Quality Over Quantity? LLM-Based Curation for a Data-Efficient Audio–Video Foundation Model

Ali Vosoughi*
University of Rochester
Rochester, NY, USA

Dimitra Emmanouilidou
Microsoft Research
Redmond, WA, USA

Hannes Gamper
Microsoft Research
Redmond, WA, USA

***Abstract*—Integrating audio and visual data for training multimodal foundational models remains challenging. We present Audio-Video Vector Alignment (AVVA), which aligns audiovisual (AV) scene content beyond mere temporal synchronization via a Large Language Model (LLM)-based *data curation* pipeline. Specifically, AVVA scores and selects high-quality training clips using Whisper (speech-based audio foundation model) for audio and DINOv2 for video within a dual-encoder contrastive learning framework. Evaluations on AudioCaps, VALOR, and VGGSound demonstrate that this approach can achieve significant accuracy gains with substantially less curated data. For instance, AVVA yields a 7.6% improvement in top-1 accuracy for audio-to-video retrieval on VGGSound compared to ImageBind, despite training on only 192 hours of carefully filtered data (vs. $5800+$ hours). Moreover, an ablation study highlights that *trading data quantity for data quality* improves performance, yielding respective top-3 accuracy increases of 47.8, 48.4, and 58.0 percentage points on AudioCaps, VALOR, and VGGSound over uncurated baselines. While these results underscore AVVA's data efficiency, we also discuss the overhead of LLM-driven curation and how it may be scaled or approximated in larger domains. Overall, AVVA provides a viable path toward more robust, text-free audiovisual learning with improved retrieval accuracy.**

***Index Terms*—Audio-Video Vector Alignment (AVVA), Multimodal Learning, Audio-Visual Retrieval, Scene Understanding**

I. Introduction and Motivation

Humans seamlessly merge auditory and visual information without relying on explicit textual mediation. When watching a video, we integrate visual cues and corresponding sounds to form a cohesive perception of the scene [1], [2]. In contrast, most multimodal AI systems—such as CLIP [3], CLAP [4], and related models [5]–[11]—rely on text captions to align audio and visual information [12]–[21].This text-centric alignment does not fully reflect human-like sensory integration, where no textual intermediary is needed [1], [2], [22]–[24].

Recent approaches, e.g., Wav2CLIP [7], AudioCLIP [8], and ImageBind [9], also aim to fuse audio and vision, yet most still incorporate text-based or audio-image pairs for alignment, possibly limiting generalization [22], [28]. To address this gap, we propose **AVVA (Audio-Video Vector Alignment)**, a framework designed for concurrent, *text-free* audiovisual learning during the *model-training* phase. In AVVA, we first perform *LLM-based data curation* to score and filter raw video clips, then train a cross-attentive embedding model using Whisper [29] for audio and DINOv2 [30] for vision. Our motivation is twofold:

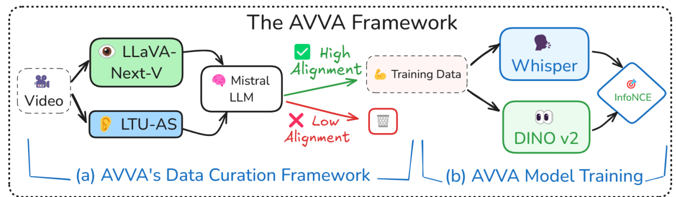


Fig. 1: **Overview of AVVA.** (a) A data-curation stage uses multimodal reasoning to filter out misaligned samples, leveraging LLaVA-Next-Video [25], LTU-AS [26], and Mistral [27]. (b) Our model then employs Whisper (audio) and DINOv2 (video) backbones to learn a *joint* audiovisual embedding *without text* during training.

- **New audiovisual tasks, e.g. video LLMs:** Modern applications often require a *merged* representation that captures temporal-spatial cues from both audio and video, rather than separate embeddings [12], [31].
- **Data Efficiency:** By filtering out low-quality pairs, we show (Sec. III) that training on fewer but higher-quality samples can still achieve strong performance on retrieval and classification benchmarks (AudioCaps [32], VALOR [33], VGGSound [34]), comparing favorably to large-scale baselines.

