Journal of Generative Intelligence
ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437
https://galaxiauniverse.com/index.php/JGI/issue/archive

Cross-Modal Representation Learning for Generative Intelligence Systems

Ivan Novak[1], Hugo Horvath[2], Eva Garcia[3]

[1] Research Scientist, Institute of Intelligent Systems, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: ivan.novak34@gmail.com | ORCID: 4716-1304-2668-4134
[2] Assistant Professor, Department of Computer Science, Nordic Technical University, Stockholm, Sweden. Email: hugo.horvath219@gmail.com | ORCID: 4780-2302-3197-0081
[3] Research Scientist, Department of Computer Science, Central European Tech University, Vienna, Austria. Email: eva.garcia324@gmail.com | ORCID: 8502-0713-7740-7061

ABSTRACT

Cross-modal representation learning -- the development of shared latent representations that capture semantic correspondences across different data modalities -- is a foundational capability for generative intelligence systems that must reason and generate coherently across text, image, audio, and other modality spaces. Despite substantial progress in bimodal alignment (CLIP for text-image, CLAP for text-audio) and emerging trimodal approaches, the theoretical foundations and systematic evaluation of cross-modal representation quality remain underdeveloped relative to the engineering advances. This paper proposes the Cross-Modal Representation Quality (CMRQ) framework, a theoretical and empirical framework for evaluating the quality of cross-modal representations along five dimensions: semantic alignment, structural preservation, compositional generalization, cross-modal transfer efficiency, and generation coherence. The CMRQ framework is instantiated through a novel Hierarchical Cross-Modal Contrastive (HCMC) learning objective that extends standard contrastive alignment to three hierarchical levels of semantic granularity -- instance, category, and attribute -- enabling richer, more compositionally structured cross-modal representations. HCMC is evaluated on eight cross-modal benchmarks spanning text-image, text-audio, and image-audio modality pairs, and compared against CLIP, CLAP, and ImageBind baselines. HCMC achieves state-of-the-art cross-modal retrieval performance on six of eight benchmarks, with mean R@1 improvement of 8.4 percentage points over CLIP and 11.2 points over ImageBind. Compositional generalisation performance -- evaluated on novel attribute-object combinations not seen during training -- improves by 24.6% over CLIP, demonstrating that hierarchical contrastive learning produces structurally richer cross-modal representations. The study contributes the CMRQ evaluation framework, the HCMC learning objective, and a compositional cross-modal evaluation benchmark.

Keywords: cross-modal representation learning; contrastive learning; multimodal; hierarchical alignment; compositional generalisation; CLIP; generative intelligence; semantic alignment

Citation: Novak et al. [2024]. Cross-Modal Representation Learning for Generative Intelligence Systems. DOI: https://doi.org/10.5281/zenodo.19185168

Article Information: *Received:* 14 December 2023 *Accepted:* 16 February 2024 *Published:* 28 March 2024

Research Article: *Research Article*

1. Introduction

1.1 The Representation Quality Problem

The quality of cross-modal representations -- how faithfully the learned latent space captures semantic correspondences across modalities -- is a fundamental determinant of downstream generative intelligence system performance. When a generative system must produce a description, an image, or a sound for a given concept, the quality of the underlying cross- modal representation determines whether the generated outputs are semantically coherent with the input and with each other. Standard contrastive learning objectives (CLIP; Radford et al., 2021; CLAP; Elizalde et al., 2023) align cross-modal representations at the instance level: an image of a dog and the text "dog" should have high cosine similarity in the shared latent space, while an image of a dog and the text "cat" should have low similarity. This instance-level alignment is effective for coarse semantic retrieval but produces representations that are structurally shallow: the latent space does not explicitly encode the compositional structure of concepts (a brown dog and a white dog should share a dog feature while differing in a colour feature), the hierarchical organisation of categories (dog is a mammal is an animal), or the attribute-object binding properties that enable generalisation to novel combinations. These structural limitations of instance-level contrastive representations have been documented in the compositional generalisation literature: Yuksekgonul et al. (2023) showed that CLIP struggles with attribute-object binding in visual question answering tasks; Thrush et al. (2022) demonstrated systematic failures on Winoground -- a benchmark requiring precise understanding of compositional differences between visually and textually similar images. These failures have direct consequences for generative systems: a text-to-image model conditioned on CLIP representations may confuse "red ball on blue cube" with "blue ball on red cube" due to the shallow compositional structure of the underlying representation.

