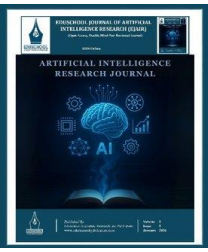
EDUSCHOOL JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH (EJAIR)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: 3139-7336



Cross-Modal Retrieval with Modality-Agnostic Embeddings Towards Universal Representation Learning

Rejina P V

Assistant professor, Co-Operative Arts and Science College, Madayi , Pazhayangadi, Kannur, India

Article information

Received: 3rd August 2026
Received in revised form: 5th September 2026
Accepted: 7th October 2026
Available online: 17th November 2026

Volume: 1
Issue: 3
DOI: https://doi.org/10.63090/EJAIR/3139.7336.0010

Abstract

A single vector space that accepts images, text, audio, video and sensor readings, and that supports retrieval between any pair of them, has become an explicit goal of multimodal research. Systems now exist that approach it by binding every modality to one anchor, which produces retrieval between pairs that were never trained together. This article examines what such spaces achieve and where they fail. We first analyse the alignment topologies available, showing that binding through an anchor reduces the requirement for paired data from a quadratic to a linear function of the number of modalities, and that the reduction has a measurable cost in the quality of pairings the anchor does not mediate. We then examine the geometry of the resulting space. Published analysis has shown that representations of different modalities occupy separate regions rather than mixing, a separation that persists throughout training and that has consequences for retrieval which are widely overlooked. Hubness and anisotropy compound the effect at the scale required for practical search. We compare seven systems by their binding topology and coverage, review the approximate search structures that make retrieval feasible over billions of items, and assess the claim that models trained on different modalities are converging on a shared statistical description of the world. Six research problems are identified.

Keywords:- Cross-Modal Retrieval, Modality-Agnostic Embeddings, Modality Gap, Binding Topology, Hubness, Universal Representation

.

I. INTRODUCTION

Retrieval across modalities was once a matter of building a bridge between two specific representations, with a separate bridge for each pair of interest. Contrastive pre-training on image and caption data replaced the bridge with a shared space in which both inputs are encoded as vectors of equal dimension and compared by inner product [1], [2]. The construction generalises. If audio is aligned to images, and text is aligned to images, then audio and text can be compared through the space even though no audio and text pair was ever presented during training. Systems built on this observation report retrieval between such pairs and describe the capability as emergent [3], [4].

The appeal is practical as well as conceptual. A single index serves every query type, a new modality can be added by training one encoder against an existing anchor, and downstream systems need to handle only one kind of vector. The appeal has carried the idea into deployment, where shared embedding spaces now support search over mixed collections of photographs, documents and recordings. What has received less attention is the internal structure of these spaces and whether the geometry supports the uses to which they are being put.

It does not, at least not without correction. Analysis of contrastive spaces has established that embeddings of different modalities occupy distinct regions separated by a substantial distance, that the separation is present at initialisation and is preserved rather than removed by training, and that it affects downstream behaviour [8]. A space in which every image lies in one region and every caption in another is not modality-agnostic in any meaningful sense, and the fact that retrieval works despite this reflects the local ordering of similarities rather than any genuine mixing. Two further properties of high dimensional spaces, namely the emergence of hub points that appear among the nearest neighbours of a large number of queries [9] and the concentration of embeddings within a narrow cone, compound the difficulty at the scale of a practical index.

This article treats the shared space as an object of study rather than as a means to an end. Section II compares alignment topologies and quantifies the paired data each requires. Section III examines the geometry of the resulting space and the three properties that govern retrieval quality. Section IV compares published systems and reviews the search structures needed at scale. Section V assesses the claim that representations from different modalities are converging on a common description of the world, and states six open problems.

II. ALIGNMENT TOPOLOGIES

Three arrangements are possible when more than two modalities are involved, and Fig. 1 shows them. Pairwise alignment trains a separate objective for every pair, which requires paired data for each and produces the strongest correspondence for the pairs that receive it. Anchor binding aligns every modality to one designated modality, usually images or text, and relies on transitivity for the rest. Joint encoding maps all modalities through a single network with modality-specific input adapters, so that alignment is a consequence of shared parameters rather than of an explicit pairwise objective [10], [11], [12].