Our contributions are threefold:

1) **Joint multimodal embedding:** AVVA produces a *single* fused representation for audio and video, in contrast to prior methods using isolated encoders (like CLAP + CLIP).
2) **LLM-driven data curation:** We introduce a pipeline that uses text, audio, and video LLMs (e.g., LLaVA-Next-Video [25], [35], LTU-AS [26]) to automatically score alignment in the raw data, removing misleading or mismatched clips.
3) **Competitive results with less data:** Despite using fewer training samples (curated sets), AVVA attains results on par with or better than state-of-the-art approaches on standard audiovisual retrieval tasks (Sec. III).

Notably, while we rely on LLM prompts for *data filtering*, AVVA's *model-training* process does not require text captions.

* Work completed during an internship at Microsoft Research, Redmond, WA, USA.

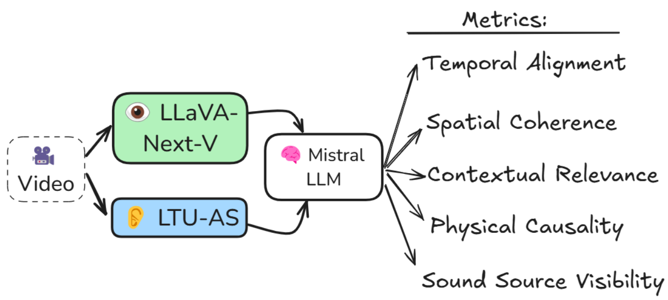


Fig. 2: The architecture of the MRE. The design integrates outputs of an audio-LLM and a video-LLM into a Mistral LLM to reason over audiovisual scene alignment by integrating 5 alignment scores that were calculated on the AV pairs.

This distinction enables a *text-free* alignment objective in the final stage, reflecting more human-like audiovisual fusion. The rest of the paper is structured as follows: Sec. II describes our data curation and model design, Sec. III reports experimental findings, and Sec. IV discusses limitations and possible extensions.

II. AVVA: Audio-Video Vector Alignment

A key aim of **AVVA** is to produce a *unified audio-video embedding*, rather than relying on separate encoders (e.g., CLAP for audio, CLIP for video) whose semantic spaces often diverge in tasks such as video LLMs. By integrating both modalities into one joint representation, AVVA can better support downstream applications that require *merged* audio-video context (Sec. III), such as generation [36], event detection [37], and other tasks requiring detailed audiovisual analysis [38]. In the following, we describe our data curation pipeline and training approach.

A. Multimodal Reasoning Engine (MRE)

High-quality, aligned audiovisual data is critical for learning robust embeddings [39], [40]. We therefore introduce a **Multimodal Reasoning Engine (MRE)** to curate training pairs. MRE prompts three LLMs:

1) An audio-LLM (LTU-AS [26], LLaMA 2-based) for generating audio captions,
2) A video-LLM (LLaVa-NeXT-Video [25], [35], LLaMA 3-based) for video captions,
3) A Mistral 7B Instruct model for final scoring.

We use **five metrics**—Temporal Alignment, Spatial Coherence, Contextual Relevance, Physical Causality, and Sound Source Visibility—to evaluate each audio-video pair, see Fig. 2. Each metric is scored on 0–10. An overall average alignment score is computed, and pairs below a threshold (e.g., 6.2 or 7.6) are *excluded*.

a) Datasets and Processing.: We sample diverse datasets (Epic-Kitchens, HowTo100M, VGGSound, *etc.*) to cover various domains: egocentric *vs.* exocentric, music *vs.* ambient audio, speech, and more (see Table I for statistics). Videos are cut into 3 s clips; each clip's audio/video is fed to MRE for captioning and scoring. The final curated set typically retains 70–90% of the original data.

TABLE I: Summary of Dataset Statistics Used in Analysis

Dataset	Set	Segments	Hours	% of Total
AudioCaps [32]	Train	39,167	32.64	2.56%
AudioSet [41]	Train	65,799	54.83	4.30%
AVE [23]	Train	12,001	10.00	0.78%
HD-VILA-100M [42]	Train	754,374	628.65	49.34%
VGGSound [34]	Train	36,280	30.23	2.37%
Epic-Kitchens [43]	Train	1,638	1.37	0.11%
HowTo100M [44]	Train	9,326	7.77	0.61%
Music-MIT [45]	Train	2,572	2.14	0.17%
VALOR [33]	Train	570,823	475.69	37.34%
Total (Train)		**1,034,639**	**1,243.32**	**97.53%**
AudioCaps	Val	478	0.40	0.03%
VALOR	Val	16,364	13.64	1.07%
Total (Val)		**16,842**	**14.04**	**1.10%**
AudioCaps	Test	1,199	1.00	0.08%
VALOR	Test	16,300	13.58	1.07%
VGGSound	Test	3,384	2.82	0.22%
Total (Test)		**20,883**	**17.40**	**1.36%**
Grand Total		**1,072,364**	**1,274.76**	**100%**

