Representation Learning for Semantic Alignment of Language, Audio, and Visual Modalities

Parthasaarathy Sudarsanam, Irene Martín-Morató, Tuomas Virtanen

Audio Research Group, Tampere University, Tampere, Finland

{parthasaarathy.ariyakulamsudarsanam, irene.martinmorato, tuomas.virtanen}@tuni.fi

Abstract—This paper proposes a single-stage training approach that semantically aligns three modalities - audio, visual, and text using a contrastive learning framework. Contrastive training has gained prominence for multimodal alignment, utilizing large-scale unlabeled data to learn shared representations. Existing deep learning approach for trimodal alignment involves two-stages, that separately align visual-text and audio-text modalities. This approach suffers from mismatched data distributions, resulting in suboptimal alignment. Leveraging the AVCaps dataset, which provides audio, visual and audio-visual captions for video clips, our method jointly optimizes the representation of all the modalities using contrastive training. Our results demonstrate that the single-stage approach outperforms the two-stage method, achieving a two-fold improvement in audio based visual retrieval, highlighting the advantages of unified multimodal representation learning.

Index Terms—Multimodal representation, AVCaps, Audio-visual

I. INTRODUCTION

The increasing prevalence of multimodal content and large language models demands the development of models capable of jointly processing and understanding audio, visual (video frames without sound), and textual modalities. Contrastive learning techniques [1, 2] have proven effective in bridging modality gaps. They achieve better semantic alignment across modalities compared to earlier deep learning approaches. For example, multimodal deep Boltzmann machines [3] relied on generative modeling, [4] demonstrated that CNNs trained to predict words from image captions could learn useful multimodal representations. Contrastive learning frameworks explicitly optimize for multimodal similarity achieving superior alignment.

Multimodal representation learning has seen significant advancements in aligning two modalities. Models like ALIGN [5], CLIP [1], Florence [6] achieve impressive results by aligning image-text pairs; and CLAP [2] performs effectively in aligning audio-text pairs, using the contrastive learning frameworks. However, extending this success to trimodal alignment remains a challenge. Two-stage approaches, such as wav2CLIP [7], first align visual and textual modalities in a shared embedding space and then introduce audio alignment in a second stage. While effective for certain tasks, this sequential process often leads to suboptimal audio-visual alignment, as the model is biased towards the initial visual-text relationship. Moreover, differing dataset distributions between the two stages amplify the alignment issues, making it difficult to
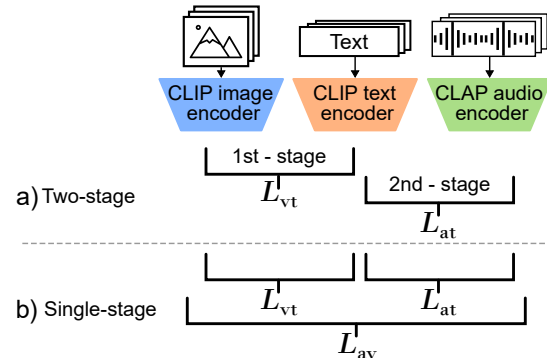

Fig. 1: Schematic representation of the two-stage reference method (a) and the proposed single-stage method (b). Our proposed method aligns the representations from the three modalities by jointly optimizing the visual-text (L_{vt}), audio-text (L_{at}), and audio-visual (L_{av}) contrastive losses.

fully capture the intricate interactions among audio, visual, and textual modalities.

Trimodal representation learning is further hampered by the lack of datasets with modality-specific textual annotations for audio and visual data for the same video clip. Existing datasets, focus on visual-text [8–10] or audio-text [11–14] pairs, leaving a gap in capturing the complex relationships required for unified trimodal models. AudioCLIP [15] attempts to bridge this gap by extending CLIP with an audio branch and aligning all three modalities using sound event classes from AudioSet [16] as textual descriptions. However, these annotations are primarily sound-centric and lack visual context. These limitations highlight the need for unified frameworks and datasets that enable simultaneous learning across all three modalities. Recently, we released the AVCaps dataset [17] that contains modality-specific captions for audio, visual, and audiovisual modalities associated with video clips.

