Novelty Assessment Report

Paper: Distributional Vision-Language Alignment by Cauchy-Schwarz Divergence
PDF URL: https://openreview.net/pdf?id=UUAjF4xL0e
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

Vision-language alignment is crucial for various downstream tasks such as cross-modal generation and retrieval. Previous multimodal approaches like CLIP utilize InfoNCE to maximize mutual information, primarily aligning pairwise samples across modalities while overlooking distributional differences. In addition, InfoNCE has inherent conflict in terms of alignment and uniformity in multimodality, leading to suboptimal alignment with modality gaps. To overcome the limitations, we propose CS-Aligner, a novel framework that performs distributional vision-language alignment by integrating Cauchy-Schwarz (CS) divergence with mutual information. CS-Aligner captures both the global distribution information of each modality and the pairwise semantic relationships. We find that the CS divergence seamlessly addresses the InfoNCE's alignment-uniformity conflict and serves complementary roles with InfoNCE, yielding tighter and more precise alignment. Moreover, by introducing distributional alignment, CS-Aligner enables incorporating additional information from unpaired data and token-level representations, enhancing flexible and fine-grained alignment in practice. Experiments on text-to-image generation and cross-modality retrieval tasks demonstrate the effectiveness of our method on vision-language alignment.

Core Task Landscape

This paper addresses: **vision-language alignment across modalities**
A total of **50 papers** were analyzed and organized into a taxonomy with **19 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Alignment Objectives and Training Frameworks**
- **Multimodal Architecture and Integration**
- **Cross-Modal Representation and Semantic Analysis**
- **Domain-Specific Vision-Language Alignment**
- **Downstream Task Applications and Adaptation**
- **Survey and Review Literature**

Complete Taxonomy Tree

- vision-language alignment across modalities Survey Taxonomy
- Alignment Objectives and Training Frameworks
  - Contrastive and Distributional Alignment ★ (6 papers)
  - [0] Distributional Vision-Language Alignment by Cauchy-Schwarz Divergence (Anon et al., 2026) View paper
  - [10] Flava: A foundational language and vision alignment model (Amanpreet Singh, 2022) View paper
  - [18] Vlmixer: Unpaired vision-language pre-training via cross-modal cutmix (Wang Teng, 2022) View paper
  - [25] COTS: Collaborative two-stream vision-language pre-training model for cross-modal retrieval (Haoyu Lu, 2022) View paper
  - [35] VISTA: Enhancing Vision-Text Alignment in MLLMs via Cross-Modal Mutual Information Maximization (LI Mingxiao, 2025) View paper
  - [41] Pyramidclip: Hierarchical feature alignment for vision-language model pretraining (Gao Yuting, 2022) View paper
  - Multi-Granularity and Hierarchical Alignment (4 papers)
  - [1] Multi-granularity cross-modal alignment for generalized medical visual representation learning (Wang Fuying, 2022) View paper
  - [6] Multi-modal alignment using representation codebook (Jiali Duan, 2022) View paper
  - [24] Beyond general alignment: Fine-grained entity-centric image-text matching with multimodal attentive experts (Yaxiong Wang, 2025) View paper
  - [43] Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders (Amato, 2021) View paper
  - Preference Optimization and Post-Training Alignment (3 papers)
  - [12] Aligning modalities in vision large language models via preference fine-tuning (Zhou, 2024) View paper
  - [26] Visual Representation Alignment for Multimodal Large Language Models (Jung Jaewoo, 2025) View paper
  - [37] Post-pre-training for modality alignment in vision-language foundation models (Yamaguchi, 2025) View paper
- Multimodal Architecture and Integration
  - Cross-Modal Attention and Fusion (5 papers)
  - [23] Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis (Luwei Xiao, 2023) View paper
  - [38] Unsupervised and pseudo-supervised vision-language alignment in visual dialog (Fei-Long Chen, 2022) View paper
  - [45] Multimodal Prompt-Guided Bidirectional Fusion for Referring Remote Sensing Image Segmentation (Yingjie Li, 2025) View paper
  - [47] mplug: Effective and efficient vision-language learning by cross-modal skip-connections (Chenliang Li, 2022) View paper