b) Limitations.: Running MRE is not trivial: each clip calls multiple LLMs, which may be expensive at scale. In practice, we focus on moderate-sized data. Future work could replace MRE with approximate approaches or conduct partial captioning for large-scale pipelines.

B. Model Architecture and Language-free Training

a) Why a Joint Embedding?: Unlike methods that use separate encoders (CLAP audio + CLIP video, e.g., [31], [46]), our approach explicitly merges both signals. This is beneficial for *emerging tasks* such as video LLMs [35] where audio cues (speech, music) and visual features must be processed *together*, enabling more coherent temporal/spatial understanding than separate embeddings [12], [31], [47].

b) Encoders and Cross-Attention.: Figure 3 illustrates our approach, where we employ **Whisper** (the 32-layer model [29]) for audio processing, utilizing the features of all layers, except the first layer, which undergoes a average pooling [26], [48]. For video processing, we employ **DINOv2** [30], which has demonstrated superior ability in capturing local features [22]. Each produces a feature sequence. We feed these into *aligner layers* (MLPs: $(\mathrm{dim}, 1024, 768)$ with ReLU, dropout), then a **bidirectional cross-modal attention** module with 8 heads, dimension 768. This yields a *single* fused representation for each modality, capturing cross-modal context.

c) InfoNCE Training.: We adopt an InfoNCE loss [3], temperature 0.07, to align audio-video pairs. Freezing Whisper/DINOv2 [49] avoids overfitting and reduces GPU memory usage. Only the cross-attention and aligner parameters are trained (AdamW, lr=10^{-4}). Our experiments (§III) show that combining *MRE-curated data* with *joint audio-video embedding* yields strong performance, even with fewer training hours compared to large-scale baselines.

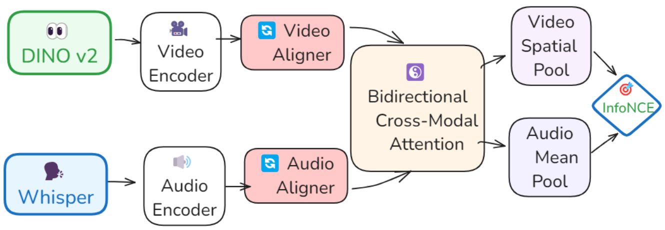


Fig. 3: **AVVA Architecture.** Whisper (audio) and DINOv2 (video) are frozen encoders. Outputs are passed to aligner MLPs, then a bidirectional cross-attention block. The final fused embeddings are used for contrastive learning.

III. Experiments

We conducted three sets of experiments to evaluate the performance of our method in diverse scenarios, focusing on both small-scale retrieval feasibility and data curation effects.

A. Cross-modal Retrieval

In this experiment, we assess the ability of **AVVA** to retrieve audio from video and vice versa across three datasets: AudioCaps [32], VALOR [33], and VGG-Sound [34]. We compare our model against Wav2CLIP [7], DenseAV [22], a Random baseline, and ImageBind [9]. Each test is performed on 3-second video segments containing embedded audio and is repeated five times, sampling 100 video files per iteration. While some files may be repeated across different iterations, no duplication occurs within the same set of 100 samples. Results are reported as statistical averages.

Why Cross-attention at Test? Our model leverages a bidirectional cross-attention mechanism to enrich audio-video embeddings for fine-grained alignment, even during inference. Although this improves accuracy, it can be computationally expensive in large-scale deployments. For that reason, we focus on a small-scale setting of 100 samples per iteration, which keeps inference overhead manageable. In real-world scenarios with potentially millions of items, a two-stage retrieval pipeline (coarse candidate selection followed by cross-attentive re-ranking) could be employed to mitigate computational cost.