This paper presents SLAVA, a Single-stage Language Audio Visual Alignment model contrastively trained to semantically align the three modalities effectively. Unlike previous methods that use either a two-stage approach [7] or a single-stage model with only audio captions [15], we leverage the AVCaps dataset to simultaneously utilize both audio and visual captions. This allows us to align all three modalities within a single-stage framework. Our results demonstrate that SLAVA achieves superior trimodal alignment, outperforming existing methods.

II. RELATED WORKS

A. Contrastive Language-Image Pretraining (CLIP)

The CLIP model [1] is a powerful contrastive learning framework designed to align textual and visual modalities. Trained on large-scale of 400M image-text pairs, it enables the model to learn meaningful correspondences between natural language descriptions and images. However, its reliance on object-centric captions in the training data limits its capacity to capture event-level or scene-level relationships. Additionally, since the text encoder was trained solely with captions derived for static images, it has not been exposed to audio-related language. This restricts CLIP's ability to generalize to tasks involving audio-visual contexts or to represent relationships beyond the visual modality.

B. Contrastive Language-Audio Pretraining (CLAP)

CLAP [2] uses contrastive learning to align textual and audio modalities. Training on audio-text pairs, CLAP enables associations between natural language descriptions and diverse audio inputs, such as environmental sounds or speech. CLAP is trained with a relatively smaller dataset of 128K audio clips and their associated textual description sourced from various audio captioning datasets. Similar to the limitations of CLIP, The CLAP text encoder has not been exposed to vision specific language during training, constraining its ability to represent visual relationships.

C. Wav2CLIP

Wav2CLIP [7] uses a two-stage approach for audio-visual-textual alignment. It involves training a visual-text semantic alignment model in the first stage. In the second stage, one of the branches is either frozen or fine-tuned along with an audio-text or audio-visual dataset to achieve trimodal alignment. A common issue with this method arises from using datasets from different sources for the two stages, resulting in data distribution mismatches that lead to suboptimal alignment.

D. AudioCLIP

In AudioCLIP [15], the authors proposed a single stage model by utilizing AudioSet [16], a dataset that includes video clips associated with sound events, as a foundation for its trimodal training. AudioCLIP extends CLIP by integrating an audio encoder to the CLIP framework to create a unified model. While this trimodal approach represents significant progress, it has certain limitations. The textual descriptions used for aligning the three modalities are derived from Audioset sound-event classes, resulting in annotations that are primarily focused on sound-event-specific information. This focus limits the model's ability to generalize to tasks requiring balanced, scene-level, or context-rich representations across audio-visual modalities.

III. METHODS

A. Dataset

We use our recently released AVCaps [17] dataset. The dataset comprises 2,061 videos with a total duration of 28.8 hours. Each video is annotated with up to five crowdsourced captions for audio, visual, and audio-visual modalities. In addition, each video includes three audio-visual captions generated using GPT-4, leveraging the modality-specific captions. This uniqueness of AVCaps dataset enables a comprehensive study of training strategies for aligning audio, visual, and textual modalities.

B. Overview of the methods

In all our experiments, we used pretrained CLIP image and text encoders and CLAP audio encoder and finetuned their projection layers to obtain latent representations of visual, textual, and audio inputs. The CLIP image encoder processes an input video of M frames individually to produce a representation of size $\mathbb{R}^{M \times 768}$, while the text encoder generates a $\mathbb{R}^{512}$ representation for a given textual input. The CLAP audio encoder takes N chunks of 10-second audio inputs and returns a representation of size $\mathbb{R}^{N \times 768}$.

The encoded representations are passed through their respective projection layers. The CLIP image and CLAP audio projection layers, each a linear layer with 512 neurons, convert $\mathbb{R}^{M \times 768}$ and $\mathbb{R}^{N \times 768}$ to $\mathbb{R}^{M \times 512}$ and $\mathbb{R}^{N \times 512}$, respectively. The CLIP text projection layer is also a linear layer with 512 neurons and hence keeps the size of the text representation unchanged. To make the representations of the three modalities the same size, we compute the average of the audio and visual features along the time axis, resulting in a final representation of $\mathbb{R}^{512}$. The projection layers, which learn the joint multi-modal representation are finetuned using the InfoNCE [18] contrastive loss.

