2A				
	R1 ↑	R5 ↑	R10 ↑	MR ↓	R1 ↑	R5 ↑	R10 ↑	MR ↓	RSum ↑	R1 ↑	R5 ↑	R10 ↑	MR ↓	R1 ↑	R5 ↑	R10 ↑	MR ↓	RSum ↑
CLIP	15.9	38.3	50.5	10	17.2	38.8	50.8	10	211.5	13.0	33.5	44.4	14	13.3	33.0	43.5	16	180.7
+ alignment weighting	18.0	40.9	52.0	9	17.6	40.3	52.0	9	220.8	14.8	35.0	46.6	13	13.7	34.3	44.8	14	189.2
+ gradual thresholding	17.6	40.5	52.0	9	17.6	40.6	52.0	9	220.3	14.6	36.6	47.1	12	14.8	35.3	46.0	13	194.4
+ warm-up	17.9	41.1	52.6	9	17.9	40.6	52.0	9	222.1	14.2	36.1	47.3	12	14.5	35.9	46.0	13	194.0

CLIP favours scenes broadly associated with a sound. For example, given a whistle, CLIP is able to retrieve football scenes, but CLIP (+Ours) ranks the referee video first.

4.5 Zero-Shot Generalisation

We examine whether CGAAL's retrieval gains generalise to unseen data on AVE, DiDeMo and MSR-VTT, without fine-tuning on these datasets. Table 2 compares CLIP and CLIP (+Ours) trained with and without synthetic misalignment. CGAAL improves RSum on all three datasets in both settings, showing that it results in a more transferable representation. The gains are particularly strong when training includes synthetic misalignment and when the test dataset contains less clear audio-visual correspondences (MSR-VTT and DiDeMo).

4.6 Ablations

We ablate the components of CGAAL in Table 3. Starting from the CLIP baseline, we progressively add (1) weighting the loss with alignment scores, (2) introducing alignment scores with gradual thresholding, (3) adding the warm-up component prior to the gradual thresholding. Weighting positive pairs by their alignment improves every retrieval metric over CLIP. This demonstrates the effectiveness of using modality-specific captions as a proxy for alignment and incorporating this as a prior signal to reduce the weight of less aligned samples. Gradual thresholding further improves the overall results (RSum) when synthetic misalignment is present. Its greater benefit in the synthetic setting is consistent with progressively limiting the influence of the additional weakly aligned pairs. Adding the warm-up improves the overall performance without synthetic misalignment and maintains good performance with synthetic misalignment. An initial period of equal weighting may help establish shared structure when positive pairs are generally informative, while offering less benefit when more pairs have been deliberately misaligned.

5 Conclusions

In this work, we explored the challenge of partial correspondence in multimodal contrastive learning for audio-visual retrieval. We proposed CGAAL, which uses modality-specific captions as proxies to estimate the semantic alignment of the modalities of paired signals and weights pairs in the contrastive loss according to this alignment. This allows for more effective learning of the commonality between modalities, resulting in better audio-video retrieval. Furthermore, we show our method is effective with a variety of contrastive learning approaches, can handle both naturally occurring and additional synthetic misalignment, and improves generalisability to unseen data. Together, these results show the value of accounting for varying degrees of correspondence between positive pairs.

REFERENCES

Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-Supervised MultiModal Versatile Networks. In *NeurIPS*, 2020.

Edson Araujo, Andrew Rouditchenko, Yuan Gong, Saurabhchand Bhati, Samuel Thomas, Brian Kingsbury, Leonid Karlinsky, Rogerio Feris, James R. Glass, and Hilde Kuehne. Cav-mae sync: Improving contrastive audio-visual mask autoencoders via fine-grained alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, Angie Boggust, Rameswar Panda, Brian Kingsbury, Rogerio Feris, David Harwath, James Glass, Michael Picheny, and Shih-Fu Chang. Multimodal clustering networks for self-supervised learning from unlabeled videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8012–8021, October 2021.

Changan Chen, Kumar Ashutosh, Rohit Girdhar, David Harwath, and Kristen Grauman. Soundingactions: Learning how actions sound from narrated egocentric videos. In *CVPR*, 2024.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36:72842–72866, 2023.

Benoit Dufumier, Javiera Castillo Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *International Conference on Learning Representations*, volume 2025, pp. 5408–5432, 2025.

Baoyu Fan, Lu Liu, Xiaochuan Li, Runze Zhang, Liang Jin, and Jin Zhang. Fine-grained audio–visual event localization. *IEEE Transactions on Neural Networks and Learning Systems*, 37(1):52–66, January 2026. ISSN 2162-2388. doi: 10.1109/tnnls.2025.3600878. URL http://dx.doi.org/10.1109/TNNLS.2025.3600878.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck, 2020. URL https://arxiv.org/abs/2002.07017.