Key Novelty. Unlike typical two-encoder methods, AVVA uses a *joint* audio-video architecture specifically tailored to exploit alignment cues *without* needing text. We also introduce a *data curation* process (detailed in Sec. III-B) that further boosts accuracy by filtering out misaligned pairs. Together, these elements yield robust performance with fewer training hours.

Results and Data Efficiency. AVVA achieves accuracy comparable to ImageBind despite using only 192 hrs of carefully curated data (versus ImageBind's 5,800+ hrs), indicating a 30$\times$ improvement in data efficiency. AVVA also surpasses DenseAV, underscoring the effectiveness of training on high-quality, curated audiovisual pairs. Notably, all other methods in Table II rely on larger datasets (e.g., Wav2CLIP uses 278 hrs). This comparison reveals how robust curation can enhance model performance. The results for AVVA in Table II reflect a 7.6 out of 10 score, chosen by selecting the minimal loss checkpoint.

A key observation from our experiments is that increasing the amount of training data does not always yield higher accuracy. While more data often boosts performance initially, it may also introduce noise if misaligned or irrelevant samples are included. This phenomenon is evident in our runs, where models trained on larger but unfiltered data (Wav2CLIP, DenseAV) underperformed relative to AVVA, despite having access to more total hours.

B. Data Curation Impact on Performance

This experiment evaluates how our curation process influences cross-modal retrieval. As shown in Fig. 4, higher curation thresholds lead to improved performance, particularly with limited training data. Curation removes noisy pairs, allowing the model to learn better alignment. Though it raises preprocessing time overhead (about six additional seconds per segment), the quality gains translate directly to better retrieval, as seen in Table III. The ablation study reveals that Table III and §III-B precisely demonstrate the significant increases in top-k accuracies for AudioCaps, VALOR, and VGGSound. AVVA achieves substantial Top-1 improvements under the same training duration for both audio-to-video and video-to-audio tasks on AudioCaps, VALOR, and VGG-Sound.

TABLE III: Performance increases (%) with data curation at Top-1,3,10 across various datasets.

Dataset	**Audio→Video**			**Video→Audio**		
	Top1	**Top3**	**Top10**	**Top1**	**Top3**	**Top10**
AudioCaps	37.4	47.8	42.8	30.0	38.2	35.4
VALOR	41.2	48.4	35.6	34.2	44.2	39.2
VGGSound	49.6	58.0	43.8	40.4	47.8	36.2

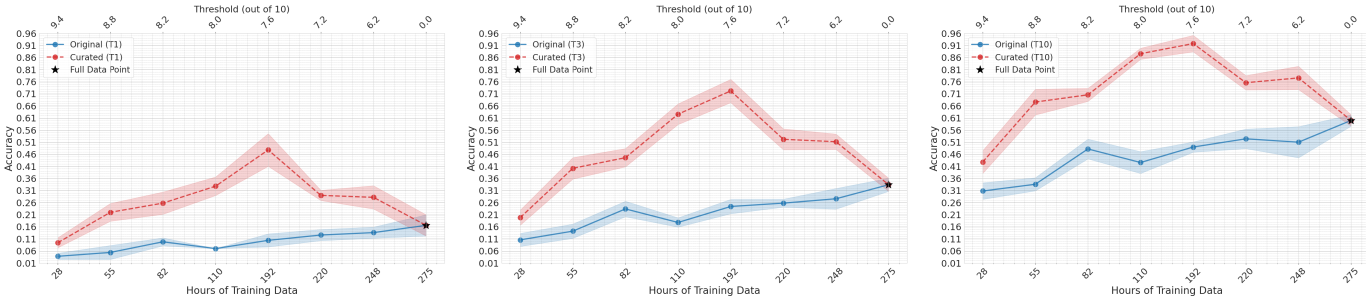
C. Temporal Alignment

Finally, to examine audio-video timing, we shift audio segments from -3.0 to +3.0 s relative to their paired videos and measure cosine similarity. Figure 5 shows a clear peak near a zero-second offset, suggesting our cross-attentive embeddings capture tight temporal correlations. Notably, events with sharp onsets (e.g., gunshots) exhibit higher sensitivity to timing shifts than slow-moving scenes (e.g., distant train noises).

Limitation and Future Work. While we show strong results in small-scale tests (100 items per run), fully cross-attentive retrieval can be computationally costly at very large scales. A dual-encoder or two-stage re-ranking architecture may be preferable in those scenarios. In future work, we plan to explore hybrid solutions that preserve our *data efficiency* and alignment quality while scaling to broader retrieval contexts.