- [49] DELAN: Dual-Level Alignment for Vision-and-Language Navigation by Cross-Modal Contrastive Learning (Mengfei Du, 2024) View paper
- Unified and Foundational Multimodal Models (4 papers)
- [7] Onellm: One framework to align all modalities with language (Jiaming Han, 2024) View paper
- [11] Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment (Zhu Bin, 2023) View paper
- [30] Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks (Zhe Chen, 2024) View paper
- [33] X-vila: Cross-modality alignment for large language model (Ye, 2024) View paper
- Vision-Language Model Projectors and Connectors (3 papers)
- [17] Alignvlm: Bridging vision and language latent spaces for multimodal understanding (Masry, 2025) View paper
- [28] Align-KD: Distilling Cross-Modal Alignment Knowledge for Mobile Vision-Language Large Model Enhancement (Qianhan Feng, 2025) View paper
- [32] Kernel-based unsupervised embedding alignment for enhanced visual representation in vision-language models (Gong, 2025) View paper

- Cross-Modal Representation and Semantic Analysis
- Modality Gap and Representation Space Analysis (4 papers)
- [16] Deciphering Cross-Modal Alignment in Large Vision-Language Models via Modality Integration Rate (Q Huang, 2025) View paper
- [31] Unified modality separation: A vision-language framework for unsupervised domain adaptation (Li Xinyao, 2025) View paper
- [34] Assessing and Learning Alignment of Unimodal Vision and Language Models (Le Zhang, 2025) View paper
- [44] Mind the Modality Gap: Towards a Remote Sensing Vision-Language Model via Cross-modal Alignment (Angelos Zavras, 2024) View paper
- Cross-Modal Task Representations and Transfer (2 papers)
- [15] Vision-Language Models Create Cross-Modal Task Representations (Luo, 2024) View paper
- [50] The in-context inductive biases of vision-language models differ across modalities (Allen, 2025) View paper
- Multimodal Integration in Neural Systems (1 papers)
- [19] Revealing vision-language integration in the brain with multimodal networks (Subramaniam Vighnesh, 2024) View paper

- Domain-Specific Vision-Language Alignment
- Medical and Scientific Imaging Alignment (2 papers)
- [21] LVLM-HBA: Large Vision-Language Model with Cross-Modal Alignment for Human Behavior Analysis (Jun Yu, 2025) View paper
- [46] Enhancing visual and semantic alignment of multimodal large models in medical images: Q. Liu et al. (Q Liu, 2025) View paper
- Remote Sensing and Geospatial Vision-Language (1 papers)
- [29] VLCA: vision-language aligning model with cross-modal attention for bilingual remote sensing image captioning (Tingting Wei, 2023) View paper
- Embodied and Interactive Vision-Language Tasks (1 papers)
- [42] Improving cross-modal alignment in vision language navigation via syntactic information (Bansal, 2021) View paper
- Multimodal Sensing Beyond Vision-Language (1 papers)
- [4] A touch, vision, and language dataset for multimodal alignment (Fu, 2024) View paper

- Downstream Task Applications and Adaptation
- Cross-Modal Retrieval and Matching (2 papers)
- [3] Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval (Ding Jiang, 2023) View paper
- [22] Vision-Language Models Struggle to Align Entities across Modalities (Alonso, 2025) View paper
- Multimodal Generation and Synthesis (2 papers)
- [2] Generating images with multimodal language models (Koh, 2023) View paper
- [27] Textmatch: Enhancing image-text consistency through multimodal optimization (Luo Yu-cong, 2024) View paper
- Referring Segmentation and Grounding (1 papers)
- [48] ReferSAM: Unleashing Segment Anything Model for Referring Image Segmentation (Sun-Ao Liu, 2025) View paper
- Multimodal Recommendation and Sentiment Analysis (1 papers)
- [20] Alignrec: Aligning and training in multimodal recommendations (Yifan Liu, 2024) View paper
- Knowledge Graph Construction and Reasoning (2 papers)
- [39] Aligning vision to language: Annotation-free multimodal knowledge graph construction for enhanced llms reasoning (Liu Jun-ming, 2025) View paper
- [40] Aligning vision to language: Text-free multimodal knowledge graph construction for enhanced llms reasoning (Liu Jun-ming, 2025) View paper

- Survey and Review Literature (6 papers)
- [5] Large vision-language model alignment and misalignment: A survey through the lens of explainability (Dong Shu, 2025) View paper
- [8] From large language models to large multimodal models: A literature review (Dawei Huang, 2024) View paper
- [9] A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges (Zongxia Li, 2025) View paper
- [13] Multimodal fusion and vision-language models: A survey for robot vision (Xiaofeng Han, 2025) View paper
- [14] Exploring the frontier of vision-language models: A survey of current methodologies and future directions (Ghosh, 2024) View paper
- [36] Multimodal large language models: A survey (Joshi, 2023) View paper