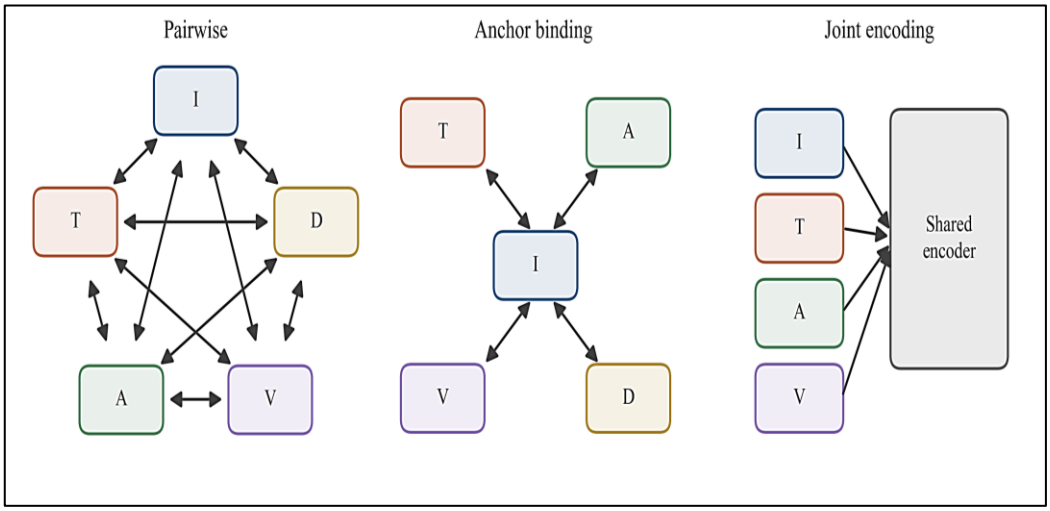


Fig. 1: Three alignment topologies for a shared embedding space over five modalities. Pairwise alignment requires paired data for every edge, anchor binding requires it only for edges incident on the anchor, and joint encoding replaces explicit pairing with shared parameters and modality-specific adapters.

The difference in data requirement is decisive. Pairwise alignment over N modalities requires paired corpora for N(N-1)/2 combinations, while anchor binding requires N-1. Fig. 2 plots both, and the divergence is severe by the time a system covers six or seven signal types.

The practical position is worse than the counting suggests, because paired corpora do not exist in comparable quantity for most combinations. Image and text pairs are available in billions [17], audio and text pairs in the tens of thousands [20], [30], and pairs joining depth or inertial measurement to language are scarce. Anchor binding is therefore not only cheaper but is the only arrangement for which the data exist, which is the reason it has become standard [3], [4].