IV. Conclusion

We have presented **AVVA**, a framework for joint audio-video alignment that avoids text captions during *model training*. Instead, our *Multimodal Reasoning Engine (MRE)* uses LLM prompts to filter out low-scoring clips, enabling AVVA to learn robust audiovisual representations from significantly less data than prior methods like ImageBind.



(a) Top-1 Accuracy (b) Top-3 Accuracy (c) Top-10 Accuracy

Fig. 4: Model performance over hours of training data, with and without data curation, shown for Top-1, Top-3, and Top-10 accuracies. Solid (blue) lines depict model performance when training on original data, while dashed (red) lines on curated data. Performance is consistently higher when training on curated data. The top x-axis shows the MRE score threshold. While these results are for audio-to-video retrieval on the AudioCaps dataset, similar patterns were observed for other datasets and video-to-audio tasks, where curation enhances performance without overfitting.

TABLE II: **Performance Comparison on Audio-Video Retrieval** (Top-$k = \{1, 3, 10\}$) across three datasets (%). Standard deviations are over five runs. **Bold** indicates the highest accuracy; underlined marks the second-highest.

Method	Retrieval Type	AudioCaps			VALOR			VGG-Sound		
		Top1	Top3	Top10	Top1	Top3	Top10	Top1	Top3	Top10
Wav2CLIP [7]	A→V	$1.20^{\pm0.98}$	$6.40^{\pm3.01}$	$18.60^{\pm3.83}$	$3.60^{\pm0.80}$	$8.60^{\pm0.49}$	$18.20^{\pm3.54}$	$3.40^{\pm1.36}$	$8.20^{\pm1.17}$	$19.80^{\pm3.06}$
	V→A	$3.80^{\pm2.14}$	$10.00^{\pm3.22}$	$20.00^{\pm3.63}$	$4.20^{\pm2.64}$	$8.00^{\pm4.24}$	$19.00^{\pm3.52}$	$3.80^{\pm1.94}$	$9.20^{\pm2.32}$	$19.60^{\pm1.62}$
Random	A→V	$1.40^{\pm0.49}$	$3.80^{\pm0.75}$	$11.80^{\pm1.17}$	$1.20^{\pm0.75}$	$3.20^{\pm0.40}$	$11.60^{\pm1.62}$	$1.20^{\pm0.40}$	$3.40^{\pm0.49}$	$11.60^{\pm2.15}$
	V→A	$1.00^{\pm0.00}$	$3.60^{\pm0.80}$	$10.80^{\pm1.17}$	$1.20^{\pm0.40}$	$3.20^{\pm0.75}$	$11.00^{\pm0.63}$	$1.00^{\pm0.00}$	$3.00^{\pm0.00}$	$10.60^{\pm0.80}$
DenseAV [22]	A→V	$10.20^{\pm2.04}$	$22.60^{\pm4.54}$	$49.40^{\pm4.54}$	$7.80^{\pm5.19}$	$19.00^{\pm5.90}$	$41.80^{\pm4.79}$	$6.80^{\pm2.64}$	$16.00^{\pm2.90}$	$43.20^{\pm3.43}$
	V→A	$1.40^{\pm0.80}$	$5.60^{\pm1.85}$	$26.40^{\pm2.73}$	$2.20^{\pm1.17}$	$5.80^{\pm2.79}$	$24.60^{\pm7.68}$	$1.60^{\pm1.02}$	$5.00^{\pm0.63}$	$22.60^{\pm2.58}$
ImageBind [9]	A→V	$\mathbf{62.00^{\pm2.28}}$	$\mathbf{83.40^{\pm3.01}}$	$92.60^{\pm1.85}$	$\mathbf{55.80^{\pm4.66}}$	$71.60^{\pm3.61}$	$85.00^{\pm3.74}$	$50.60^{\pm3.14}$	$74.00^{\pm5.93}$	$88.20^{\pm2.99}$
	V→A	$\mathbf{64.00^{\pm5.37}}$	$\mathbf{85.40^{\pm4.27}}$	$\mathbf{95.40^{\pm0.80}}$	$\mathbf{58.80^{\pm4.71}}$	$\mathbf{73.60^{\pm4.36}}$	$86.60^{\pm3.20}$	$53.20^{\pm3.31}$	$73.40^{\pm6.02}$	$85.60^{\pm3.20}$
AVVA (Ours)	A→V	$47.80^{\pm6.05}$	$72.20^{\pm4.31}$	$\mathbf{91.80^{\pm3.06}}$	$52.60^{\pm4.32}$	$\mathbf{74.80^{\pm3.66}}$	$\mathbf{92.20^{\pm1.47}}$	$\mathbf{58.20^{\pm4.79}}$	$\mathbf{81.40^{\pm3.56}}$	$\mathbf{96.00^{\pm1.79}}$
	V→A	$40.40^{\pm4.63}$	$60.20^{\pm4.53}$	$85.80^{\pm3.19}$	$46.80^{\pm5.19}$	$71.80^{\pm4.49}$	$\mathbf{92.00^{\pm3.03}}$	$51.80^{\pm5.38}$	$\mathbf{74.40^{\pm5.54}}$	$\mathbf{91.60^{\pm3.44}}$