Table I provides an overview of the reference methods alongside our proposed approaches, highlighting the trainable layers, the AVCaps data partitions used, and the loss functions employed. These methods are further detailed in the following subsections.

C. Reference Methods

*1) **Wav2CLIP-style two-stage model**:* In this approach, we first align the visual and textual representations, followed by the alignment of audio and textual representations in the next stage. We use the AVCaps dataset, which provides audio and visual captions for the same videos, to address data distribution mismatches in conventional two-stage models. We begin by fine-tuning the projection layers of a pre-trained CLIP model, using the visual content from the videos and their crowdsourced visual captions. For the second stage, we finetune the audio projection layers of a pre-trained CLAP model alongside the textual projection layers of the CLIP text encoder from stage one. Alternatively, we experimented with keeping the text projection layers frozen, only fine-tuning the audio projection layers in the second stage. This approach is similar to the training paradigm used in Wav2CLIP. In this approach, we align the audio and visual embeddings indirectly through shared textual representations, which may result in suboptimal performance for audio-visual tasks.

Methods	Stages	Trainable Layers	Training data	Objective Function
Wav2CLIP-style	Stage 1 Stage 2	Visual and text projection Audio and text projection	Visuals and visual captions Audio and audio captions	L_{vt} L_{at}
AudioCLIP-style	Single stage	Audio, visual, and text projection	Audios, visuals, audio captions	$L_{av} + L_{vt} + L_{at}$
SLAVA$_{A\&V}$ (ours)	Single stage	Audio, visual, and text projection	Audios, visuals, audio captions, visual captions	$L_{av} + (L_{vt} \text{ or } L_{at})$
SLAVA$_{AV}$ (ours)	Single stage	Audio, visual, and text projection	Audios, visuals, audio-visual captions	$L_{av} + L_{vt} + L_{at}$

TABLE I: Overview of the reference and proposed models, training stages, trainable layers, datasets, and loss functions. L_{av}, L_{at}, and L_{vt} represent the audio-visual, audio-text, and visual-text contrastive losses respectively.

2) AudioCLIP-style single-stage model: As an additional reference, we developed a single-stage model inspired by the AudioCLIP framework. In this model, we fine-tuned the pre-trained CLIP projection layers and CLAP audio projection layers using the visuals, audios, and the audio captions from the AVCaps dataset using pairwise contrastive losses for all three modalities. This approach mirrors the AudioCLIP training process, which utilized the AudioSet, as their captions predominantly focus on sound events. By aligning the three modalities directly in a single-stage framework, this method provides an additional benchmark for evaluating the performance of our proposed trimodal alignment approach. The single-stage approach addresses the indirect audio-visual alignment limitation of the two-stage approach. However, its effectiveness is limited by the fine-tuning dataset, where the textual descriptions cover only a single modality, in this case audio.

D. Proposed methods

1) SLAVA with audio and visual captions: Our proposed SLAVA model uses both audio and visual captions for single-stage training to enhance multimodal alignment. In this setup, the model benefits from both auditory and visual information present in the textual captions, enhancing the alignment between all three modalities. Figure 1b shows our proposed SLAVA model. For each minibatch, we choose corresponding audios, visuals, and randomly either audio captions or visual captions from the AVCaps dataset. When audio captions are used, the loss includes audio-text and audio-visual contrastive terms. Conversely, when visual captions are used, the loss comprises visual-text and audio-visual contrastive terms. We refer to this model as SLAVA$_{A\&V}$ denoting that it was trained with audio and visual captions. The total loss L_{total} in this case is given by

$$L_{total} = L_{av} + (L_{at} \text{ or } L_{vt}) \qquad (1)$$

where, L_{av}, L_{at}, and L_{vt} represent the audio-visual, audio-text, and visual-text contrastive losses respectively. Our approach addresses the limitations of our reference methods by 1) enabling explicit alignment of both audio and visual modalities, and 2) ensuring that the textual descriptions comprehensively cover both audio and visual modalities.