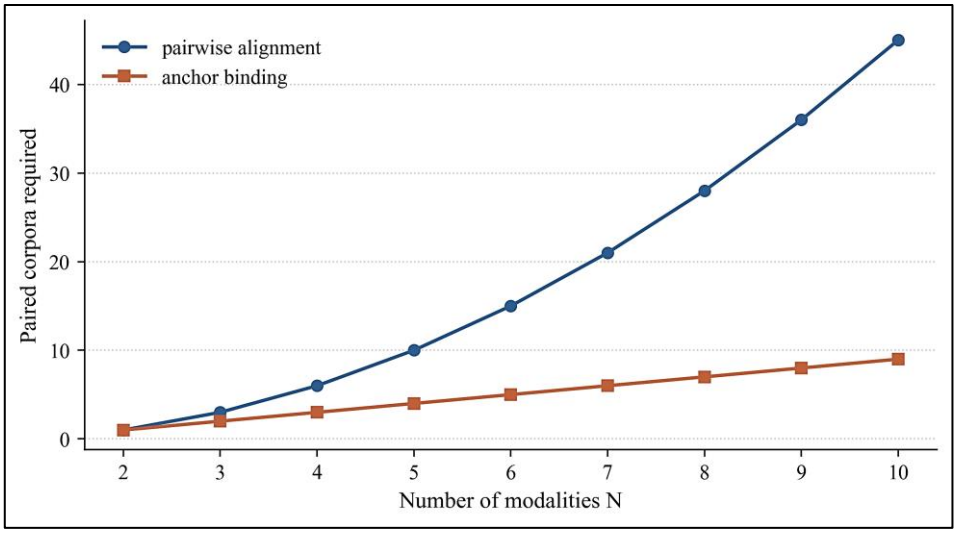


Fig. 2: Number of paired corpora required to align N modalities under pairwise alignment and under anchor binding. Values are computed from the combinatorial expressions N(N-1)/2 and N-1 respectively.

The cost of the reduction is that transitivity is not exact. Alignment of audio to images preserves the information the two share, and alignment of text to images preserves a different overlap. What survives the composition of the two is the intersection, so retrieval between audio and text through an image anchor is restricted to content that images can express. A sound that has no visual correlate, such as the tone of an utterance or a mechanical fault audible in a running engine, has no representation the anchor can carry. Reported results for emergent pairs are consistent with this account, since the accuracy attained on pairs mediated by an anchor is below that attained on directly trained pairs even when the encoders are the same [3]. Table 1 sets out the properties of the three topologies.

Table 1. Properties of the three alignment topologies

Topology	Paired corpora needed	Emergent pairs	Principal advantage	Principal limitation
Pairwise	N(N-1)/2	Not required	Strongest correspondence per pair	Data do not exist for most pairs
Anchor binding	N-1	Yes, through the anchor	New modality added by one training run	Content absent from the anchor cannot transfer
Joint encoding	Varies with supervision	Yes, through shared parameters	One network for all inputs	Capacity shared across dissimilar signals

The encoders are largely common across the three topologies, which is worth stating because it locates the differences in the objective rather than in the network. Almost every system uses transformer blocks [29] with patch or token embeddings supplying the input, following the formulation established for images [28], and almost every alignment objective descends from noise contrastive estimation applied to a batch of candidates [16]. Designs that retain a fusion encoder rather than producing independent vectors achieve finer alignment and are used for reranking a shortlist [27], [32], and the two modality retrieval systems built for video follow the same pattern with temporal pooling added [25], [26]. Unified masked modelling over image and text tokens reaches comparable quality with a single objective [22]. What separates the modality-agnostic systems from these is not the encoder but the decision to align many modalities into one space, and the properties examined in the next section are consequences of that decision.

III. THE GEOMETRY OF SHARED SPACES

A. The modality gap

The most consequential property of contrastive embedding spaces is that they do not mix. Embeddings of images occupy one region of the hypersphere and embeddings of text another, with a clear separation between them, and the separation is not an artefact of insufficient training. It arises from the narrow cone in which randomly

initialised deep networks place their outputs, combined with a contrastive objective that preserves the initial separation because moving the two groups together is not required to rank the correct pair first [8]. Fig. 3 illustrates the arrangement on the unit circle. The correct pair is closer than any incorrect pair, so retrieval succeeds, yet no image embedding is close to any text embedding in absolute terms.

