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 a challenge. The Audio-Video Vector Alignment (AVVA) framework addresses this by considering AV scene alignment beyond mere temporal synchronization, and leveraging Large Language Models (LLMs) for data curation. AVVA implements a scoring mechanism for selecting aligned training data segments. It integrates Whisper, a speech-based foundation model, for audio and DINOv2 for video analysis in a dual-encoder structure with contrastive learning on AV pairs. Evaluations on AudioCaps, VALOR, and VGGSound demonstrate the effectiveness of the proposed model architecture and data curation approach. AVVA achieves a significant improvement in top-k accuracies for video-to-audio retrieval on all datasets compared to DenseAV, while using only 192 hrs of curated training data. Furthermore, an ablation study indicates that the data curation process effectively trades data quality for data quantity, yielding increases in top-k retrieval accuracies on AudioCaps, VALOR, and VGGSound, compared to training on the full spectrum of uncurated data.**

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

I. Introduction and Motivation

Humans naturally process audiovisual information without any need for textual mediation. For instance, when watching a video, we instinctively merge visual cues with corresponding sounds to create a cohesive understanding of the scene. However, most current multimodal AI systems, like CLIP [1] and CLAP [2], and majority of other models [3]–[9], depend on textual captions to connect visual and auditory features. This reliance on text-based alignment is at odds with how humans integrate sensory information, where no explicit textual representation is required.

Replicating this human-like processing in AI is challenging [14]–[16]. Existing multimodal models primarily handle individual modalities separately, later merging them based on text-based associations [17]–[26]. This approach, evident in models like Wav2CLIP [5], AudioCLIP [6] and ImageBind [7], misses the opportunity to exploit the natural synchronization between audio and visual data. While efforts like AV-HuBERT [27] and DenseAV [28] aim to capture linguistic information along with the location of sounds from raw audiovisual pairs, they still rely on speech-image pairs for training, which may restrict their generalization.

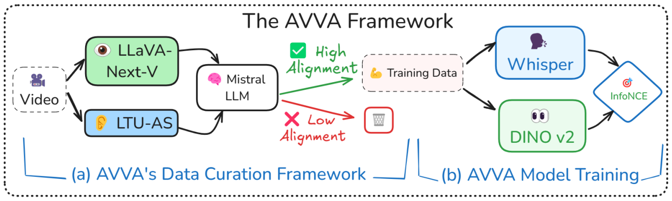

Fig. 1: Overview of the proposed audiovisual alignment approach. (a) Our method's data curation stage uses multimodal reasoning to retain only highly aligned data. It uses LLaVA-Next-Video [10], [11] for video reasoning, LTU-AS [12] for audio processing, and Mistral [13] for alignment scoring. (b) AVVA employs Whisper (audio), DINOv2 (video backbone), without the need for textual mediation during training.

To address this gap, we introduce **AVVA: Audio-Video Vector Alignment**, a framework designed to directly align AV information without any text dependency. The proposed model leverages *Whisper* [29] for audio processing and *DINOv2* [30] for visual understanding. This makes AVVA particularly effective in applications requiring concurrent, text-free audiovisual comprehension, such as video analysis and human-computer interaction. An important data curation stage takes place first, which itself relies on a text-, audio- and video-LLMs.

Our contributions are threefold: (1) AVVA is the first audio-visual foundation model that incorporates a speech foundation model to enable generalized audiovisual representation learning. (2) Unlike previous approaches that often align audio and visual features independently, AVVA introduces a mechanism for joint multimodal reasoning. (3) Our novel data curation mechanism significantly reduces the amount of required training data while still achieving competitive results with state-of-the-art models, which further demonstrates the efficiency and effectiveness of using curated data over the original datasets.

The remainder of the paper is organized as follows: Section II explains our methodology, including data and model design. Section III presents the experimental results, and Section IV concludes the paper.

II. AVVA: Audio-Video Vector Alignment

A key feature of the proposed method is the curation and selection of high-quality paired data. AVVA leverages the synergy

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

of three large models—two for multimodal inputs (audio and video) and one for joint reasoning—to compute five alignment scores. These scores will then be used to evaluate the coherence of the audiovisual data. We will explain different parts of the method. More details on the implementation and reproducibility of AVVA, including prompts, statistics of datasets, and ablations studies will be provided at https://github.com/AVVA-curation.

