Journal of Technological Information, Management & Engineering Sciences (2023),4(1), 53-62

Khan and Malik

Unified Multimodal Representation Learning: Cross-Modal Alignment of Video, Audio, and Text for Robust Artificial Intelligence

[1] **Dr. Ahmed Khan**
[2] **Sara Malik**

How to cite this article:

Khan, A., & Malik, S. (2023). Unified Multimodal Representation Learning: Cross-Modal Alignment of Video, Audio, and Text for Robust Artificial Intelligence. *Journal of Technological Information, Management & Engineering Sciences, 4(1), 53-62.*

Received	Accepted	Published
12 January 2023	18 March 2023	20 June 2023

Abstract

The explosive growth of multimodal data, which includes text, audio and video, has motivated the creation of unified representation learning approaches that can align different modalities in a common embedding space. This alignment is crucial for the development of strong AI systems that need to engage in reasoning across modalities, zero-shot transfer learning, and robustness to missing or noisy modalities. This paper summarizes recent progress in cross-modal alignment of video, audio and text, covering contrastive learning paradigms, hub-based binding strategies and large multimodal models. The analysis is based on peer-reviewed papers from 2023 to 2025, such as ImageBind, AudioCLIP, Video-LLaMA, UniAlign, CAV-MAE refinements, and in-depth studies of multimodal fusion.The analysis draws from peer-reviewed papers that span the years from 2023 to 2025, including ImageBind, AudioCLIP, Video-LLaMA, UniAlign, CAV-MAE refinements, and comprehensive surveys of multimodal fusion. The results show that image-centric hub binding and well-designed contrastive losses lead to good emergent alignment between different, previously unpaired modalities, while unified encoder architectures and reconstruction-based objectives further enhance the coherence between understanding and generation. Some challenges that remain are temporal synchronisation in long videos, modality imbalance, computation complexity and robustness to real-world noise. Finally, there are implications for building successful and robust scalable multimodal systems, as well as directions for future research, in the paper.

Keywords: *Multimodal representation learning, Cross-modal Alignment, Video-Audio-Text, contrastive learning, ImageBind, unified multimodal models, Robust Artificial Intelligence*

[1] Department of Computer Science, Islamia College Peshawar, Pakistan.
[2] Department of Computer Science, Islamia College Peshawar, Pakistan.

Copyright (c) 2023 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

1. Introduction

The human perceptual system is naturally multimodal, whereby visual scenes, sounds, and language combine to form an understanding of the world. In recent years, artificial intelligence systems have made an effort to replicate this functionality by learning joint representations of video, audio and text. These modalities can be embedded in a shared embedding space and enable many applications such as zero-shot retrieval, audiovisual event localization, video question answering, multimodal generation and embodied agents for dealing with complex real-world environments (Li & Tang, 2024; Girdhar et al., 2023).

The first multimodal systems used individual encoders for each modality and only later combined features. Thanks to large scale contrastive pre-training, such as CLIP, for image–text pairs, it is clear that a shared embedding space trained on a large number of paired data can enable robust zero-shot transfer. Later research generalised this paradigm to more sensory modalities, such as audio (AudioCLIP, CLAP), video, and other senses. Additionally, ImageBind demonstrated that an image-centric hub can cause emergent alignment between modalities not directly aligned in their training, which can minimize the need for large multimodal corpora (Girdhar et al., 2023). On the other hand, large multimodal language models (LLMs) have been developed in parallel, such as the Video-LLaMA, Qwen-VL, and the Gemini-class models, which add continuous or discrete multimodal tokens to the decoder-only architecture and allow for the understanding and generation of both video, audio, and text.

These developments have not yet made it easy to build truly integrated and strong multimodal representations, however. However, temporal misalignment in long videos, modality imbalance, noise sensitivity, high computational cost, and the lack of high-quality multi-modal datasets remain a challenge for performance and generalization. In this paper, we review the latest developments in unified multimodal representation learning and focus on the alignment of video, audio and text. It summarizes architectural approaches, goals for training, empirical results and open problems to direct researchers and practitioners towards more robust AI systems.

[1] Department of Computer Science, Islamia College Peshawar, Pakistan.
[2] Department of Computer Science, Islamia College Peshawar, Pakistan.

Copyright (c) 2023 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

2. Literature Review

Cross-Modal Alignment: Foundations.