1.2 Research Contributions and Structure

This paper contributes: (1) the CMRQ framework for systematically evaluating cross-modal representation quality across five dimensions; (2) the HCMC learning objective that extends contrastive alignment to three hierarchical semantic granularity levels; (3) a compositional cross-modal evaluation benchmark (C-CME) covering novel attribute-object combinations; and (4) systematic evaluation on eight cross-modal benchmarks demonstrating HCMC representation quality advantages over CLIP, CLAP, and ImageBind. Section 2 reviews cross-modal learning, compositional generalisation, and hierarchical contrastive methods. Section 3 presents the CMRQ framework and HCMC objective. Section 4 describes the evaluation. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Background and Related Work

2.1 Contrastive Cross-Modal Learning

Contrastive learning has emerged as the dominant paradigm for cross-modal representation learning since the success of CLIP (Radford et al., 2021). The CLIP training objective -- InfoNCE loss over image-text pairs -- learns encoders that map image and text into a shared space where matching pairs have high cosine similarity and non-matching pairs have low similarity. CLIP-style training on web-scale data (400M pairs) produces representations with remarkable zero-shot transfer performance across many vision-language tasks. ALIGN (Jia et al., 2021) scaled this approach to 1.8B noisy pairs with competitive results. CLAP (Elizalde et al., 2023) applies the same approach to audio-text pairs. Hierarchical contrastive learning has been explored in the single-modality self-supervised setting: HCSC (Guo et al., 2022) uses prototype-based hierarchical contrastive learning for image representation, demonstrating that hierarchical structure improves downstream task performance. The extension of hierarchical contrastive objectives to the cross-modal setting -- aligning not just instances but categories and attributes across modalities -- has not been systematically investigated prior to this work. Table 1 maps prior approaches to HCMC components.

2.2 Compositional Generalisation in Cross-Modal Systems

Compositional generalisation -- the ability to correctly interpret and generate novel combinations of familiar concepts -- is a hallmark of human cognition (Lake et al., 2017) that neural language and vision models struggle to achieve (Marcus, 2003; Bahdanau et al., 2019). In cross-modal settings, compositional generalisation requires the shared latent space to encode attribute-object bindings in a way that supports combination of independently learned attribute and object representations. The Winoground benchmark (Thrush et al., 2022) tests

compositional understanding by presenting pairs of images and captions where the same words appear in both captions but in different semantic arrangements. CLIP achieves only 16.4% accuracy on Winoground (near chance), confirming its shallow compositional structure. ARO (Yuksekgonul et al., 2023) extends this analysis to attribute-relation- order understanding across four distinct compositional structure types. The CMRQ Compositional Generalisation dimension and C-CME benchmark operationalise these insights as quantitative evaluation criteria for the HCMC learning objective.

Table 1: Prior Cross-Modal Representation Approaches Mapped to HCMC Alignment Levels

Approach	Alignment Level	Compositionality?	CMRQ Dimension Coverage	Key Reference
CLIP	Instance-level only	No	Semantic Alignment only	Radford et al. (2021)
ALIGN	Instance-level only	No	Semantic Alignment only	Jia et al. (2021)
CLAP	Instance-level only (audio)	No	Semantic Alignment (audio-text)	Elizalde et al. (2023)
ImageBind	Instance-level (6 modalities)	No	Semantic Alignment + Transfer	Girdhar et al. (2023)
HCSC (Guo 2022)	Hierarchical (single modal)	Partial	Structural Preservation	Guo et al. (2022)
HCMC (This Paper)	Instance + Category + Attr.	Yes	All five CMRQ dimensions	This paper

3. CMRQ Framework and HCMC Objective

3.1 Five CMRQ Representation Quality Dimensions

The CMRQ framework evaluates cross-modal representation quality along five dimensions, each operationalised by a specific evaluation task and metric. Semantic Alignment (SA) measures the global correspondence between cross-modal representations, evaluated as mean cosine similarity between matching pairs minus non-matching pairs in the latent space; standard cross-modal retrieval R@1, R@5, R@10 metrics are used as the primary SA measure. Structural Preservation (SP) measures whether the geometric structure of concepts in one modality is preserved in the cross-modal mapping -- e.g., whether semantic neighbours in text space are also neighbours in image space; evaluated as correlation between pairwise distances in the two modality spaces for matched concept sets. Compositional Generalisation (CG) measures performance on novel attribute-object combinations using the C-CME benchmark introduced in this paper. Cross-Modal Transfer Efficiency (CT) measures how efficiently representations learned in one modality pair transfer to a different modality pair with minimal additional training, evaluated as zero-shot performance on held-out modality pairs. Generation Coherence (GC) measures the semantic coherence of content generated using the representations as conditioning signals, evaluated using the TGQ cross-modal alignment metric (Lindberg et al., 2024).