Narrative

Core task: vision-language alignment across modalities. The field centers on learning joint representations that bridge visual and textual information, enabling models to understand and reason about images and language together. The taxonomy reveals several major branches: Alignment Objectives and Training Frameworks explore how to define and optimize cross-modal correspondence through contrastive, distributional, or other learning signals; Multimodal Architecture and Integration addresses the design of encoders, fusion mechanisms, and unified models that process multiple modalities; Cross-Modal Representation and Semantic Analysis investigates how semantic structures and fine-grained correspondences emerge in shared embedding spaces; Domain-Specific Vision-Language Alignment tailors methods to specialized contexts such as medical imaging or robotics; and Downstream Task Applications and Adaptation examines how aligned representations transfer to retrieval, captioning, and reasoning tasks. Representative works like FLAVA[10] and

LanguageBind[11] illustrate foundational architectures, while surveys such as Alignment Misalignment Survey[5] and Vision Language Survey[9] synthesize progress across these dimensions.

Within the Alignment Objectives and Training Frameworks branch, a particularly active line focuses on contrastive and distributional alignment strategies. Many studies refine how similarity metrics and loss functions shape the learned embedding geometry, balancing global alignment with fine-grained semantic structure. The Cauchy-Schwarz Divergence[0] paper situates itself in this cluster, proposing an alternative divergence measure for aligning distributions across modalities. It shares thematic ground with works like VLMixer[18] and COTS[25], which also emphasize training objectives that go beyond standard contrastive losses, yet differs in its specific mathematical formulation and focus on distributional properties. Nearby efforts such as VISTA[35] explore hierarchical or multi-scale alignment, highlighting ongoing questions about how to capture both coarse semantic agreement and detailed correspondences. These contrasting emphases reflect broader trade-offs in the field: whether to prioritize scalability and simplicity or to incorporate richer structural priors into the alignment process.

Related Works in Same Category

The following **5 sibling papers** share the same taxonomy leaf node with the original paper:

1. Flava: A foundational language and vision alignment model

Authors: Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, et al. (7 authors total) | **Year/Venue**: 2022 | **URL**: View paper

Abstract

State-of-the-art vision and vision-and-language models rely on large-scale visio-linguistic pretraining for obtaining good performance on a variety of downstream tasks. Generally, such models are often either cross-modal (contrastive) or multi-modal (with earlier fusion) but not both; and they often only target specific modalities or tasks. A promising direction would be to use a single holistic universal model, as a â□□foundationâ□□, that targets all modalities at once-a true vision and languag...

Relationship Analysis

Both papers belong to the Contrastive and Distributional Alignment category, employing contrastive learning methods for vision-language alignment. FLAVA shares the original paper's focus on aligning vision and language modalities through contrastive objectives (global contrastive loss similar to CLIP's InfoNCE), and both address distributional aspects of multimodal alignment. However, FLAVA differs by using a three-encoder architecture (separate image, text, and multimodal encoders) with multiple pretraining objectives (contrastive, masked multimodal modeling, image-text matching), while the original paper specifically introduces Cauchy-Schwarz divergence to explicitly minimize distributional gaps and resolve the alignment-uniformity conflict inherent in InfoNCE.

---

2. Vlmixer: Unpaired vision-language pre-training via cross-modal cutmix

Authors: Wang Teng, Jiang, Wenhao, Teng Wang, Lu, et al. (18 authors total) | **Year/Venue**: 2022 | **URL**: View paper

Abstract

Existing vision-language pre-training (VLP) methods primarily rely on paired image-text datasets, which are either annotated by enormous human labors, or crawled from the internet followed by elaborate data cleaning techniques. To reduce the dependency on well-aligned image-text pairs, it is promising to directly leverage the large-scale text-only and image-only corpora. This paper proposes a data augmentation method, namely cross-modal CutMix (CMC), for implicit cross-modal alignment learning i...

Relationship Analysis

Both papers belong to the Contrastive and Distributional Alignment category, focusing on vision-language alignment through contrastive learning mechanisms. They overlap in addressing the modality gap problem and utilizing contrastive objectives for cross-modal alignment, with both methods incorporating distributional considerations. However, the original paper (CS-Aligner) explicitly introduces Cauchy-Schwarz divergence as a distributional alignment measure to complement InfoNCE and resolve its alignment-uniformity conflict, while VLMixer focuses on cross-modal data augmentation (Cross-Modal CutMix) to create multi-modal representations from unpaired data, using contrastive learning between original and augmented views rather than explicit distributional divergence measures.