The goal of cross-modal alignment is to perform a mapping of the data from different modalities into a common latent space, in such a way that the semantically similar samples are grouped together, while the dissimilar are pushed away. The current paradigm is known as the contrastive approach. Methods like CLIP output pairs that are as similar as possible when they are matched, but as dissimilar as possible when they are not matched in the same batch (Radford et al., 2021). Similar to this, extensions to audio–text (CLAP) and audio–image–text (AudioCLIP) are based on the same idea, typically with InfoNCE or similar losses (Guzhov et al., 2022; Elizalde et al., 2023). There are two types of multimodal alignment: data-level fusion, feature-level fusion, and decision-level fusion; contrastive methods, attention-based methods, generative methods, and methods based on large language models (Li & Tang, 2024, 2025). The surveys, summarizing over 200-260 papers, highlight the need for both geometric proximity in embedding space, and semantic consistency with probes conducted downstream.

2.3 Local, Regional, and Global Alignment

To overcome this, ImageBind came up with a promising alternative to running pairwise training. The model is able to create emergent alignment between multiple modalities (audio, depth, thermal, IMU, text) without ever seeing pairs of these modalities together with the frozen image encoder (Girdhar et al., 2023). This hub approach significantly lowered the data needs and has impacted later on unified embedders. Contrastive audio-visual masked autoencoders (CAV-MAE and subsequent variations like CAV-MAE Sync) has shown that self-supervised reconstruction tasks on unlabelled video can further enhance fine-grained temporal and semantic alignment of audio and visual inputs. Preliminary studies have shown that the accuracy of the cross-modal retrieval task can be significantly boosted by projecting different types of audio encoders into the CLIP space, as some encoders have seen more than 40 percentage points after alignment.

2.3 Unified Multimodal Architectures

[1] Department of Computer Science, Islamia College Peshawar, Pakistan.
[2] Department of Computer Science, Islamia College Peshawar, Pakistan.

Copyright (c) 2023 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

In recent architectures, the idea has been to try to encode an arbitrary number of modalities using a single encoder, or more tightly coupled set of components. The model, UniAlign, scales a shared encoder with low-rank adaptation and modality-aware mixture-of-experts with relatively few trainable parameters (around 7.8 million) and without sharing weights across tasks (CVPR 2025). Other methods involve transforming video and audio into text representations in a structured format that can be processed by the large language models, aiming to obtain semantic and temporal alignment through a common language representation (UMaT framework, 2025). Video-LLaMA fuses audio and imagery features through either Q-Formers or linear projectors into the frozen or lightly tuned language-model backbone, enabling understanding and limited generation. However, recent multi-modal models that optimize both understanding and generation by either reconstruction or auto-encoder techniques have also demonstrated cross benefits of both abilities, obtaining measurable gains in generation and editing tasks after training with reconstruction alignment.

2.4 Temporal and Semantic Alignment Challenges

A video, as opposed to a static image–text combination, adds another dimension to time. For long videos with few salient events, there should be mechanisms for temporal segmentation, redundant filtering and hierarchical alignment. Systems that first generate short clip captions and/or automatic speech recognition (ASR) transcripts, and then organize the transcript into a temporal order of segments have shown significant improvements on long video question answering (absolute accuracy improvements of 9-17% on selected benchmarks depending on video length). Moreover, asymmetric hierarchical anchoring techniques tackle the ambiguity of information allocation between the audio and visual streams to better generalize across modalities in limited-supervised setting where supervision is only drawn from paired AV data. It is important to note from these results that embedding alignment is not enough for temporally extended content when embedding globally.

3. Research Gap

[1] Department of Computer Science, Islamia College Peshawar, Pakistan.
[2] Department of Computer Science, Islamia College Peshawar, Pakistan.
Copyright (c) 2023 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

While pairwise and hub-based alignment approaches have come to fruition, there are still missing elements in the quest for truly unified trimodal (video–audio–text) representations, which are kept robust when certain modalities are missing, the time difference between modalities is large, and the domain mismatch is large. A lot of high-performing systems depend on carefully curated large paired corpora, which are expensive to collect, and may not be easily generalizable to real world noise, or low resource languages and cultures. Comparisons of emergent versus explicitly-supervised alignment to the same benchmark are still limited. Furthermore, the relationship between the representations' quality of alignment and the fidelity of generative models has only started to be quantitatively examined recently. The gaps must be closed to enable reliable deployment of multimodal AI in unconstrained environments like robotics, autonomous systems, and assistive technologies.