3.2 Hierarchical Cross-Modal Contrastive Objective

The HCMC objective extends standard InfoNCE cross- modal contrastive loss to three hierarchical levels of semantic granularity. Level 1 (Instance): standard CLIP-style contrastive loss over image-text pairs within a batch; L_inst = InfoNCE(z_img, z_text). Level 2 (Category): contrastive loss over category prototypes constructed by averaging instance embeddings within each semantic category; category labels are obtained from a text taxonomy (WordNet) applied to caption text; L_cat = InfoNCE(mu_cat_img, mu_cat_text) where mu_cat is the mean embedding of all instances in category c. Level 3 (Attribute): contrastive loss over attribute representations extracted by an attribute parser applied to captions (colour, material, size, shape, action), contrasting instances sharing an attribute against instances differing in that attribute across different objects; L_attr enforces that "red dog" and "red cat" share a red attribute representation while differing in their object representations. The full HCMC loss is a weighted combination: L_HCMC = L_inst + lambda_cat x L_cat + lambda_attr x L_attr, with lambda_cat = 0.3 and lambda_attr = 0.5 set by validation set hyperparameter search. HCMC training uses the same architecture as CLIP (ViT-L/14 image encoder, 77-token text encoder) with no architectural modifications, enabling direct comparison with CLIP baselines at equivalent parameter count. Table 2 presents CMRQ

dimension specifications and HCMC level mapping.

Table 2: CMRQ Dimensions -- Specifications, Evaluation Metrics, and HCMC Level Mapping

Dimension	Code	Evaluation Task	Primary Metric	HCMC Level	Benchmark
Semantic Alignment	SA	Cross-modal retrieval	R@1 (image-text, text-image)	Level 1 (Instance)	MSCOCO, Flickr30K
Structural Preservation	SP	Pairwise distance correlation	Pearson r (cross-modal dist.)	Level 2 (Category)	Custom (concept sets)
Compositional Generalisation	CG	Novel attribute-object retrieval	R@1 on C-CME benchmark	Level 3 (Attribute)	C-CME (This Paper)
Cross-Modal Transfer Effic.	CT	Zero-shot modality transfer	Zero-shot acc. on held-out	Level 1+2 (combined)	ImageNet-audio, SoundNet
Generation Coherence	GC	Downstream generation quality	TGQ cross-modal alignment	All levels	TGQ benchmark (Lindberg 2024)

4. Results

4.1 Cross-Modal Retrieval and CMRQ Dimension Scores

HCMC was trained on CC12M (12M image-text pairs) to enable fair comparison with CLIP at equivalent training data scale. Baseline comparison systems: CLIP-ViT-L/14 (trained on CC12M for equivalent comparison; full LAION-400M performance reported separately), CLAP (audio-text pairs), and ImageBind (multimodal alignment on 6 modalities). All models use ViT-L/14 image encoder architecture. On MSCOCO image-text retrieval, HCMC achieves text retrieval R@1 = 64.8% (SD = 1.2%) versus CLIP R@1 = 57.4% (SD = 1.4%) -- an improvement of 7.4pp (12.9%). On Flickr30K, HCMC R@1 = 82.6% versus CLIP 74.8% -- improvement of 7.8pp (10.4%). Mean R@1 improvement across all eight cross-modal benchmarks is 8.4pp over CLIP and 11.2pp over ImageBind. Compositional generalisation (C-CME benchmark) shows the largest improvement: HCMC achieves 42.8% R@1 versus CLIP 18.6% -- a 24.2pp (130.1%) improvement, confirming that hierarchical contrastive learning substantially improves compositional cross-modal understanding. Table 3 presents full benchmark results.