2) SLAVA with audio-visual captions: In another set of experiments, we used the audios, visuals and the LLM-generated audio-visual captions. We refer to this model as SLAVA$_{AV}$ denoting that it was trained on audio-visual captions. For these experiments, we designed two training configurations. In the first configuration, training was performed using two contrastive losses L_{at} and L_{vt}. This approach allows us to compare how well the captions containing both audio and visual information could indirectly improve audio-visual alignment without explicitly applying L_{av}. In this configuration, the loss function is given by

$$L_{total} = L_{at} + L_{vt} \qquad (2)$$

In the second configuration, we included the L_{av}, thereby explicitly aligning all three modalities. The loss function is given by

$$L_{total} = L_{av} + L_{at} + L_{vt} \qquad (3)$$

IV. EVALUATION

To evaluate our reference and proposed methods, we performed a series of retrieval tasks designed to assess bimodal and trimodal alignment. All our models were finetuned on the AVCaps training split for 20 epochs, using the AdamW optimizer with a learning rate of $1e-5$ and a weight decay of 0.1 using the InfoNCE loss [18]. The best model is selected based on its performance on the validation split of AVCaps.

We evaluated all our models on the test split of AVCaps on several retrieval tasks with the widely used recall@10 metric. In a retrieval task, a textual caption or an audio signal serves as the input query, which is compared against all items in the database to retrieve the top-10 most similar results. The recall@10 metric calculates the proportion of relevant items among the top-10 retrieved results by the total of items in the dataset.

The bimodal retrieval tasks evaluate the model's ability to retrieve one modality using queries from another, such as retrieving audio with audio captions or visuals with visual captions. For crossmodal retrieval, we investigate tasks like audio captions based visual retrieval and visual captions based audio retrieval, which require the model to align representations across different modalities. We also evaluate audio-visual captions based audio and visual retrieval. Additionally,

Retrieve	Based On	Wav2CLIP-style (2 stage)		AudioCLIP-style	SLAVA$_{\text{A\&V}}$	SLAVA$_{\text{AV}}$ (ours)	
		Frozen Text	Trainable Text	(single stage)	(ours)	$L_{\text{at}} + L_{\text{vt}}$	$L_{\text{av}} + L_{\text{at}} + L_{\text{vt}}$
Visual	Visual Captions	0.84	0.82	0.69	0.81	0.83	0.81
Visual	Audio Captions	0.26	0.29	0.29	0.30	0.29	0.29
Visual	Audio-Visual Captions	0.86	0.86	0.75	0.85	0.89	0.86
Audio	Audio Captions	0.35	0.40	0.36	0.37	0.33	0.34
Audio	Visual Captions	0.27	0.29	0.26	0.32	0.32	0.33
Audio	Audio-Visual Captions	0.41	0.41	0.37	0.40	0.44	0.45
Visual	Audio	0.22	0.27	0.45	0.52	0.40	0.50

TABLE II: **Recall@10** retrieval performance across various models and configurations on the AVCaps test split.

to compare our models in terms of audio-visual similarities, we analyze the performance of audio based visual retrieval to assess how the models align and transfer knowledge between these two modalities. These tasks provide a comprehensive evaluation of the model's ability in creating a shared multi-modal representation for all three modalities involved.

V. RESULTS

Table II shows the retrieval performance of all our models on the AVCaps test split evaluated using the recall@10 metric. Wav2CLIP-style results show the retrieval performance of our reference two-stage model in two configurations: frozen and trainable text projection layers in stage 2. In the first approach, the text encoder is trained only with visual captions and CLIP training data, which mainly contain vision-specific language. When audio captions are introduced in stage 2, the frozen text encoder produces suboptimal representations, constraining the audio representation to align with these suboptimal text embeddings. This results in a weaker audio-visual alignment. In contrast, updating the text projection layer in stage 2 enables the audio encoder to learn better representations of the input audio, resulting in better alignment. Finetuning text projection layers slightly lowers visual retrieval with visual captions but improves audio retrieval and other crossmodal tasks.

AudioCLIP-style results present the performance of our reference single-stage model, which aligns all three modalities using audio captions. It improves the audio-visual alignment, with audio-based visual retrieval increasing to 0.45 compared to 0.27 of the two-stage approach. However, the visual retrieval scores drop significantly in comparison to the two-stage approach. This is because the captions contain audio-specific information which shifts the distribution of the text encoder away from its original visual-centric pretraining, reducing its ability to capture visual features.