---

3. COTS: Collaborative two-stream vision-language pre-training model for cross-modal retrieval

Authors: Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, et al. (7 authors total) | **Year/Venue**: 2022 | **URL**: View paper

Abstract

Large-scale single-stream pre-training has shown dramatic performance in image-text retrieval. Regrettably, it faces low inference efficiency due to heavy attention layers. Recently, two-stream methods like CLIP and ALIGN with high inference efficiency have also shown promising performance, however, they only consider instance-level alignment between the two streams (thus there is still room for improvement). To overcome these limitations, we propose a novel COllaborative Two-Stream vision-langu...

Relationship Analysis

Both papers belong to the Contrastive and Distributional Alignment category, employing contrastive learning methods for vision-language alignment. COTS focuses on enhancing two-stream architectures through multi-level interactions (instance, token, and task-level) using momentum contrastive learning and masked vision-language modeling, while the original paper addresses distributional alignment by integrating Cauchy-Schwarz divergence with InfoNCE to resolve alignment-uniformity conflicts. The key difference is that COTS emphasizes architectural efficiency and multi-level interaction design within the contrastive framework, whereas the original paper introduces a novel divergence metric to explicitly align modality distributions and address theoretical limitations of InfoNCE.

---

4. VISTA: Enhancing Vision-Text Alignment in MLLMs via Cross-Modal Mutual Information Maximization

Authors: LI Mingxiao, Su Na, Mingxiao Li, Qu Fang, Na Su, et al. (17 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Current multimodal large language models (MLLMs) face a critical challenge in modality alignment, often exhibiting a bias towards textual information at the expense of other modalities like vision. This paper conducts a systematic information-theoretic analysis of the widely used cross-entropy loss in MLLMs, uncovering its implicit alignment objective. Our theoretical investigation reveals that this implicit objective has inherent limitations, leading to a degradation of cross-modal alignment as...

Relationship Analysis

Both papers belong to the Contrastive and Distributional Alignment category, focusing on improving vision-language alignment through distributional measures and mutual information maximization. They overlap in addressing the limitations of InfoNCE/contrastive learning for multimodal alignment and both propose explicit alignment objectives to bridge modality gaps. The key difference is that the original paper introduces Cauchy-Schwarz divergence as a novel distributional alignment metric with parameter-efficient adapters,

while the candidate paper (VISTA) focuses on theoretical analysis of cross-entropy's implicit alignment degradation and proposes an explicit mutual information maximization loss directly integrated into MLLM training without additional modules.

---

5. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining

Authors: Gao Yuting, Yuting Gao, Liu, Jinfeng, Jinfeng Liu, et al. (16 authors total) | **Year/Venue**: 2022 | **URL**: View paper

Abstract

Large-scale vision-language pre-training has achieved promising results on downstream tasks. Existing methods highly rely on the assumption that the image-text pairs crawled from the Internet are in perfect one-to-one correspondence. However, in real scenarios, this assumption can be difficult to hold: the text description, obtained by crawling the affiliated metadata of the image, often suffers from the semantic mismatch and the mutual compatibility. To address these issues, we introduce Pyrami...

Relationship Analysis

Both papers belong to the Contrastive and Distributional Alignment category, employing contrastive learning methods for vision-language alignment. They overlap in addressing modality gap issues and using contrastive objectives (InfoNCE) for alignment, but differ fundamentally in their approaches: the original paper introduces Cauchy-Schwarz divergence for distributional alignment to explicitly minimize distribution distance between modalities, while PyramidCLIP constructs hierarchical input pyramids with multi-level semantic alignment (global, local, and object-level) and uses label smoothing to handle negative samples. The original paper focuses on distribution-level alignment theory and unpaired data flexibility, whereas PyramidCLIP emphasizes hierarchical semantic matching across different granularities.

Contributions Analysis

Overall novelty summary. The paper proposes CS-Aligner, a framework integrating Cauchy-Schwarz divergence with mutual information for distributional vision-language alignment. It resides in the 'Contrastive and Distributional Alignment' leaf under 'Alignment Objectives and Training Frameworks', alongside five sibling papers. This leaf represents a moderately populated research direction within the broader taxonomy of 50 papers across approximately 36 topics, indicating active but not overcrowded exploration of contrastive and distributional training objectives for cross-modal alignment.

The taxonomy tree reveals that CS-Aligner's leaf sits within a parent branch focused on alignment objectives and training paradigms, distinct from architectural integration (e.g., cross-modal attention mechanisms) and downstream applications (e.g., retrieval or generation tasks). Neighboring leaves include 'Multi-Granularity and Hierarchical Alignment' and 'Preference Optimization and Post-Training Alignment', which address complementary aspects of training but differ in scope: the former targets multi-level semantic correspondence, while the latter refines models after initial pre-training. CS-Aligner's focus on distributional divergence measures positions it at the intersection of contrastive learning refinement and global distribution matching.