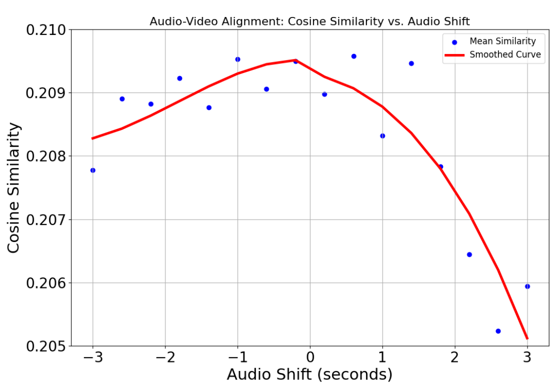


Fig. 5: Cosine similarity as a function of audio shift. The data points show mean similarity scores at each shift level, with the red line representing a smoothed trend. This plot illustrates the sensitivity of our method to audio-video alignment shifts on the AudioCaps dataset, highlighting the temporal precision in audiovisual synchronization.

a) Key Results.: Despite training on only ∼192 hours of curated samples, AVVA achieves comparable or superior retrieval accuracy to large-scale baselines (5,800+ hrs). In particular, AVVA attains 58.20% Top-1 on VGG-Sound (vs. 50.60% for ImageBind) for audio-to-video retrieval and 92.00% Top-10 on VALOR (vs. 86.60%) for video-to-audio retrieval. These findings highlight the importance of *data quality* over brute-force quantity in advancing multimodal AI.

b) Limitations and Future Work.: While our LLM-based curation greatly reduces the need for large labeled datasets, running multiple models for scoring can be computationally demanding. In large-scale deployments, approximate or staged filtering might be needed. Additionally, we keep Whisper and DINOv2 encoders frozen, which can limit ultimate performance if end-to-end fine-tuning is feasible. Future directions include extending AVVA embeddings to *video LLM* applications, exploring partial finetuning of the encoders, and broadening domain coverage beyond our current datasets.

By centering on *joint* audio-video embeddings and high-quality data selection, AVVA bridges a gap in text-free audiovisual learning while preserving strong performance. We hope these results encourage further research on *LLM-guided data filtering* and *data efficiency* in multimodal foundation models.

REFERENCES

[1] David Harwath, Antonio Torralba, and James Glass, "Unsupervised Learning of Spoken Language with Visual Context," *Adv. Neural Inform. Process. Syst.*, Vol. 29, 2016.

[2] Yapeng Tian, Dingzeyu Li, and Chenliang Xu, "Unified Multisensory Perception: Weakly-supervised Audio-Visual Video Parsing," *Eur. Con. Comput. Vis.* Springer, pp. 436–454, 2020.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, *et al.*, "Learning transferable visual models from natural language supervision," *Int. Con. Machine Learning*, 2021.

[4] Benjamin Elizalde, Soham Deshmukh, *et al.*, "CLAP Learning Audio Concepts from Natural Language Supervision," *Int. Con. Acoustics, Speech, and Sig. Process.* IEEE, pp. 1–5, 2023.

[5] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," *Int. Con. Machine Learning*. PMLR, pp. 19730–19742, 2023.

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, *et al.*, "Uniter: Universal Image-Text Representation Learning," *Eur. Con. Comput. Vis.* Springer, pp. 104–120, 2020.