A. Multimodal Reasoning Engine (MRE)

The potential of most AI techniques for LLMs and multimodal learning often hinges on the diversity and quality of the data they interact with [31], [32]. In this work, we curate the training data via the introduction of a Multimodal Reasoning Engine (MRE), which is a set of prompts for obtaining detailed reasoning of audio, video, and finally to score the level of alignment between audio and video from their textual descriptions, given a set of five metrics.

We used multiple audiovisual datasets that cover diverse scenes from both egocentric and exocentric perspectives, and various forms of audio, including natural, music, ambient, and speech, and other complex scenarios. The datasets are: Epic-Kitchens (1.37 hrs) [33], HowTo100M (7.77 hrs) [34], Music-MIT (2.14 hrs) [35], VALOR (train/test 94.57/13.58 hrs) [36], VGGSound (train/test 30.23/2.82 hrs) [37], AVE (10.00 hrs) [38], AudioSet (54.83 hrs) [39], AudioCaps (train/test 32.64/1.00 hrs) [40], HD-VILA-100M (51.45 hrs) [41]. All input videos were segmented into 3-sec clips; for longer videos, up to 20 random clips were kept. To achieve joint audio-speech reasoning, each segment was processed using LLaVa-NeXT-Video with LLaMA 3 [10], [11] for video reasoning and LTU-AS [12] with LLaMA 2 for audio reasoning. We used Mistral 7B Instruct v0.3 for prompting and measuring alignment.

Given a video sample, we obtain one caption from LTU-AS describing the audio, and one caption from LLaVA-Next-V describing the video, see Fig. 2. The two captions are then fed to Mistral, along with a prompt request to obtain five separate scores in a scale of [0,10] that aim to capture caption alignment. These scores come from the five metrics: Temporal Alignment, Spatial Coherence, Contextual Relevance, Physical Causality, and Sound Source Visibility. For Temporal Alignment, we ask the system to assess how well the events described in the audio caption match the timing of events in the video caption (e.g., a clap sound syncing with hands meeting). Spatial Coherence evaluates how well the audio description aligns with the spatial layout and objects described in the video (e.g., a car's engine sound moving from left to right as it passes). Contextual Relevance refers to how closely the subject matter and theme of the audio align with those of the video (e.g., kitchen sounds matching cooking activities). Physical Causality assesses the extent to which the described sounds can be logically attributed to the objects, actions, or events depicted in the video (e.g., glass breaking sound matching the visual shattering). Sound Source Visibility considers that some visual objects may produce sound without being visible and others may be visible but silent. The prompt details can be found in the Appendix (see GitHub). These alignment scores are then averaged with equal weights, and a final alignment score is produced, with the assumption that a higher score represents better audiovisual alignment. A scoring threshold is subsequently applied for final data curation. For reference, retaining 90% or 70% of the original training data corresponded to a score threshold of 6.2 and 7.6 respectively.

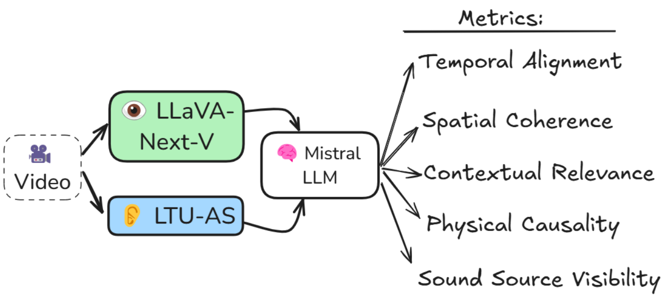

Fig. 2: The architecture of the MRE. The design integrates outputs of an audio-LLM and a video-LLM into a Mistral LLM to reason over audiovisual scene alignment by integrating 5 alignment scores that were calculated on the AV pairs.

B. Model Architecture and Language-free Training

The AVVA model employs a bidirectional cross-modal attention mechanism to integrate audio and video modalities using dual encoders—Whisper for audio and DINOv2 for video. We selected DINOv2 [30] over models like CLIP [1] due to its ability to capture local visual features, which are crucial for producing high-quality global representations through feature pooling [28]. For the audio encoder, we utilize Whisper, concatenating 32 layers while discarding the first layer, as applied in [12]. The architecture is illustrated in Fig. 3.