Among 22 candidates examined, the contribution-level analysis shows varied novelty profiles. The core CS-Aligner framework (3 candidates examined, 0 refutable) and the InfoNCE alignment-uniformity conflict analysis (10 candidates examined, 0 refutable) appear relatively novel within the limited search scope. However, the extension to unpaired data and token-level alignment (9 candidates examined, 1 refutable) encounters at least one prior work with overlapping ideas. These statistics reflect a targeted semantic search, not an exhaustive survey, suggesting that while the core divergence-based approach may be distinctive, certain practical extensions have precedent in the examined literature.

Given the limited search scope of 22 candidates, the analysis suggests moderate novelty for the core distributional alignment framework and theoretical conflict analysis, with some overlap in the unpaired data extension. The taxonomy context indicates the paper contributes to an active but not saturated research direction, where refinements to contrastive objectives remain an open question. A broader literature search might reveal additional related work, particularly in distributional alignment methods or token-level correspondence techniques.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: CS-Aligner framework for distributional vision-language alignment

Description: The authors introduce CS-Aligner, a framework that combines Cauchy-Schwarz divergence with mutual information to align vision and language representations at both distributional and sample-wise levels, addressing the modality gap problem in existing methods like CLIP.

This contribution was assessed against **3 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Contrastive Alignment with Semantic Gap-Aware Corrections in Text-Video Retrieval

URL: View paper

Brief Assessment

Semantic Gap-Aware[60] focuses on text-video retrieval using pair-specific increments to address optimization tension in contrastive learning, not on distributional alignment using Cauchy-Schwarz divergence and mutual information as in the original paper.

---

2. New divergence measures and their application in multimodal image registration

URL: View paper

Brief Assessment

Divergence Measures Registration[61] applies Cauchy-Schwarz divergence to multimodal medical image registration, not vision-language alignment. The candidate focuses on aligning medical images from different modalities (e.g., MRI, CT scans) for spatial registration, while the original paper addresses semantic alignment between vision and language representations in a shared embedding space for tasks like text-to-image generation and retrieval.

---

3. Supplementary Materialâ□□On the Effects of Self-supervision and Contrastive Alignment in Deep Multi-view Clustering

URL: View paper

Brief Assessment

Deep Multi-view Clustering[62] focuses on multi-view clustering using Cauchy-Schwarz divergence for clustering tasks, not vision-language alignment. The candidate applies CS divergence to cluster representations within multi-view data, while the original paper uses it to align vision and language modality distributions.

Contribution 2: Analysis of InfoNCE alignment-uniformity conflict in multimodality

Description: The authors analyze and demonstrate that InfoNCE loss contains inherent conflicts between alignment and uniformity terms in multimodal settings, and show that CS divergence resolves this conflict while remaining compatible with InfoNCE through kernel density estimation.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Contrastive Alignment with Semantic Gap-Aware Corrections in Text-Video Retrieval

URL: View paper

Brief Assessment

Semantic Gap-Aware[60] analyzes gradient tension in InfoNCE from a modality gap perspective but does not demonstrate the inherent conflict between alignment and uniformity terms or propose CS divergence as a resolution.

---

2. Model-Aware Contrastive Learning: Towards Escaping the Dilemmas

URL: View paper

Brief Assessment

Model-Aware Contrastive[68] addresses InfoNCE issues in unimodal contrastive learning (uniformity-tolerance dilemma, gradient reduction), not the multimodal alignment-uniformity conflict analyzed in the original paper. The candidate focuses on temperature adaptation and negative sampling in general contrastive learning, while the original specifically analyzes how InfoNCE's uniformity term conflicts with alignment across different modalities (vision-language).

---

3. f-MICL: Understanding and Generalizing InfoNCE-based Contrastive Learning

URL: View paper

Brief Assessment

f-MICL[70] analyzes InfoNCE's alignment-uniformity conflict in general contrastive learning settings, not specifically in multimodal contexts. The original paper focuses on the unique challenges of multimodal alignment where alignment and uniformity terms conflict across different modalities (vision-language), while f-MICL[70] addresses uniformity within single-modality feature spaces.

---

4. Semantic item graph enhancement for multimodal recommendation

URL: View paper

Brief Assessment

Semantic Item Graph[63] focuses on multimodal recommendation systems using item-item semantic graphs and applies InfoNCE for representation alignment. It does not analyze or demonstrate inherent conflicts between alignment and uniformity terms in InfoNCE for multimodal settings, nor does it propose CS divergence as a resolution mechanism.