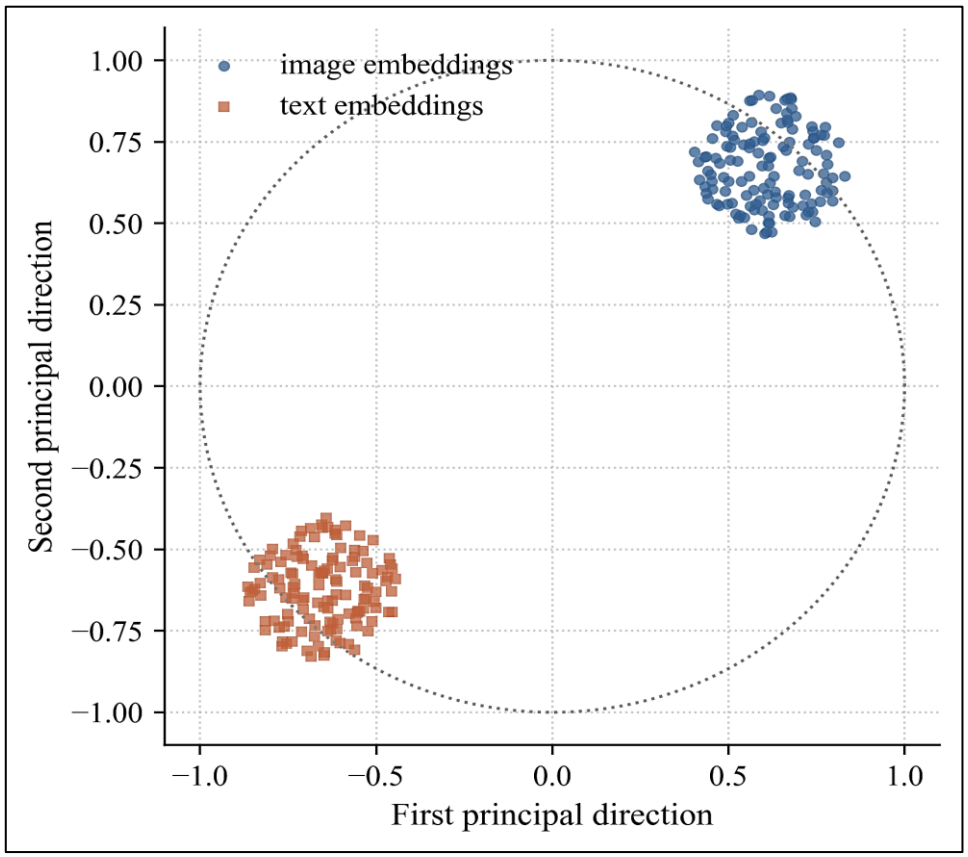


Fig. 3: Schematic illustration of the modality gap on the unit sphere, drawn from a fixed seed. Embeddings of the two modalities occupy separate regions while the ranking within each region supports retrieval. The arrangement follows the analysis reported in [8].

Three consequences follow for retrieval. Scores are not comparable across modality pairs, so a threshold calibrated for image to text search is wrong for audio to text search, and any system that fuses results from several pair types must recalibrate. Arithmetic in the space is unreliable, since adding a text vector to an image vector produces a point in the empty region between the two groups where no encoder places anything. Fine tuning that moves the groups together can improve or degrade downstream accuracy depending on the task, which means the gap is not simply a defect to be removed [8]. Practical remedies include centring each modality by subtracting its mean before comparison, whitening within modality, and training with a formulation that scores each pair independently rather than through a softmax over the batch [21].

B. Hubness and anisotropy

Two further properties appear at scale. Hubness is the tendency in high dimensional spaces for a small number of points to appear in the neighbour lists of a disproportionate share of queries, a phenomenon established for general vector data and directly relevant to retrieval [9]. In a cross-modal index, hub items are returned for queries they do not match, which lowers precision in a way that averages conceal because the affected queries are a minority. Corrections that normalise similarity by the average similarity of each candidate to a sample of queries are effective and inexpensive, and their absence from most published retrieval pipelines is an oversight rather than a considered choice.

Anisotropy is the concentration of embeddings within a narrow cone, which compresses the range of achievable similarity scores and reduces the separation between a correct match and a plausible distractor. The effect interacts with quantisation, since a compression scheme that allocates codewords uniformly over the sphere wastes capacity on regions no embedding occupies [15]. Whitening or applying a learned rotation before quantisation recovers part of the loss and is standard practice in text retrieval, though it is applied inconsistently in the cross-modal setting.

