Novelty Assessment Report

Paper: Multimodal Aligned Semantic Knowledge for Unpaired Image-text Matching
PDF URL: https://openreview.net/pdf?id=d3CISVVO6v
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

While existing approaches address unpaired image-text matching by constructing cross-modal aligned knowledge, they often fail to identify semantically corresponding visual representations for Out-of-Distribution (OOD) words. Moreover, the distributional variance of visual representations associated with different words varies significantly, which negatively impacts matching accuracy. To address these issues, we propose a novel method namely Multimodal Aligned Semantic Knowledge (MASK), which leverages word embeddings as bridges to associate words with their corresponding prototypes, thereby enabling semantic knowledge alignment between the image and text modalities. For OOD words, the representative prototypes are constructed by leveraging the semantic relationships encoded in word embeddings. Beyond that, we introduce a prototype consistency contrastive loss to structurally regularize the feature space, effectively mitigating the adverse effects of variance. Experimental results on the Flickr30K and MSCOCO datasets demonstrate that MASK achieves superior performance in unpaired matching.

Core Task Landscape

This paper addresses: **Unpaired Image-Text Matching**

A total of **50 papers** were analyzed and organized into a taxonomy with **17 categories**.

Taxonomy Overview

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

- **Contrastive Learning and Alignment Frameworks**
- **Generation-Based Unpaired Matching**
- **Semantic Knowledge and Prototype-Based Methods**
- **Hashing and Efficient Retrieval**
- **Domain-Specific and Application-Oriented Methods**
- **Transfer Learning and Adaptation**
- **Specialized Architectures and Auxiliary Tasks**

Complete Taxonomy Tree

- Unpaired Image-Text Matching Survey Taxonomy
- Contrastive Learning and Alignment Frameworks
  - General Vision-Language Contrastive Learning (6 papers)
  - [6] MedCLIP: Contrastive Learning from Unpaired Medical Images and Text (Wang Zifeng, 2022) View paper
  - [11] Quality-Aware Image-Text Alignment for Opinion-Unaware Image Quality Assessment (Agnolucci, 2024) View paper
  - [13] The "Law" of the Unconscious Contrastive Learner: Probabilistic Alignment of Unpaired Modalities (Yunlong Che, 2025) View paper
  - [19] PyramidCLIP: Hierarchical Feature Alignment for Vision-language Model Pretraining (Gao Yuting, 2022) View paper
  - [40] Leveraging unpaired data for vision-language generative models via cycle consistency (Li Tianhong, 2023) View paper
  - [42] S-CLIP: Semi-supervised Vision-Language Learning using Few Specialist Captions (Mo, 2023) View paper
  - Medical Domain Contrastive Learning (4 papers)
  - [21] Backdoor Attack on Un-paired Medical Image-Text Pretrained Models: A Pilot Study on MedCLIP (R Jin, 2024) View paper
  - [27] Relaxing Binary Constraints in Contrastive Vision-Language Medical Representation Learning (Xiaoyang Wei, 2025) View paper
  - [30] Can Language-Guided Unsupervised Adaptation Improve Medical Image Classification Using Unpaired Images and Texts? (Umaima Rahman, 2025) View paper
  - [31] Self-supervised image-text pre-training with mixed data in chest x-rays (Wang Xiao-song, 2021) View paper
  - Multimodal and Cross-Lingual Alignment (4 papers)
  - [3] UniAlign: Scaling Multimodal Alignment within One Unified Model (Bo Zhou, 2025) View paper
  - [22] Better Together: Leveraging Unpaired Multimodal Data for Stronger Unimodal Models (Gupta, 2025) View paper
  - [25] CL2CM: Improving Cross-Lingual Cross-Modal Retrieval via Cross-Lingual Knowledge Transfer (Wang Ya-Bing, 2023) View paper
  - [37] Multimodal LLM Enhanced Cross-lingual Cross-modal Retrieval (Yabing Wang, 2024) View paper
- Generation-Based Unpaired Matching
  - Pseudo-Pair Generation for Captioning (4 papers)
  - [7] Improving cross-modal alignment with synthetic pairs for text-only image captioning (Liu Zhi-yue, 2024) View paper
  - [8] Prompt-based learning for unpaired image captioning (Peipei Zhu, 2023) View paper
  - [12] Unpaired image captioning via scene graph alignments (Jiuxiang Gu, 2019) View paper
  - [23] Unpaired image captioning with semantic-constrained self-learning (Huixia Ben, 2021) View paper