---

5. Multi-Level Contrastive Learning for Multimodal Sentiment Analysis

URL: View paper

Brief Assessment

Multi-Level Contrastive[65] focuses on multimodal sentiment analysis and does not analyze InfoNCE's alignment-uniformity conflict. The candidate mentions replacing InfoNCE with SupCon but provides no theoretical analysis of conflicts between alignment and uniformity terms in multimodal settings.

---

6. Enhancing Recommendation Representations Through Alignment and Uniformity with Integrated Contrastive Learning and Collaborative Filtering

URL: View paper

Brief Assessment

Alignment Uniformity Recommendation[69] focuses on recommendation systems using collaborative filtering, not multimodal vision-language alignment. The paper analyzes alignment-uniformity in the context of user-item interactions, not the multimodal distributional conflicts that the original paper addresses.

---

7. Improving Contrastive Learning of Sentence Embeddings with Focal InfoNCE

URL: View paper

Brief Assessment

Focal InfoNCE[71] focuses on sentence embeddings in unimodal settings (text-only), addressing hard negative mining through self-paced re-weighting. The original paper analyzes InfoNCE's alignment-uniformity conflict specifically in multimodal (vision-language) settings, which is a fundamentally different context not addressed by Focal InfoNCE[71].

---

8. A Principled Framework for Multi-View Contrastive Learning

URL: View paper

Brief Assessment

Multi-View Contrastive Framework[67] addresses alignment-uniformity coupling in multi-view unimodal settings (multiple augmented views of the same modality), not multimodal settings (text-image pairs). The candidate focuses on extending contrastive learning to handle multiple views of single-modality data, whereas the original paper analyzes conflicts specifically arising in cross-modal alignment between vision and language distributions.

---

9. Open-set Cross Modal Generalization via Multimodal Unified Representation

URL: View paper

Brief Assessment

Unified Representation[64] focuses on open-set cross-modal generalization using masked InfoNCE and jigsaw puzzles for handling unseen classes. It does not analyze or address the alignment-uniformity conflict inherent in InfoNCE loss for multimodal settings.

---

10. The Role of Local Alignment and Uniformity in Image-Text Contrastive Learning on Medical Images

URL: View paper

Brief Assessment

Local Alignment Uniformity[66] focuses on decomposing global vs. local contrastive losses for medical image-text tasks, not on the inherent conflict between alignment and uniformity terms in InfoNCE for general multimodal settings. The candidate examines how local and global losses relate within a specific medical imaging framework (LoVT), while the original analyzes InfoNCE's structural conflict across modalities and proposes CS divergence as a resolution.

Contribution 3: Extension to unpaired data and token-level alignment

Description: The authors extend their framework to leverage unpaired multimodal data (including multiple captions per image and independently sampled data) and introduce token-level alignment between vision and language tokens for more fine-grained multimodal correspondence.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Deep boosting learning: A brand-new cooperative approach for image-text matching

URL: View paper

Brief Assessment

Deep Boosting Learning[53] focuses on image-text matching through cooperative peer-branch training with adaptive margin constraints, not on unpaired multimodal data or token-level alignment strategies.

---

2. SPSD: Similarity-preserving self-distillation for videoâ□□text retrieval

URL: View paper

Brief Assessment

SPSD[55] focuses on self-distillation for video-text retrieval with similarity preservation. The candidate's limited context does not demonstrate prior work on distributional alignment with unpaired multimodal data or token-level CS divergence-based alignment as proposed in the original paper.

---

3. Vlmixer: Unpaired vision-language pre-training via cross-modal cutmix

URL: View paper

Prior Art Analysis

VLMixer[18] demonstrates prior work on both unpaired multimodal data and token-level alignment. The candidate paper explicitly addresses unpaired vision-language pre-training, including handling multiple captions per image and independently sampled data. Additionally, VLMixer[18] proposes token-level alignment between vision and language tokens through their cross-modal cutmix approach, which aligns all tokens rather than just CLS tokens for fine-grained alignment.

Evidence

Evidence 1 - **Rationale**: VLMixer[18] explicitly addresses unpaired vision-language pre-training by leveraging text-only and image-only corpora, demonstrating prior work on unpaired data alignment before the original paper. - **Original**: Moreover, by introducing distributional alignment, cs-aligner enables incorporating additional information from unpaired data and token-level representations, enhancing flexible and fine-grained alignment in practice. - **Candidate**: to reduce the dependency on well-aligned imagetext pairs, it is promising to directly leverage the large-scale text-only and image-only corpora. this paper proposes a data augmentation method, namely cross-modal cutmix (cmc), for implicit cross-modal alignment learning in unpaired vlp.