IV. SYSTEMS AND RETRIEVAL AT SCALE

Table 2 compares seven systems that construct shared spaces across more than two modalities. The pattern is consistent, with images or text serving as the anchor and additional modalities attached by contrastive objectives against the anchor. Coverage has grown from three modalities to six, and the newer systems use language rather than images as the anchor on the argument that language expresses more of what the other modalities carry [4]. Systems that avoid the anchor entirely by encoding all inputs through shared parameters have demonstrated broad task coverage [10], [11], [12], [23], though their retrieval accuracy on standard cross-modal benchmarks has not matched that of dedicated contrastive models.

Table 2. Systems constructing shared embedding spaces over multiple modalities

System	Year	Modalities covered	Anchor	Alignment objective	Ref.
CLIP	2021	Image, text	Not applicable	Symmetric InfoNCE	[1]
VATT	2021	Video, audio, text	Video	Hierarchical contrastive	[6]
AudioCLIP	2022	Image, text, audio	Image and text	Contrastive extension	[5]
CLAP	2023	Audio, text	Not applicable	Contrastive	[7], [24]
ImageBind	2023	Image, text, audio, depth, thermal, inertial	Image	Pairwise contrastive to anchor	[3]
LanguageBind	2024	Video, infrared, depth, audio, text	Language	Pairwise contrastive to anchor	[4]
Meta-Transformer	2023	Twelve input types	Not applicable	Shared encoder with adapters	[12]

Evaluation of these systems relies on benchmarks built for two modalities, principally curated image and caption sets [18], [19] and audio caption collections [20]. No benchmark measures the property that distinguishes a modality-agnostic space, namely consistency of the geometry across all pairs. A system could score well on every pair separately while placing the modalities in mutually inconsistent arrangements, and no reported protocol would detect it. A suitable measure would compare the similarity assigned to a triple of items describing the same content across three modalities and report the variance of those similarities, which would test transitivity directly.

Retrieval at deployment scale depends on approximate search. Exhaustive comparison against a billion vectors is infeasible for interactive use, so systems use graph based indexes that traverse a hierarchy of neighbour links [14] or partitioned indexes with compressed residuals, with product quantisation the standard compression [15], and implementations on parallel hardware make billion scale search practical [13]. The interaction between these structures and the geometry described in Section III is the part most often ignored. A graph index built over embeddings that occupy two separated regions will place few edges between the regions, so a query from one modality enters the graph far from its target and the search must cross the gap before it can descend. Building separate indexes per modality and centring each one removes the problem and is the recommended arrangement.

V. TOWARDS UNIVERSAL REPRESENTATIONS

The strongest form of the universal representation claim is that models trained on different data and different modalities converge on the same underlying description of the world, so that alignment amounts to discovering a rotation between spaces that already agree. Evidence has been offered for this position in the form of measurements showing that representational similarity between independently trained models increases with their scale and performance, across vision and language alike [34]. The analogy with unsupervised alignment of word embeddings between languages is instructive, since translation without parallel data was achieved by finding a linear map between separately trained spaces on the basis of their shared structure [33].

The evidence supports a weaker reading than the strong claim requires. Convergence is measured between models trained on data that overlap in content, and the shared structure may reflect a shared world rather than a shared representational format. The modality gap is direct evidence against complete convergence, since spaces that agreed would not remain separated after alignment. A reasonable position is that a common core exists for content that several modalities can express, that this core is what anchor binding transfers, and that each modality retains information outside it which no shared space will carry.

Six problems follow. The first is a measure of transitivity, without which the central property of a modality-agnostic space is unmeasured. The second is representation of modality-specific content, which a single shared

vector cannot hold and which a composite representation with a shared and a private component could. The third is calibration of similarity across pairs, so that scores from different modality pairs can be compared and combined. The fourth is the treatment of hubness and anisotropy as part of the retrieval pipeline rather than as post hoc corrections. The fifth is the addition of modalities without retraining, since the anchor arrangement requires a new contrastive run for each and cheaper adaptation methods have been demonstrated for the two modality case [31]. The sixth is evaluation on modalities with weak visual correlates, such as inertial measurement and thermal imaging, where anchor binding through images is least likely to transfer and where reported results are thinnest [3].