[7] Ho-Hsiang Wu, Prem Seetharaman, *et al.*, "Wav2CLIP: Learning Robust Audio Representations from CLIP," *Int. Con. Acoustics, Speech, and Sig. Process.* IEEE, pp. 4563–4567, 2022.

[8] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel, "AudioCLIP: Extending CLIP to Image, Text and Audio," *Int. Con. Acoustics, Speech, and Sig. Process.* IEEE, pp. 976–980, 2022.

[9] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, *et al.*, "ImageBind: One Embedding Space to Bind Them All," *IEEE Con. Comput. Vis. Pattern Recog.*, pp. 15180–15190, 2023.

[10] Ali Vosoughi, Luca Bondi, Ho-Hsiang Wu, and Chenliang Xu, "Learning Audio Concepts from Counterfactual Natural Language," *Int. Con. Acoustics, Speech, and Sig. Process.* IEEE, pp. 366–370, 2024.

[11] Zhe Chen, Hongcheng Liu, and Yu Wang, "DialogMCF: Multimodal Context Flow for Audio Visual Scene-Aware Dialog," *ACM Trans. Aud., Spe., Lang. Process.*, Vol. 32, pp. 753–764, 2023.

[12] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, *et al.*, "Video Understanding with Large Language Models: A Survey," *arXiv preprint arXiv:2312.17432*, 2024.

[13] Jing Bi, Yunlong Tang, Luchuan Song, Ali Vosoughi, *et al.*, "EAGLE: Egocentric AGgregated Language-video Engine," *ACM Int. Con. Multimedia*, 2024.

[14] Nguyen Nguyen, Jing Bi, Ali Vosoughi, *et al.*, "OSCaR: Object State Captioning and State Change Representation," *Proc. Annual Con. North American Chapter Assoc. for Comput. Linguistics*, pp. 3565–3576, 2024.

[15] Kun Su, Xiulong Liu, and Eli Shlizerman, "From Vision to Audio and Beyond: A Unified Model for Audio-Visual Representation and Generation," *Int. Con. Machine Learning*, 2024.

[16] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, *et al.*, "SALMONN: Towards Generic Hearing Abilities for Large Language Models," *Int. Con. Learn. Represent.*, 2024.

[17] Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, *et al.*, "Audio Flamingo: A Novel Audio Language Model with Few-Shot Learning and Dialogue Abilities," *Int. Con. Machine Learning*, 2024.

[18] Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang, "Pengi: An Audio Language Model for Audio Tasks," *Adv. Neural Inform. Process. Syst.*, Vol. 36, pp. 18090–18108, 2023.

[19] Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, *et al.*, "Listen, Think, and Understand," *Int. Con. Learn. Represent.*, 2024.

[20] Bin Huang, Xin Wang, Hong Chen, Zihan Song, *et al.*, "VTimeLLM: Empower LLM to Grasp Video Moments," *IEEE Con. Comput. Vis. Pattern Recog.*, pp. 14271–14280, 2024.

[21] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, *et al.*, "LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment," *Int. Con. Learn. Represent.*, 2024.

[22] Mark Hamilton, Andrew Zisserman, John R Hershey, and William T Freeman, "Separating the" Chirp" from the" Chat": Self-supervised Visual Grounding of Sound and Language," *IEEE Con. Comput. Vis. Pattern Recog.*, pp. 13117–13127, 2024.

[23] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu, "Audio-Visual Event Localization in Unconstrained Videos," *Eur. Con. Comput. Vis.*, pp. 247–263, 2018.

[24] David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, *et al.*, "Jointly discovering visual objects and spoken words from raw sensory input," *Eur. Con. Comput. Vis.*, pp. 649–665, 2018.

[25] Yuanhan Zhang, Bo Li, Haotian Liu, *et al.*, "LLaVA-NeXT: A Strong Zero-shot Video Understanding Model," https://llava-vl.github.io/blog/2024-04-30-llava-next-video/, April 2024, A: March 13, 2025.

[26] Yuan Gong, Alexander H Liu, Hongyin Luo, Leonid Karlinsky, *et al.*, "Joint Audio and Speech Understanding," *W. Automatic Speech Recog. and Understanding*. IEEE, pp. 1–8, 2023.

[27] Mistral AI, "Mistral Inference Model," https://github.com/mistralai/mistral-inference, 2023, A: March 13, 2025.