The bidirectional attention mechanism, implemented with 8 attention heads and a 768-dimensional hidden state, ensures a robust flow of information between the audio and video streams, treating both modalities as equally important. This design makes AVVA particularly effective for complex multimodal tasks requiring detailed audiovisual understanding, such as synchronized multimedia content generation [42], event detection [43], and other tasks requiring detailed audiovisual analysis [44]. By aligning the features through learnable aligner layers - implemented as MLPs with dimensions (input_dim, 1024, 768) and layer normalization, $ReLU$, and dropout 0.2 between layers - and pooling the outputs, the model generates compact embeddings suitable for contrastive learning.

We use the InfoNCE loss function [1] with a temperature setting of 0.07 to help the model learn correlations between audio and video features. For optimization, we adopt the AdamW optimizer with a learning rate and weight decay set to 10^{-4}. To maintain computational efficiency and prevent catastrophic forgetting, both the DINOv2 and Whisper encoders are frozen [45] during training, focusing on training only the alignment and cross-modal layers.

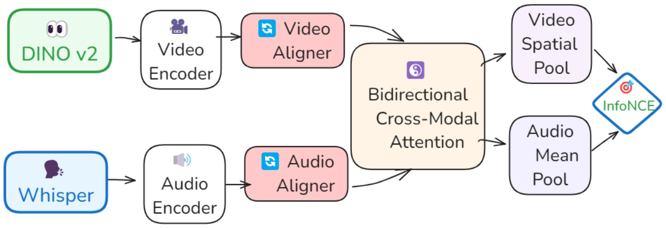

Fig. 3: The AVVA model training. Audio (Whisper) and video (DINOv2) encoders process raw inputs, which are aligned via learnable parameters in aligner layers. A Bidirectional Cross-Modal Attention helps capture the interaction between audio and video features, which are pooled to generate final embeddings for contrastive learning.

III. Experiments

We conducted three sets of experiments to evaluate the performance of our method in diverse scenarios.

A. *Cross-modal Retrieval*

In this experiment, we assess the ability of AVVA to retrieve audio from video input and vice versa, across three datasets: AudioCaps [40], VALOR [36], and VGG-Sound [37]. Each test is performed on 3-second video segments containing embedded audio, and we compare our model against Wav2CLIP [5], DenseAV [28], Random weights, and ImageBind [7]. Each retrieval test is evaluated K=100 times on N=100 random audio/video files per iteration. No duplication occurs within each set of 100 samples per run. Results are reported as statistical averages.

AVVA achieves audio-to-video accuracy comparable to DenseAV, with significantly improved video-to-audio accuracy, despite using only 192 hrs of carefully curated data compared to DenseAV's 5,800 hrs, demonstrating a 30x improvement in data efficiency. This showcases the effectiveness of high-quality, curated audiovisual pairs curated by our system. Notably, all competing methods in Table I were trained on larger datasets. For instance, Wav2CLIP was trained on approximately 278 hrs of data. This comparison highlights the impact of effective data curation on enhancing model performance. The results for AVVA in Table I reflect an MRE threshold score of 7.6 out of 10, based on selecting the epoch with the minimum validation los. A key observation from our experiments is that increasing the amount of training data does not always lead to better performance. While more data should generally improve model accuracy initially, adding data can introduce noise, particularly when the additional data is less curated or includes irrelevant or misaligned audiovisual pairs. This phenomenon is portrayed in our experiments, where models trained on large but uncurated datasets such as Wav2CLIP and DenseAV performed equivalent or worse than AVVA, especially in V2A retrieval, despite having access to more data.