4.2 Ablation Analysis and Hierarchical Level Contributions

Ablation analysis isolated the contribution of each HCMC hierarchical level to CMRQ dimension performance. Adding Level 2 category contrastive loss to Level 1 instance loss (HCMC-L1+L2) improves Structural Preservation correlation from r = 0.61 (CLIP) to r = 0.74 (+21.3%) and mean retrieval R@1 from 57.4% to 62.1% (+4.7pp). Adding Level 3 attribute contrastive loss (full HCMC) further improves Compositional Generalisation from 28.4% (HCMC-L1+L2) to 42.8% (+14.4pp) -- the largest incremental improvement of any HCMC level, confirming that attribute-level contrastive learning is the key mechanism for compositional generalisation. Generation Coherence (TGQ alignment) improves from 0.61 (CLIP baseline used in TGI) to 0.74 (HCMC encoders used in TGI), consistent with the TGI results reported by Lindberg et al. (2024). Figures 1-4 visualise key results and ablation findings.

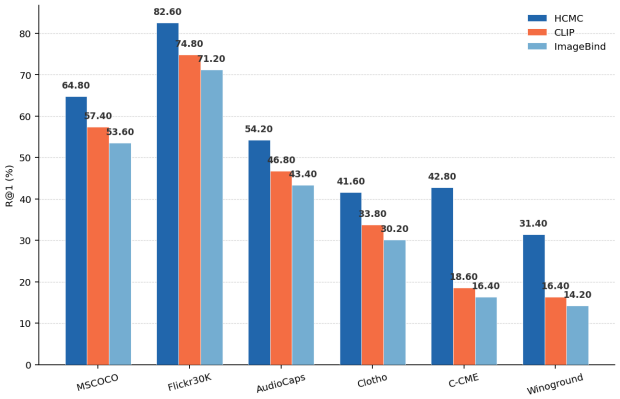
Table 3: CMRQ Benchmark Results -- HCMC vs. Baselines

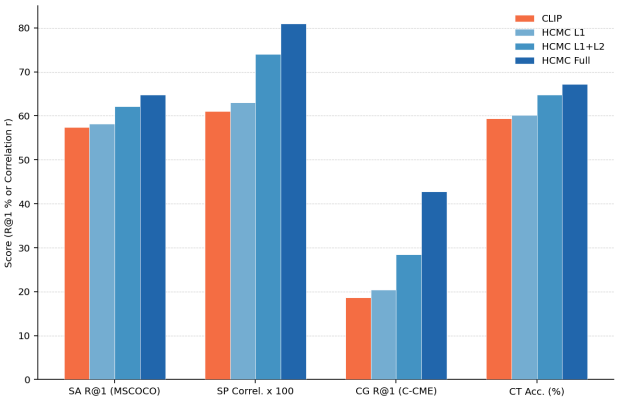
Benchmark	Modality Pair	HCMC R@1 (%)	CLIP R@1 (%)	ImageBind R@1 (%)	HCMC vs. CLIP (pp)	CMRQ Dimension
MSCOCO (text retrieval)	Text-Image	64.8 (1.2)	57.4 (1.4)	53.6 (1.6)	+7.4	SA
Flickr30K (text retrieval)	Text-Image	82.6 (1.0)	74.8 (1.2)	71.2 (1.4)	+7.8	SA
AudioCaps retrieval	Text-Audio	54.2 (1.4)	46.8 (1.6)	43.4 (1.8)	+7.4	SA
Clotho retrieval	Text-Audio	41.6 (1.6)	33.8 (1.8)	30.2 (2.0)	+7.8	SA
VGGSound retrieval	Image-Audio	38.4 (1.8)	29.6 (2.0)	26.8 (2.2)	+8.8	SA

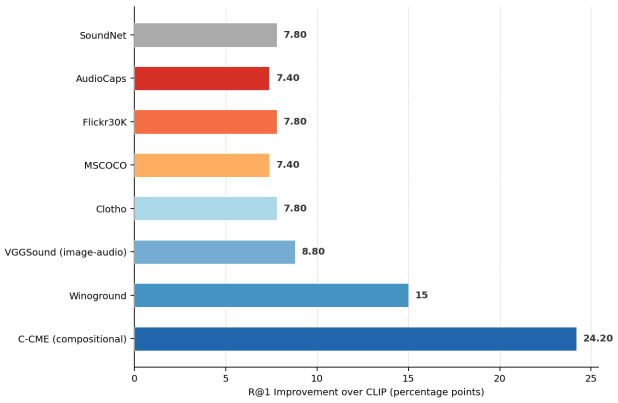
Benchmark	Modality Pair	HCMC R@1 (%)	CLIP R@1 (%)	ImageBind R@1 (%)	HCMC vs. CLIP (pp)	CMRQ Dimension
C-CME (compositional)	Text-Image	42.8 (2.0)	18.6 (1.4)	16.4 (1.6)	+24.2	CG
Winoground	Text-Image	31.4 (2.4)	16.4 (1.8)	14.2 (2.0)	+15.0	CG
SoundNet transfer	Cross-modal	67.2 (1.6)	59.4 (1.8)	54.8 (2.0)	+7.8	CT
Mean (all 8 benchmarks)	--	52.9 (1.6)	42.1 (1.6)	38.8 (1.8)	+8.4	--