---

4. Self-supervised multimodal learning: A survey

URL: View paper

Brief Assessment

Self-supervised Multimodal Survey[51] is a survey paper that reviews existing methods in multimodal learning. While it discusses unpaired data and fine-grained alignment techniques in sections 5.1.2 and 5.2, it does not present these as novel contributions but rather surveys existing approaches in the field. The original paper proposes a specific framework (CS-Aligner) with novel extensions for unpaired data and token-level alignment, which is distinct from a survey's role of reviewing prior work.

---

5. InfoMAE: Pair-Efficient Cross-Modal Alignment for Multimodal Time-Series Sensing Signals

URL: View paper

Brief Assessment

InfoMAE[56] focuses on cross-modal alignment for time-series sensing signals (e.g., seismic, acoustic, IMU sensors) in IoT applications, not vision-language alignment. The candidate addresses distribution-level alignment of multimodal time-series data with limited synchronized pairs, which is a fundamentally different domain and technical approach from the original paper's vision-language token-level alignment and unpaired image-text data.

---

6. Image-Text Retrieval via Green Explainable Multi-modal Alignment (GEMMA)

URL: View paper

Brief Assessment

GEMMA[57] focuses on stage-wise alignment using frozen encoders with clustering and feature selection modules, not on distributional alignment with unpaired data or token-level CS divergence as in the original paper.

---

7. Cross-modal Full-mode Fine-grained Alignment for Text-to-Image Person Retrieval

URL: View paper

Brief Assessment

Full-mode Fine-grained[54] focuses on text-to-image person retrieval with explicit fine-grained alignment between text and image features, not on unpaired multimodal data or token-level alignment in the distributional sense proposed by the original paper.

---

8. Fine-grained text-based person re-identification via interlaced cross-attention and LoRA fine-tuning: M. Hu et al.

URL: View paper

Brief Assessment

Interlaced Cross-attention LoRA[52] focuses on person re-identification using cross-attention mechanisms for text-image matching, not on distributional alignment frameworks or unpaired multimodal data leveraging as in the original paper.