- Cross-Modal Retrieval with Pseudo-Pairs (3 papers)
- [20] Text-based person search without parallel image-text data (Yang Bai, 2023) View paper
- [26] Cross-modal generation and alignment via attribute-guided prompt for unsupervised text-based person retrieval (Zhihao Gong, 2024) View paper
- [28] From coarse to fine: a two-stage common semantic space construction for unpaired cross modal retrieval (Zhanyang Liang, 2025) View paper
- Diffusion and Generative Modeling (4 papers)
- [5] Categorical SchrÃ¶dinger Bridge Matching (Korotin, 2025) View paper
- [9] PathDiff: Histopathology Image Synthesis with Unpaired Text and Mask Conditions (Bhosale, 2025) View paper
- [17] Generating Multimodal Images with GAN: Integrating Text, Image, and Style (Chaoyi Tan, 2025) View paper
- [29] Doracycle: Domain-oriented adaptation of unified generative model in multimodal cycles (Rui Zhao, 2025) View paper

- Semantic Knowledge and Prototype-Based Methods
  - Prototype and Conceptual Knowledge Alignment ★ (3 papers)
  - [0] Multimodal Aligned Semantic Knowledge for Unpaired Image-text Matching (Anon et al., 2026) View paper
  - [14] MACK: Multimodal Aligned Conceptual Knowledge for Unpaired Image-text Matching (Yan Huang, 2022) View paper
  - [47] Unpaired Image-Text Matching via Multimodal Aligned Conceptual Knowledge (Yan Huang, 2024) View paper
  - Scene Graph and Structural Alignment (1 papers)
  - [4] Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval (Ding Jiang, 2023) View paper
  - Prompt-Based and Language-Guided Learning (3 papers)
  - [36] MadCLIP: Few-shot Medical Anomaly Detection with CLIP (Beyan Cigdem, 2025) View paper
  - [39] Language quantized autoencoders: Towards unsupervised text-image alignment (Liu Hao, 2023) View paper
  - [44] MAtch, eXpand and Improve: Unsupervised Finetuning for Zero-Shot Action Recognition with Language Knowledge (Wei Lin, 2023) View paper
- Hashing and Efficient Retrieval
  - Unsupervised Cross-Modal Hashing (3 papers)
  - [41] UCMH: Unpaired cross-modal hashing with matrix factorization (Jing Gao, 2020) View paper
  - [43] Unsupervised Dual Deep Hashing With Semantic-Index and Content-Code for Cross-Modal Retrieval (Bin Zhang, 2024) View paper
  - [46] Unsupervised Dual Hashing Coding (UDC) on Semantic Tagging and Sample Content for Cross-Modal Retrieval (Hongmin Cai, 2024) View paper
  - Adversarial and Self-Supervised Hashing (2 papers)
  - [16] Revising similarity relationship hashing for unsupervised cross-modal retrieval (You Wu, 2025) View paper
  - [48] Self-supervised adversarial hashing networks for cross-modal retrieval (Chao Li, 2018) View paper
- Domain-Specific and Application-Oriented Methods
  - Domain-Specific Retrieval Applications (4 papers)
  - [2] Revamping cross-modal recipe retrieval with hierarchical transformers and self-supervised learning (Amaia Salvador, 2021) View paper
  - [10] Zero-Shot Cross-Modal Retrieval for Remote Sensing Images With Minimal Supervision (Ushasi Chaudhuri, 2022) View paper
  - [35] Start from Video-Music Retrieval: An Inter-Intra Modal Loss for Cross Modal Retrieval (Chen Zeyu, 2024) View paper
  - [38] Surface material retrieval using weakly paired cross-modal learning (Huaping Liu, 2018) View paper
  - General Cross-Modal Retrieval (3 papers)
  - [15] Enhanced Partially Relevant Video Retrieval through Inter- and Intra-Sample Analysis with Coherence Prediction (Zhang Gang-jian, 2025) View paper
  - [18] Breaking Through the Noisy Correspondence: A Robust Model for Image-Text Matching (Haitao Shi, 2024) View paper
  - [50] Descriptive Image-Text Matching with Graded Contextual Similarity (Jang, 2025) View paper
- Transfer Learning and Adaptation
  - Zero-Shot and Few-Shot Learning (1 papers)
  - [32] Ternary Adversarial Networks With Self-Supervision for Zero-Shot Cross-Modal Retrieval (Xing Xu, 2020) View paper
  - Unsupervised Domain Adaptation (2 papers)
  - [33] Dual alignment unsupervised domain adaptation for video-text retrieval (Xiao-Shuai Hao, 2023) View paper
  - [49] Deep multimodal transfer learning for cross-modal retrieval (Liangli Zhen, 2020) View paper
- Specialized Architectures and Auxiliary Tasks
  - Hierarchical and Attention-Based Architectures (1 papers)
  - [24] Twoâ□□stage partial imageâ□□text clustering (TPITâ□□C) (Dongjin Guo, 2022) View paper
  - Auxiliary Task Learning (3 papers)
  - [1] LITA: lmm-guided image-text alignment for art assessment (Kaede Shiohara, 2024) View paper
  - [34] Multimodal llms as customized reward models for text-to-image generation (Zhou Shi-jie, 2025) View paper
  - [45] Improving Joint Speech-Text Representations Without Alignment (Cal Peyser, 2023) View paper