4. Objectives

This synthesis has four objectives:

1. To identify the key training and architectural approaches to aligning video, audio, and text representations.
2. To review empirical data about the effectiveness of contrastive, hubs, and reconstruction based alignment techniques in retrieval, classification, and generation tasks.
3. To find remaining limitations that impair robustness in realistic scenarios such as temporal misalignment, modality imbalance and noise sensitivity.
4. To develop design guidelines for scalable and robust multimodal systems which can function with partial or degraded sensory inputs.

5. Methodology

The work is a structured literature synthesis of peer-reviewed papers, CVPR, related IEEE/CVF papers, NeurIPS-adjacent multimodal preprints from 2023 to 2025, which has some connections to cross-modal alignment involving a minimum of two of three target modalities (i.e., video, audio, text). Only those results of retrieval, classification, or generation tasks which were explicitly reported in terms of inclusion objectives, architectural features, and quantitative results were

[1] Department of Computer Science, Islamia College Peshawar, Pakistan.
[2] Department of Computer Science, Islamia College Peshawar, Pakistan.
Copyright (c) 2023 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

not excluded. Primary sources are: ImageBind, AudioCLIP, CLAP, Video-LLaMA and its successors (multimodal LLM), UniAlign, CAV-MAE variants, UMaT, reconstruction-alignment studies, multimodal alignment surveys of over 200 papers and related unified multimodal model studies. The findings have been grouped thematically instead of chronologically, focusing on commonalities and differences in the methods, and the empirical results. No experimental data was gathered, all the performance claims are due to the original studies.

6. Results

The effectiveness of Constrastive and Hub Based Alignment (CBA) is presented.Effectiveness of Constrastive and Hub Based Alignment (CBA) is presented.

Contrastive image–text models provided a solid foundation for zero-shot transfer. Applying the same goal to audio, competitive sound classification and cross-modal retrieval performance is achieved. ImageBind was shown to enable emergent alignment between the other modalities without using any audio–text pairs—it shows how to align additional modalities when all you have is an image hub. Later testing showed that when the audio encoders are realigned into a CLIP-compatible space, the top-1 audio-to-video retrieval rate can be greatly enhanced (up to 40% higher for some encoders after projection). The language-supervised audio encoders (CLAP, Whisper) also show increased generative compatibility after the realignment compared to purely audio-visual models, indicating that the linguistic supervision in pretraining provides long term benefits to multimodal generation.

By the end of this unit, students should be able to: 6.2 Understand the concept of Unified Architectures and the efficiency of parameters.

By complementing LoRA with a modality-aware mixture-of-experts design, UniAlign was able to achieve a comparable or superior performance to specialized state-of-the-art models for 6 or more modalities with fewer than 8 million trainable parameters on several benchmarks. The video-to-text unification strategies based on text show that large language models can achieve significant improvements in reasoning with long videos, especially when redundant

[1] Department of Computer Science, Islamia College Peshawar, Pakistan.
[2] Department of Computer Science, Islamia College Peshawar, Pakistan.
Copyright (c) 2023 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

information and the structure of the video are removed. Supporting generation and understanding as reconstruction, be it through a post-training approach or an auto-encoder, further improves both generation precision and cross-modal coherence with only a few tens of GPU hours for reported cases. The results suggest that unification does not need to involve extensive retraining of the joint network from the ground up.

Temporal Alignment and Long-Form Video (6.3)

Temporal approaches such as clip-level captioning, ASR integration, hierarchical anchoring, or masked autoencoding are more effective than global embedding-based methods in scenarios involving fine-grained event localization, or long context question-answering. Absolute gains on long-video QA tests range from 9-17% depending on the distribution of video lengths and QA test. These improvements demonstrate that the gains come not just from semantic alignment but also from ensuring that the correspondence is maintained or reconstructed within the right time frame when dealing with the realistic duration of video systems.

6.4 Robustness Considerations

Most published results are under relatively clean laboratory conditions. A less well quantified performance is observed in missing modalities, background noise, domain shift and asynchronous streams. Models that explicitly include robustness objectives as well as models able to degrade gracefully when using only a subsample of the modalities are typically more robust, but systematic stress-testing is still a missing requirement in the real world. There is still a lack of standard robustness measures for trimodal systems that make comparisons possible.