---

9. L-MCAT: Unpaired Multimodal Transformer with Contrastive Attention for Label-Efficient Satellite Image Classification

URL: View paper

Brief Assessment

L-MCAT[58] focuses on unpaired satellite image modalities (SAR-optical) for remote sensing classification, not general vision-language alignment. The token-level alignment in L-MCAT[58] operates on spatial patch tokens within satellite images, whereas the original paper aligns vision and language tokens across different modalities (text and images) for tasks like text-to-image generation and retrieval.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Distributional Vision-Language Alignment by Cauchy-Schwarz Divergence View paper
- [1] Multi-granularity cross-modal alignment for generalized medical visual representation learning View paper
- [2] Generating images with multimodal language models View paper
- [3] Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval View paper
- [4] A touch, vision, and language dataset for multimodal alignment View paper
- [5] Large vision-language model alignment and misalignment: A survey through the lens of explainability View paper
- [6] Multi-modal alignment using representation codebook View paper
- [7] Onellm: One framework to align all modalities with language View paper
- [8] From large language models to large multimodal models: A literature review View paper
- [9] A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges View paper
- [10] Flava: A foundational language and vision alignment model View paper
- [11] Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment View paper
- [12] Aligning modalities in vision large language models via preference fine-tuning View paper
- [13] Multimodal fusion and vision-language models: A survey for robot vision View paper
- [14] Exploring the frontier of vision-language models: A survey of current methodologies and future directions View paper
- [15] Vision-Language Models Create Cross-Modal Task Representations View paper
- [16] Deciphering Cross-Modal Alignment in Large Vision-Language Models via Modality Integration Rate View paper
- [17] Alignvlm: Bridging vision and language latent spaces for multimodal understanding View paper
- [18] Vlmixer: Unpaired vision-language pre-training via cross-modal cutmix View paper
- [19] Revealing vision-language integration in the brain with multimodal networks View paper
- [20] Alignrec: Aligning and training in multimodal recommendations View paper
- [21] LVLM-HBA: Large Vision-Language Model with Cross-Modal Alignment for Human Behavior Analysis View paper
- [22] Vision-Language Models Struggle to Align Entities across Modalities View paper
- [23] Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis View paper
- [24] Beyond general alignment: Fine-grained entity-centric image-text matching with multimodal attentive experts View paper
- [25] COTS: Collaborative two-stream vision-language pre-training model for cross-modal retrieval View paper
- [26] Visual Representation Alignment for Multimodal Large Language Models View paper
- [27] Textmatch: Enhancing image-text consistency through multimodal optimization View paper
- [28] Align-KD: Distilling Cross-Modal Alignment Knowledge for Mobile Vision-Language Large Model Enhancement View paper
- [29] VLCA: vision-language aligning model with cross-modal attention for bilingual remote sensing image captioning View paper
- [30] Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks View paper
- [31] Unified modality separation: A vision-language framework for unsupervised domain adaptation View paper
- [32] Kernel-based unsupervised embedding alignment for enhanced visual representation in vision-language models View paper
- [33] X-vila: Cross-modality alignment for large language model View paper
- [34] Assessing and Learning Alignment of Unimodal Vision and Language Models View paper
- [35] VISTA: Enhancing Vision-Text Alignment in MLLMs via Cross-Modal Mutual Information Maximization View paper
- [36] Multimodal large language models: A survey View paper
- [37] Post-pre-training for modality alignment in vision-language foundation models View paper
- [38] Unsupervised and pseudo-supervised vision-language alignment in visual dialog View paper
- [39] Aligning vision to language: Annotation-free multimodal knowledge graph construction for enhanced llms reasoning View paper
- [40] Aligning vision to language: Text-free multimodal knowledge graph construction for enhanced llms reasoning View paper
- [41] Pyramidclip: Hierarchical feature alignment for vision-language model pretraining View paper
- [42] Improving cross-modal alignment in vision language navigation via syntactic information View paper
- [43] Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders View paper
- [44] Mind the Modality Gap: Towards a Remote Sensing Vision-Language Model via Cross-modal Alignment View paper
- [45] Multimodal Prompt-Guided Bidirectional Fusion for Referring Remote Sensing Image Segmentation View paper
- [46] Enhancing visual and semantic alignment of multimodal large models in medical images: Q. Liu et al. View paper
- [47] mplug: Effective and efficient vision-language learning by cross-modal skip-connections View paper
- [48] ReferSAM: Unleashing Segment Anything Model for Referring Image Segmentation View paper
- [49] DELAN: Dual-Level Alignment for Vision-and-Language Navigation by Cross-Modal Contrastive Learning View paper
- [50] The in-context inductive biases of vision-language models differ across modalities View paper
- [51] Self-supervised multimodal learning: A survey View paper
- [52] Fine-grained text-based person re-identification via interlaced cross-attention and LoRA fine-tuning: M. Hu et al. View paper
- [53] Deep boosting learning: A brand-new cooperative approach for image-text matching View paper
- [54] Cross-modal Full-mode Fine-grained Alignment for Text-to-Image Person Retrieval View paper
- [55] SPSD: Similarity-preserving self-distillation for videoâ□□text retrieval View paper
- [56] InfoMAE: Pair-Efficient Cross-Modal Alignment for Multimodal Time-Series Sensing Signals View paper
- [57] Image-Text Retrieval via Green Explainable Multi-modal Alignment (GEMMA) View paper
- [58] L-MCAT: Unpaired Multimodal Transformer with Contrastive Attention for Label-Efficient Satellite Image Classification View paper
- [59] SoftCLIP: Softer Cross-modal Alignment Makes CLIP Stronger View paper
- [60] Contrastive Alignment with Semantic Gap-Aware Corrections in Text-Video Retrieval View paper
- [61] New divergence measures and their application in multimodal image registration View paper
- [62] Supplementary Materialâ□□On the Effects of Self-supervision and Contrastive Alignment in Deep Multi-view Clustering View paper
- [63] Semantic item graph enhancement for multimodal recommendation View paper
- [64] Open-set Cross Modal Generalization via Multimodal Unified Representation View paper
- [65] Multi-Level Contrastive Learning for Multimodal Sentiment Analysis View paper

- [66] The Role of Local Alignment and Uniformity in Image-Text Contrastive Learning on Medical Images View paper
- [67] A Principled Framework for Multi-View Contrastive Learning View paper
- [68] Model-Aware Contrastive Learning: Towards Escaping the Dilemmas View paper
- [69] Enhancing Recommendation Representations Through Alignment and Uniformity with Integrated Contrastive Learning and Collaborative Filtering View paper
- [70] f-MICL: Understanding and Generalizing InfoNCE-based Contrastive Learning View paper
- [71] Improving Contrastive Learning of Sentence Embeddings with Focal InfoNCE View paper