B. *Data Curation Impact on Performance*

This experiment evaluates the effect of data curation on model performance in cross-modal retrieval tasks. As illustrated in Fig. 4, higher curation thresholds lead to improved performance. We argue that meaningful data curation reduces noise in training data, allowing the model to focus on high-quality examples, resulting in more accurate retrieval across modalities. Similar plots were obtained for the other test sets and the video-to-audio retrieval task. Despite being computationally expensive - typically increasing preprocessing time by 6 seconds per GPU time per segment - the improved data quality curation enhances the model's ability to generalize, which showcases the importance of the data selection in multimodal training. The findings are summarized in Table II, in terms of performance improvement (%) as compared to training on full data . AVVA achieves top-1 performance increases relative to original dataset with same hrs of training across datasets for both audio-to-video and video-to-audio tasks. For audio-to-video retrieval, AVVA achieves increases of 18.0, 16.24, and 13.57 percentage points (%) in top 1, 3, and 10 for AudioCaps; for VALOR, increases of 22.67, 23.97, and 15.42 % respectively; and for VGGSound, increases of 23.25, 15.79, and 10.44 % in top 1, 3, and 10. The proposed method also shows merit in the V2A task, in this case more moderate improvements than for the A2V task, shown in the second column of the Table.

TABLE II: Performance increases in cross-modal retrieval tasks with data curation. Top-$k = \{1, 3, 10\}$ % increase across various datasets as compared to training on full original data.

Dataset	**Audio-to-Video ↑ (%)**			**Video-to-Audio ↑ (%)**		
	Top 1	**Top 3**	**Top 10**	**Top 1**	**Top 3**	**Top 10**
AudioCaps	18.0	16.24	13.57	11.08	11.29	14.71
VALOR	22.67	23.97	15.42	10.44	4.00	8.50
VGGSound	23.25	15.79	10.44	1.76	3.41	1.86

C. *Temporal Alignment*

To investigate the audio-video temporal alignment, we conducted simulations where audio segments were systematically shifted relative to their corresponding video segments across a defined range of shifts from -3.0 to 3.0 sec, in increments of 0.4 sec. For each shift, cosine similarity between audio and video embeddings was computed to assess the alignment quality. The multimodal audiovisual embeddings were extracted using our pre-trained model.

Figure 5 shows cosine similarity scores for the video and audio embeddings of 200 samples across audio shifts. Mean similarity values at each shift are represented by markers, with a smoothed trend line fitted using a Savitzky-Golay filter to highlight the underlying pattern while preserving key variations. The analysis reveals a clear peak in similarity at around 0 sec shift between audio and video, providing evidence of meaningful audio-video learning.

It is important to consider the nature of the data when interpreting these results. Events like hammering or gunshots, which involve sharp and temporally precise correlations between sound and image, exhibit a different behavior compared to more continuous or slower-changing video scenes. For example, considering a video of a train moving in the distance, subtle audio-video delays may be less perceptually disruptive but still affect cosine similarity scores. In such cases, lower cosine

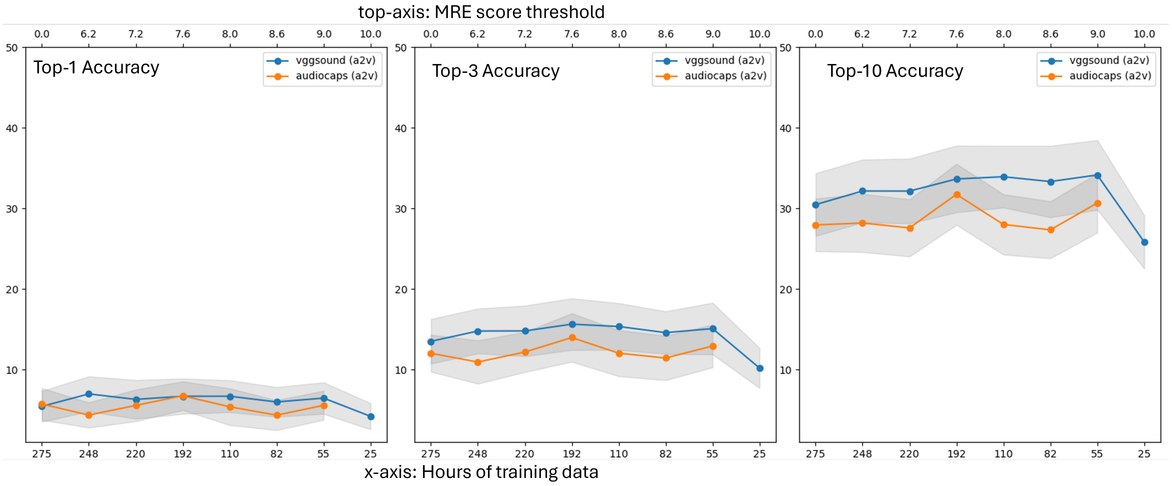

Fig. 4: Audio-to-video model performance over hours of training data, as determined by varying the selection of the MRE score threshold, shown for Top-$k = \{1, 3, 10\}$ accuracies.