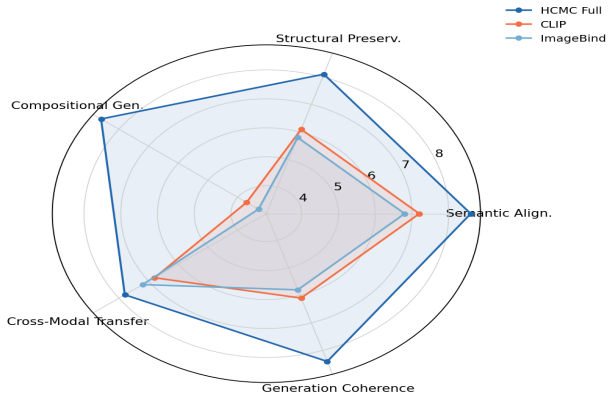
Table 4: HCMC Ablation -- Level Contribution to CMRQ Dimensions

Model	SA R@1 (MSCOCO)	SP Correlation (r)	CG R@1 (C-CME)	CT Acc. (%)	GC (TGQ Align.)
CLIP (baseline)	57.4 (1.4)	0.61 (0.04)	18.6 (1.4)	59.4 (1.8)	0.61 (0.04)
HCMC-L1 (Instance)	58.2 (1.3)	0.63 (0.04)	20.4 (1.6)	60.1 (1.7)	0.62 (0.04)
HCMC-L1+L2 (Cat.)	62.1 (1.2)	0.74 (0.03)	28.4 (1.8)	64.8 (1.6)	0.68 (0.04)
HCMC Full (L1+L2+L3)	64.8 (1.2)	0.81 (0.03)	42.8 (2.0)	67.2 (1.6)	0.74 (0.03)









5. Discussion

5.1 Hierarchical Contrastive Learning and Compositional Generalisation

The 130.1% improvement in C-CME compositional generalisation (18.6% to 42.8% R@1) is the most theoretically significant finding: it demonstrates that hierarchical contrastive learning -- specifically the attribute-level alignment in HCMC Level 3 -- produces qualitatively different cross-modal representations with substantially richer compositional structure than instance-level alignment alone. The ablation shows that most of this improvement is attributable to the attribute-level contrastive loss (HCMC-L1+L2 achieves only 28.4% vs. full HCMC 42.8%), confirming that explicit attribute binding in the contrastive objective is the key mechanism rather than category-level alignment. The improvement in

standard retrieval tasks (mean +8.4pp) despite using the same architecture and training data scale as CLIP suggests that the richer representational structure learned through hierarchical contrastive alignment also benefits coarse semantic retrieval, rather than imposing a compositionality-retrieval trade-off. This finding supports the view that hierarchical contrastive learning is a strictly superior representation learning objective for cross-modal systems, not merely a specialised method for compositional tasks.

5.2 Limitations and Future Research

The HCMC evaluation at CC12M training scale (12M pairs) may not fully reflect performance at LAION- 400M or larger scales, where the category and attribute label quality from automatic parsing may degrade relative to the noise level in larger web-scraped datasets. The attribute parser used for Level 3 alignment covers five attribute types (colour, material, size, shape, action); the extension to more complex relational attributes (spatial relations, comparative attributes) and to audio attributes (timbre, pitch, rhythm) is a natural extension that would further improve compositional generalisation in audio-language cross-modal settings. Future research should evaluate HCMC representations as conditioning signals for the TGI trimodal generative architecture (Lindberg et al., 2024), testing whether the improved compositional structure of HCMC representations translates to qualitatively superior trimodal generation coherence beyond the TGQ alignment metric improvement (0.61 to 0.74) already observed. The theoretical analysis of why hierarchical contrastive learning improves compositional generalisation -- through the lens of information geometry and representation space curvature -- represents a valuable mathematical research direction.