Philip Fradkin, Puria Azadi, Karush Suri, Frederik Wenkel, Ali Bashashati, Maciej Sypetkowski, and Dominique Beaini. How molecules impact cells: Unlocking contrastive phenomolecular retrieval. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 110667–110701. Curran Associates, Inc., 2024. doi: 10.52202/079017-3513. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c793577b644268259b1416464a6cdb8c-Paper-Conference.pdf.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190. IEEE, 2023.

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)*, pp. 976–980. IEEE, 2022.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with temporal language. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1380–1390, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1168. URL https://aclanthology.org/D18-1168/.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3942–3951, 2021a.

Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 29406–29419. Curran Associates, Inc., 2021b. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f5e62af885293cf4d511ceef31e61c80-Paper.pdf.

Guangzhao Li, Yanming Yang, Chenxi Song, Xiaohong Liu, and Chi Zhang. Flowdirector: Training-free flow steering for precise text-to-video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7805–7815, June 2026.

Paul Pu Liang, Zihao Deng, Martin Q Ma, James Y Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. *Advances in Neural Information Processing Systems*, 36:32971–32998, 2023.

Jing Liu, Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, and Jinhui Tang. Valor: Vision-audio-language omni-perception pretraining model and dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):708–724, 2024.

MingYu Lu, Ethan Weinberger, Chanwoo Kim, and Su-In Lee. Cellclip - learning perturbation effects in cell painting via text-guided contrastive learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 124505–124537. Curran Associates, Inc., 2025. doi: 10.52202/085713-4150. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/b4689a8e6ac04c0919f7162365f1248d-Paper-Conference.pdf.

Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, pp. 638–647, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392037. doi: 10.1145/3503161.3547910. URL https://doi.org/10.1145/3503161.3547910.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pp. 529–544, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-19808-3. doi: 10.1007/978-3-031-19809-0_30. URL https://doi.org/10.1007/978-3-031-19809-0_30.

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

Kranti Parida, Omar Emara, Hazel Doughty, and Dima Damen. Segmenting collision sound sources in egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026.

Sneha Paul, Zachary Patterson, and Nizar Bouguila. Point cloud as a foreign language for multi-modal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Learning Representations (ICLR)*, 2021. doi: 10.48550/ARXIV.2103.00020. URL https://arxiv.org/abs/2103.00020.

Jiahua Rao, Hanjing Lin, Leyu Chen, Jiancong Xie, Shuangjia Zheng, and Yuedong Yang. Multimodal contrastive learning with negative sampling calibration for phenotypic drug discovery. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30752–30762, 2025. doi: 10.1109/CVPR52734.2025.02864.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL https://aclanthology.org/P18-1238/.

Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once–multi-modal fusion transformer for video retrieval. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19988–19997. IEEE, 2022.

Rakshith Sharma Srinivasa, Jaejin Cho, Chouchang Yang, Yashas Malur Saidutta, Ching-Hua Lee, Yilin Shen, and Hongxia Jin. Cwcl: Cross-modal transfer with continuously weighted contrastive loss. *Advances in Neural Information Processing Systems*, 36:78496–78513, 2023.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision (ECCV)*, 2019. doi: 10.48550/ARXIV.1906.05849. URL https://arxiv.org/abs/1906.05849.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling foundation models for multimodal video understanding. pp. 396–416, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73012-2. doi: 10.1007/978-3-031-73013-9_23. URL https://doi.org/10.1007/978-3-031-73013-9_23.

Liangjian Wen, Qun Dai, Jianzhuang Liu, Jiangtao Zheng, Yong Dai, Dongkai Wang, Zhao Kang, Jun Wang, Zenglin Xu, and Jiang Duan. Infmasking: Unleashing synergistic information by contrastive multimodal interactions. *Advances in Neural Information Processing Systems*, 38: 15529–15555, 2026.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5288–5296, 2016. doi: 10.1109/CVPR.2016.571.

Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6995–7004, June 2021.

Yiling Wu Junshu Sun Yaowei Wang Shuhui Wang Yue Wu, Zhaobo Qi. Learning fine-grained representations through textual token disentanglement in composed video retrieval. *ICLR*, 2025.

Jiawei Zhou, Xiaoguang Li, Lifeng Shang, Xin Jiang, Qun Liu, and Lei Chen. Retrieval-based disentangled representation learning with natural language supervision. In *International Conference on Learning Representations (ICLR)*, 2024. doi: 10.48550/ARXIV.2212.07699. URL https://arxiv.org/abs/2212.07699.

A Appendix

A.1 AI Statement

AI was used to polish some drafted sections and figures in the paper