TABLE I: **Performance on Audio (A) - Video (V) Retrieval** (Top-$k = \{1, 3, 10\}$) (%). Standard deviations shown as superscripts depict performance variation over K=100 retrieval repetitions of random test subsets of size N=100.

Method	**Retrieval Type**	**AudioCaps**			**VALOR**			**VGG-Sound**		
		Top1	**Top3**	**Top10**	**Top1**	**Top3**	**Top10**	**Top1**	**Top3**	**Top10**
Wav2CLIP [5]	A→V	$1.20^{\pm 0.98}$	$6.40^{\pm 3.01}$	$18.60^{\pm 3.83}$	$3.60^{\pm 0.80}$	$8.60^{\pm 0.49}$	$18.20^{\pm 3.54}$	$3.40^{\pm 1.36}$	$8.20^{\pm 1.17}$	$19.80^{\pm 3.06}$
	V→A	$3.80^{\pm 2.14}$	$10.00^{\pm 3.22}$	$20.00^{\pm 3.63}$	$4.20^{\pm 2.64}$	$8.00^{\pm 4.24}$	$19.00^{\pm 3.52}$	$3.80^{\pm 1.94}$	$9.20^{\pm 2.32}$	$19.60^{\pm 1.62}$
Random	A→V	$1.40^{\pm 0.49}$	$3.80^{\pm 0.75}$	$11.80^{\pm 1.17}$	$1.20^{\pm 0.75}$	$3.20^{\pm 0.40}$	$11.60^{\pm 1.62}$	$1.20^{\pm 0.40}$	$3.40^{\pm 0.49}$	$11.60^{\pm 2.15}$
	V→A	$1.00^{\pm 0.00}$	$3.60^{\pm 0.80}$	$10.80^{\pm 1.17}$	$1.20^{\pm 0.40}$	$3.20^{\pm 0.75}$	$11.00^{\pm 0.63}$	$1.00^{\pm 0.00}$	$3.00^{\pm 0.00}$	$10.60^{\pm 0.80}$
DenseAV [28]	A→V	$10.20^{\pm 2.04}$	$22.60^{\pm 4.54}$	$49.40^{\pm 4.54}$	$7.80^{\pm 5.19}$	$19.00^{\pm 5.90}$	$41.80^{\pm 4.79}$	$6.80^{\pm 2.64}$	$16.00^{\pm 2.90}$	$43.20^{\pm 3.43}$
	V→A	$1.40^{\pm 0.80}$	$5.60^{\pm 1.85}$	$26.40^{\pm 2.73}$	$2.20^{\pm 1.17}$	$5.80^{\pm 2.79}$	$24.60^{\pm 7.68}$	$1.60^{\pm 1.02}$	$5.00^{\pm 0.63}$	$22.60^{\pm 2.58}$
ImageBind [7]	A→V	$62.00^{\pm 2.28}$	$83.40^{\pm 3.01}$	$92.60^{\pm 1.85}$	$55.80^{\pm 4.66}$	$71.60^{\pm 3.61}$	$85.00^{\pm 3.74}$	$50.60^{\pm 3.14}$	$74.00^{\pm 5.93}$	$88.20^{\pm 2.99}$
	V→A	$64.00^{\pm 5.37}$	$85.40^{\pm 4.27}$	$95.40^{\pm 0.80}$	$58.80^{\pm 4.71}$	$73.60^{\pm 4.36}$	$86.60^{\pm 3.20}$	$53.20^{\pm 3.31}$	$73.40^{\pm 6.02}$	$85.60^{\pm 3.20}$
AVVA (Ours)	A→V	$6.57^{\pm 2.30}$	$13.84^{\pm 2.80}$	$31.68^{\pm 3.57}$	$6.69^{\pm 2.13}$	$15.63^{\pm 3.52}$	$33.67^{\pm 4.40}$	$6.71^{\pm 1.91}$	$15.02^{\pm 2.73}$	$33.86^{\pm 4.23}$
	V→A	$6.23^{\pm 2.09}$	$14.70^{\pm 3.17}$	$31.06^{\pm 3.52}$	$7.75^{\pm 2.61}$	$16.64^{\pm 3.65}$	$34.27^{\pm 4.71}$	$6.86^{\pm 2.34}$	$14.47^{\pm 3.15}$	$32.84^{\pm 3.89}$