SLAVA$_{\text{A\&V}}$ results summarize the performance of our proposed single-stage model trained on audio and visual captions. It uses audio-visual contrastive loss combined with either audio-text or visual-text contrastive loss, based on the textual input in the mini-batch. By directly incorporating audio-visual contrastive loss, the single-stage approach improves audio-based visual retrieval, increasing recall@10 to 0.52 from 0.27, while maintaining performance across other retrieval

tasks. These results clearly establish that single-stage trimodal alignment using both audio and visual captions enabled by AVCaps dataset create a robust multimodal representation compared to the traditional two-stage approach.

Finally, we present the results of our SLAVA$_{\text{AV}}$ model, a single-stage model trained with audio-visual captions. The first approach, using only audio-text and visual-text contrastive losses, is designed for comparison with the two-stage model lacking explicit audio-visual alignment. Despite this, our proposed model achieves a recall@10 of 0.40 for audio-based visual retrieval, compared to 0.27 for the two-stage model. In the second approach, the model incorporates an additional audio-visual contrastive loss alongside audio-text and visual-text losses, enabling explicit alignment across the modalities. This model outperforms the reference single stage model on almost all the retrieval tasks across modalities.

We also trained an audio-visual alignment model using the CLIP image encoder and the CLAP audio encoder, with audio and visuals from the AVCaps training split. This model achieved a score of 0.50 in audio-based visual retrieval on the test split. Our proposed SLAVA models match this performance while also effectively aligning the textual modality with both the audio and visual modalities.

VI. CONCLUSION

In this paper, we proposed SLAVA, a single-stage approach for multimodal alignment that jointly optimizes audio, visual, and textual representations using contrastive training. Our method addresses the limitations of two-stage approaches, where separate training of visual-text and audio-text models often leads to suboptimal alignment due to mismatched data distributions. Using modality-specific captions in the AVCaps dataset, we demonstrated that our approach significantly improves alignment across modalities. Evaluation results on a variety of retrieval tasks, show that our single-stage method outperforms the traditional two-stage approach, achieving a recall@10 score of 0.52 in audio-based visual retrieval compared to 0.27 of the reference model. These findings highlight the advantages of AVCaps dataset and a unified framework for multimodal representation learning, demonstrating its potential for advancing tasks such as audio-visual captioning, video understanding, and multimodal question answering.

REFERENCES

[1] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 8748–8763.

[2] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP learning audio concepts from natural language supervision," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[3] N. Srivastava and R Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25, Curran Associates, Inc., 2012.

[4] A. Joulin, L. Van Der Maaten, A. Jabri, and N. Vasilache, "Learning visual features from large weakly supervised data," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, Springer, 2016, pp. 67–84.

[5] C. Jia *et al.*, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning*, PMLR, 2021, pp. 4904–4916.

[6] L. Yuan *et al.*, "Florence: A new foundation model for computer vision," *arXiv preprint arXiv:2111.11432*, 2021.

[7] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, "Wav2CLIP: Learning robust audio representations from CLIP," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 4563–4567.

[8] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, Proceedings, Part V 13*, Springer, 2014, pp. 740–755.

[9] R. Krishna *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, pp. 32–73, 2017.

[10] B. Thomee *et al.*, "YFCC100M: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[11] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.

[12] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 736–740.

[13] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 2019, pp. 119–132.

[14] I. Martín-Morató and A. Mesaros, "What is the ground truth? reliability of multi-annotator data for audio tagging," in *29th European Signal Processing Conference (EUSIPCO)*, IEEE, 2021, pp. 76–80.

[15] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "AudioCLIP: Extending CLIP to image, text and audio," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 976–980.

[16] J. F. Gemmeke *et al.*, "Audio Set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017.

[17] P. Sudarsanam, I. Martín-Morató, A. Hakala, and T. Virtanen, "Avcaps: An audio-visual dataset with modality-specific captions," *IEEE Open Journal of Signal Processing*, pp. 1–15, 2025. DOI: 10.1109/OJSP.2025.3578296.

[18] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.