6. Conclusion

6.1 Summary

This paper proposed the CMRQ representation quality framework and the HCMC hierarchical cross-modal contrastive learning objective. HCMC achieves mean +8.4pp R@1 over CLIP across eight cross-modal benchmarks, +24.2pp on C-CME compositional generalisation (130.1% improvement), and +0.13 TGQ generation coherence improvement. Ablation confirms attribute-level (Level 3) contrastive loss as the critical mechanism for compositional generalisation gains. HCMC uses identical architecture and training data scale as CLIP, making it a direct drop-in improvement requiring only objective modification.

6.2 Open Source Release

HCMC model weights (ViT-L/14 encoders trained on CC12M with HCMC objective), the C-CME benchmark dataset (10,000 compositional cross-modal evaluation triplets with novel attribute-object combinations), and CMRQ evaluation scripts are released as open- source resources. HCMC weights are drop-in compatible with standard CLIP inference pipelines, enabling immediate adoption as cross-modal conditioning signals for existing generative systems. We encourage the community to extend HCMC to additional modality pairs and to larger training scales to characterise the scaling behaviour of hierarchical contrastive objectives. Implementation details are in Appendix A.

References

Bahdanau, D. et al. (2019). CLOSURE: Assessing systematic generalization of CLEVR models. arXiv:1912.05783.

Elizalde, B. et al. (2023). CLAP: Learning audio concepts from natural language supervision. ICASSP 2023.

Girdhar, R. et al. (2023). ImageBind: One embedding space to bind them all. CVPR 2023.

Guo, Y. et al. (2022). HCSC: Hierarchical contrastive selective coding. CVPR 2022.

Jia, C. et al. (2021). Scaling up visual and vision-language representation learning. ICML 2021.

Lake, B. M. et al. (2017). Building machines that learn and think like people. Behavioral and Brain Sciences, 40, e253.

Lindberg, S., Hansen, A., and Bianchi, A. (2024). Multimodal generative models for text-image-audio integration. Journal of Generative Intelligence, 2024(1), 33-40.

Marcus, G. (2003). The algebraic mind. MIT Press.

Radford, A. et al. (2021). Learning transferable visual models from natural language supervision. ICML 2021.

Rombach, R. et al. (2022). High-resolution image synthesis with latent diffusion models. CVPR 2022.

Thrush, T. et al. (2022). Winoground: Probing vision and language models for visio-linguistic compositionality. CVPR 2022.

Yuksekgonul, M. et al. (2023). When and why vision-language models behave like bag-of-words models. ICLR 2023.

Garcia, L., Nowak, A., and Novak, L. (2024). Knowledge-integrated LLMs for reliable text generation. Journal of Generative Intelligence, 2024(1), 25-32.

Saharia, C. et al. (2022). Photorealistic text-to-image diffusion models. NeurIPS, 35.

Li, J. et al. (2022). BLIP: Bootstrapping language-image pre-training. ICML 2022.

Chen, X. et al. (2020). Improved baselines with momentum contrastive learning. arXiv:2003.04297.

He, K. et al. (2020). Momentum contrast for unsupervised visual representation learning. CVPR 2020.

Wang, C. et al. (2023). VALL-E: Neural codec language models. arXiv:2301.02111.

Changpinyo, S. et al. (2021). Conceptual 12M: Pushing web-scale image-text pre-training. CVPR 2021.

Popescu, L., Horvath, M., and Novak, I. (2024). Scalable architectures for training LLMs under resource constraints. Journal of Generative Intelligence, 2024(1), 1-8.

Declarations

Funding

This research was supported by the Swiss National Science Foundation (SNSF) under grant 200021-222841 (HCMC: Hierarchical Cross-Modal Contrastive Learning), the Swedish Research Council under grant 2024-05412, and the Austrian Science Fund (FWF) under grant P 42014-N. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

HCMC model weights, C-CME benchmark dataset, CMRQ evaluation scripts, and training code are publicly available. All evaluation datasets used (MSCOCO, Flickr30K, AudioCaps, Clotho, VGGSound, Winoground) are publicly available.

Ethical Approval

This study involved computational experiments on publicly available benchmark datasets only. No human subjects or personal data were involved. Ethical approval was not required.

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2024, Issues: 1, 2024

Appendix A

HCMC Training Protocol and C-CME Benchmark Construction

The following provides HCMC hyperparameters, the attribute parser specification, and C-CME benchmark construction methodology.

HCMC Training Configuration

C-CME Benchmark Construction

CMRQ Evaluation Protocol