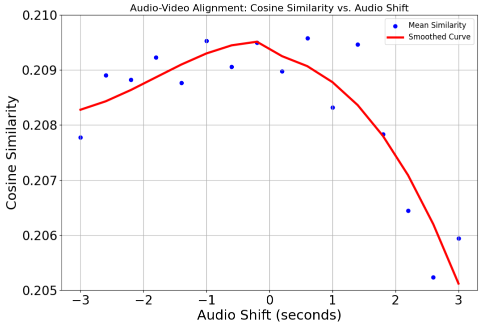

Fig. 5: Cosine similarity between AVVA embeddings (video versus shifted audio) as a function of audio shift. The data points show mean similarity scores at each shift level.

similarity may not necessarily imply poor alignment but rather may reflect the characteristics of the content, emphasizing the critical role of data context in assessing audiovisual alignment.

IV. Conclusion

AVVA addresses the challenges of joint multimodal learning by directly processing and curating multi-faceted aligned AV data without linguistic mediation in model training. Our approach, utilizing a speech foundation model backbone, demonstrates significant improvements in AV retrieval tasks. The LLM-based MRE module for data curation rejects audiovisual pairs of low-scoring alignment and helps the model achieve comparable performance to state-of-the-art methods with substantially less training data. AVVA matches DenseAV's performance using only $\sim$ 192 hrs of curated data, compared to DenseAV's 5800+ hrs – a 30x gain in data utilization. Experiments across multiple datasets showcase the merit of AVVA's methodology on reducing data utilization while maintaining or improved performance. While more work is needed to render the data curation process less computationally expensive, including more efficient reasoning engines, the five metrics comprising the proposed MRE score show a lot of promise, and the overall results highlight the importance of data quality in advancing multimodal AI models.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, *et al.*, "Learning Transferable Visual Models from Natural Language Supervision," PMLR, pp. 8748–8763, 2021.

[2] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang, "CLAP Learning Audio Concepts from Natural Language Supervision," *Int. Con. Acoustics, Speech, and Sig. Process.* IEEE, pp. 1–5, 2023.

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

[4] 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.

[5] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello, "Wav2CLIP: Learning Robust Audio Representations from CLIP," *Int. Con. Acoustics, Speech, and Sig. Process.* IEEE, pp. 4563–4567, 2022.

[6] 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.

[7] 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," *IEEE Con. Comput. Vis. Pattern Recog.*, pp. 15180–15190, 2023.

[8] 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.

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

[10] 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, Accessed: June 21, 2025.

[11] Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, *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, Accessed: June 21, 2025.

[12] 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.

[13] Mistral AI, "Mistral Inference Model," https://github.com/mistralai/mistral-inference, 2023, Accessed: June 21, 2025.

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

[15] 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.

[16] 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.

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

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

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

[20] Kun Su, Xiulong Liu, and Eli Shlizerman, "From Vision to Audio and Beyond: A Unified Model for Audio-Visual Representation and Generation," 2024.

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

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

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, *et al.*, "Robust Speech Recognition via Large-Scale Weak Supervision," 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] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, *et al.*, "Beyond Neural Scaling Laws: Beating Power Law Scaling via Data Pruning," *Adv. Neural Inform. Process. Syst.*, Vol. 35, pp. 19523–19536, 2022.

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

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

[34] 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.

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

[36] 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.

[37] 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.

[38] 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.

[39] Google Research, "AudioSet," 2017, Available online: https://research.google.com/audioset/download.html [Accessed: June 21, 2025].

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

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

[42] 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.

[43] Davide Berghi, Peipei Wu, Jinzheng Zhao, Wenwu Wang, *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.

[44] 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.

[45] 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.