[28] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed, "Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction," *Int. Con. Learn. Represent.*, 2022.

[29] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, *et al.*, "Robust Speech Recognition via Large-Scale Weak Supervision," *Int. Con. Machine Learning*. PMLR, pp. 28492–28518, 2023.

[30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, *et al.*, "DINOv2: Learning Robust Visual Features without Supervision," *Trans. on Machine Learning Research*, 2024.

[31] Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Jun Chen, *et al.*, "Meerkat: Audio-visual large language model for grounding in space and time," *Eur. Con. Comput. Vis.* Springer, pp. 52–70, 2024.

[32] Chris Dongjoo Kim, Byeongchang Kim, *et al.*, "Audiocaps: Generating Captions for Audios in the Wild," *Proc. Annual Con. North American Chapter Assoc. for Comput. Linguistics*, pp. 119–132, 2019.

[33] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, *et al.*, "VALOR: Vision-Audio-Language Omni-Perception Pretraining Model and Dataset," *arXiv preprint arXiv:2304.08345*, 2023.

[34] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman, "VGGSound: A Large-Scale Audio-Visual Dataset," *Int. Con. Acoustics, Speech, and Sig. Process.* IEEE, pp. 721–725, 2020.

[35] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, *et al.*, "LLaVA-NeXT: Improved Reasoning, OCR, and World Knowledge," https://llava-vl.github.io/blog/2024-01-30-llava-next/, January 2024, A: March 13, 2025.

[36] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, *et al.*, "Seeing and Hearing: Open-Domain Visual-Audio Generation with Diffusion Latent Aligners," *IEEE Con. Comput. Vis. Pattern Recog.*, pp. 7151–7161, 2024.

[37] Davide Berghi, Peipei Wu, Jinzheng Zhao, *et al.*, "Fusion of Audio and Visual Embeddings for Sound Event Localization and Detection," *Int. Con. Acoustics, Speech, and Sig. Process.* IEEE, pp. 8816–8820, 2024.

[38] Yiyang Su, Ali Vosoughi, Shijian Deng, Yapeng Tian, *et al.*, "Separating Invisible Sounds Toward Universal Audiovisual Scene-Aware Sound Separation," *Int. Con. Comput. Vis.*, 2023.

[39] Ben Sorscher, Robert Geirhos, Shashank Shekhar, *et al.*, "Beyond Neural Scaling Laws: Beating Power Law Scaling via Data Pruning," *Adv. Neural Inform. Process. Syst.*, Vol. 35, pp. 19523–19536, 2022.

[40] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, *et al.*, "Textbooks Are All You Need," 2023.

[41] Google Research, "AudioSet," 2017, Available online: https://research.google.com/audioset/download.html [A: March 13, 2025].

[42] Hongwei Xue, Tiankai Hang, Yanhong Zeng, *et al.*, "Advancing High-Resolution Video-Language Representation with Large-Scale Video Transcriptions," *IEEE Con. Comput. Vis. Pattern Recog.*, 2022.

[43] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, *et al.*, "The Epic-Kitchens Dataset: Collection, Challenges and Baselines," *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 43, No. 11, pp. 4125–4141, 2020.

[44] Antoine Miech, Ivan Laptev, and Josef Sivic, "HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips," *Int. Con. Comput. Vis.*, 2019.

[45] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba, "The Sound of Motions," *Int. Con. Comput. Vis.*, pp. 1735–1744, 2019.

[46] Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, *et al.*, "video-SALMONN: speech-enhanced audio-visual large language models," *Int. Con. Machine Learning*, pp. 47198–47217, 2024.

[47] Yunlong Tang, Daiki Shimada, Jing Bi, and Chenliang Xu, "Avicuna: Audio-visual llm with interleaver and context-boundary alignment for temporal referential dialogue," *arXiv e-prints*, pp. arXiv–2403, 2024.

[48] Yuan Gong, Sameer Khurana, *et al.*, "Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers," *Proc. Interspeech 2023*, pp. 2798–2802, 2023.

[49] Guangzhi Wang, Yixiao Ge, Xiaohan Ding, Mohan Kankanhalli, *et al.*, "What Makes for Good Visual Tokenizers for Large Language Models?," *arXiv preprint arXiv:2305.12223*, 2023.