VI. CONCLUSION

Shared embedding spaces have made cross-modal retrieval a matter of encoding rather than of engineering, and binding through an anchor has reduced the paired data requirement from a quadratic to a linear function of the number of modalities, which is what makes coverage of six or more signal types feasible. The arrangement carries a cost that is intrinsic rather than incidental, since only content the anchor can express passes between the modalities it mediates.

The geometry of these spaces deserves more attention than it receives. Modality separation persists through training and affects score calibration, vector arithmetic and index construction. Hubness and anisotropy degrade precision at the scale at which these systems are deployed, and inexpensive corrections exist but are inconsistently applied. The claim that independently trained models are converging on a universal representation is supported for content that several modalities share and is contradicted by the separation the same spaces exhibit. Progress requires a measure of transitivity, a representation that admits modality-specific content alongside the shared component, and evaluation on signals for which no visual anchor is available.

REFERENCES

[1] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

[2] C. Jia et al., “Scaling up visual and vision-language representation learning with noisy text supervision,” in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 4904–4916.

[3] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, “ImageBind: One embedding space to bind them all,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 15180–15190.

[4] B. Zhu et al., “LanguageBind: Extending video-language pretraining to N-modality by language-based semantic alignment,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

[5] A. Guzhov, F. Raue, J. Hees, and A. Dengel, “AudioCLIP: Extending CLIP to image, text and audio,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2022, pp. 976–980.

[6] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, “VATT: Transformers for multimodal self-supervised learning from raw video, audio and text,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021, pp. 24206–24221.

[7] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, “CLAP: Learning audio concepts from natural language supervision,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023, pp. 1–5.

[8] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Zou, “Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 17612–17625.

[9] M. Radovanovic, A. Nanopoulos, and M. Ivanovic, “Hubs in space: Popular nearest neighbors in high-dimensional data,” *J. Mach. Learn. Res.*, vol. 11, pp. 2487–2531, 2010.

[10] A. Jaegle et al., “Perceiver IO: A general architecture for structured inputs and outputs,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[11] J. Lu, C. Clark, R. Zellers, R. Mottaghi, and A. Kembhavi, “Unified-IO: A unified model for vision, language, and multi-modal tasks,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.

[12] Y. Zhang et al., “Meta-Transformer: A unified framework for multimodal learning,” *arXiv preprint* arXiv:2307.10802, 2023.

[13] J. Johnson, M. Douze, and H. Jegou, “Billion-scale similarity search with GPUs,” *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, 2021.

[14] Y. A. Malkov and D. A. Yashunin, “Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 824–836, 2020.

[15] H. Jegou, M. Douze, and C. Schmid, “Product quantization for nearest neighbor search,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 117–128, 2011.

[16] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint* arXiv:1807.03748, 2018.

[17] C. Schuhmann et al., “LAION-5B: An open large-scale dataset for training next generation image-text models,” in *Adv. Neural Inf. Process. Syst. (NeurIPS) Datasets and Benchmarks Track*, 2022.

[18] T.-Y. Lin et al., “Microsoft COCO: Common objects in context,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.

[19] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 2641–2649.
[20] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL-HLT)*, 2019, pp. 119–132.
[21] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 11975–11986.
[22] W. Wang et al., "Image as a foreign language: BEiT pretraining for vision and vision-language tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 19175–19186.
[23] S. Reed et al., "A generalist agent," *Trans. Mach. Learn. Res.*, 2022.
[24] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023, pp. 1–5.
[25] H. Xu et al., "VideoCLIP: Contrastive pre-training for zero-shot video-text understanding," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2021, pp. 6787–6800.
[26] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 1728–1738.
[27] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-language transformer without convolution or region supervision," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 5583–5594.
[28] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
[29] A. Vaswani et al., "Attention is all you need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.
[30] D. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2020, pp. 736–740.
[31] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022.
[32] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, 2023, pp. 19730–19742.
[33] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jegou, "Word translation without parallel data," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.
[34] M. Huh, B. Cheung, T. Wang, and P. Isola, "The platonic representation hypothesis," in *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, 2024.