7. Discussion

These findings suggest that a common embedding space created by contrastive or hub-based tasks is a strong basis for cross-modal transfer. Here, image-centric binding provides an elegant solution to the combinatorial data problem, unified encoder designs and reconstruction goals drive towards true multi-modality, rather than simply pairwise association. However, there are three pragmatic

[1] Department of Computer Science, Islamia College Peshawar, Pakistan.
[2] Department of Computer Science, Islamia College Peshawar, Pakistan.
Copyright (c) 2023 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

challenges. First, the alignment is still very sensitive to the distribution of the training pairs and whether they are noisy or not—noisy data or poorly aligned data can lead to spurious "associations" that manifest as hallucinations or retrieval errors. Second, there are still limited and incomplete solutions to temporal dynamics in video, with global embeddings failing to capture fine-grained event structure, while higher computational costs are required for hierarchical or sequential models to partially recover event structure. Third, the gap between lab settings and real settings, where there is noise, missing streams, cultural or linguistic diversity, etc., is still significant.

In terms of the systems perspective, parameter-efficient adaptation methods (LoRA, modality-aware MoE) enable the extension of existing large models without retraining. While text-mediated unification can be a practical approach for harnessing the reasoning power of LLMs, it may compromise the fidelity of the continuous signal. It is particularly interesting for the perception side to feature continuous embeddings and to utilize language as the medium of reason.Combining continuous embeddings for perception with language as the reasoning medium seems promising for robust agents. Multimodal capacity also increases security, fairness and hallucination risks. Modality-specific inputs that are misaligned or adversarially perturbed can have an impact on decisions seemingly made in a different modality. It is therefore critical that these failure modes are further explored through evaluation protocols in order to ensure trustworthy deployment.

8. Conclusion

From pairwise-contrastive models to hub-based emergent alignment and parameter-efficient models with video, audio, and text in the same framework, the development of unified multimodal representation learning has come leaps and bounds. The overall progress in retrieval accuracy and generative coherence has been achieved through the imageBind-style binding, the contrastive objectives, the techniques of temporal structuring, and the techniques of reconstruction-based unification. Issues such as temporal synchronization, modality imbalance, data efficiency, and real-world robustness persist as

[1] Department of Computer Science, Islamia College Peshawar, Pakistan.
[2] Department of Computer Science, Islamia College Peshawar, Pakistan.
Copyright (c) 2023 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

motivators for research. Practical systems must also utilize hub or contrastive pretraining, parameter efficient adaptation and explicit robustness testing. These ingredients can provide a starting point for more robust and widely applicable artificial intelligence that could function in each sensory modality that humans use.

References

Elizalde, B., et al. (2023). CLAP: Learning audio concepts from natural language supervision. IEEE International Conference on Acoustics, Speech and Signal Processing.

Girdhar, R., et al. (2023). ImageBind: A single embedding space to bind them all. In the IEEE/CVF Computer Vision and Pattern Recognition Conference.

Guzhov, A., et al. (2022). The extension of CLIP to image, text and audio: AudioCLIP. IEEE International Conference on Acoustics, Speech and Signal Processing.

Li, S., & Tang, H. (2024). Multimodal Alignment and Fusion: A survey. arXiv preprint arXiv:2411.17040.

Li, S., & Tang, H. (2025/2026). Multimodal Alignment and Fusion: A survey. International Journal of Computer Vision, 134, Article 103. https://doi.org/10.1007/s11263-025-02667-1

Radford, A., et al. (2021). Learning visual models that can be transferred. International Conference on Machine Learning.

Shukor, M., Dancette, C., Ramé, A., & Cord, M. (2023). Single model for image, video, audio and language tasks. arXiv preprint arXiv:2307.16184.

Various authors. (2025a). Scaling multimodal alignment with a single model: UniAlign. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Various authors. (2025b). A simple unified multi-modal framework with semantic and temporal alignment (UMaT) is able to describe everything in words. arXiv preprint arXiv:2503.09081.

Various authors. (2025c). They discovered that alignment during reconstruction improves the unified multimodal models. arXiv preprint arXiv:2509.07295.

[1] Department of Computer Science, Islamia College Peshawar, Pakistan.
[2] Department of Computer Science, Islamia College Peshawar, Pakistan.

Copyright (c) 2023 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Various authors. (2025d). Unified multimodal model: auto-encoder. arXiv preprint arXiv:2509.09666.

Various authors. (2025e). Is it possible to replace visual information with sound in LLaVA with token substitution? arXiv preprint arXiv:2506.10416.

Zhang, H., et al. (2023–2025). Multimodal large language model architectures with visual and auditory streams, like Video-LLaMA and later

[1] Department of Computer Science, Islamia College Peshawar, Pakistan.
[2] Department of Computer Science, Islamia College Peshawar, Pakistan.

Copyright (c) 2023 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.