Narrative

Core task: unpaired image-text matching addresses the challenge of learning cross-modal correspondences when images and texts are not explicitly paired during training. The field's taxonomy reveals several complementary research directions. Contrastive Learning and Alignment Frameworks emphasize metric learning and embedding space optimization, often leveraging large-scale pretraining strategies. Generation-Based Unpaired Matching explores synthesis approaches, using captioning or image generation to bridge modalities. Semantic Knowledge and Prototype-Based Methods incorporate structured knowledge, conceptual prototypes, or external semantic resources to guide alignment without direct supervision. Hashing and Efficient Retrieval focuses on compact representations for scalable search, while Domain-Specific and Application-Oriented Methods tailor solutions to specialized contexts such as medical imaging (MedCLIP[6]) or remote sensing (Zero-Shot Remote Sensing[10]). Transfer Learning and Adaptation investigates how pretrained models can be fine-tuned or adapted to unpaired scenarios, and Specialized Architectures and Auxiliary Tasks introduce novel network designs or auxiliary objectives to improve matching quality.

Within Semantic Knowledge and Prototype-Based Methods, a particularly active line of work leverages conceptual knowledge and prototype alignment to impose semantic structure on learned embeddings. Multimodal Aligned Semantic[0] exemplifies this approach by integrating semantic prototypes to align image and text representations in a shared conceptual space, closely related to efforts like MACK[14] and Unpaired Conceptual Knowledge[47], which similarly exploit structured knowledge to guide unpaired matching. In contrast, UniAlign[3] and Quality-Aware Alignment[11] emphasize alignment robustness and quality assessment across modalities,

highlighting trade-offs between semantic richness and computational efficiency. The original paper sits naturally within this prototype-driven cluster, sharing with MACK[14] and Unpaired Conceptual Knowledge[47] a focus on leveraging external semantic structures, yet it appears to place greater emphasis on multimodal alignment mechanisms that explicitly coordinate conceptual representations across vision and language.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. MACK: Multimodal Aligned Conceptual Knowledge for Unpaired Image-text Matching

Authors: Yan Huang, Yuming Wang, Yukun Zeng, Liang Wang | **Year/Venue**: 2022 • Neural Information Processing Systems | **URL**: View paper

Abstract

N/A

Relationship Analysis

Both papers belong to the Prototype and Conceptual Knowledge Alignment category, utilizing prototypes to bridge unpaired image-text representations. They share the core approach of constructing multimodal aligned conceptual knowledge through prototypical region representations aligned with words. The key difference is that MASK (original paper) extends beyond MACK by introducing word embeddings as semantic bridges for OOD word handling, incorporating prototype consistency contrastive learning to mitigate distributional variance, and employing a modality transfer model to capture semantic relationships between words, whereas MACK focuses primarily on basic prototype-word alignment without these advanced mechanisms.

---

2. Unpaired Image-Text Matching via Multimodal Aligned Conceptual Knowledge

Authors: Yan Huang, Yu-Ming Wang, Y. F. Wang, Yunan Zeng, Yuming Wang, et al. (8 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Recently, the accuracy of image-text matching has been greatly improved by multimodal pretrained models, all of which use millions or billions of paired images and texts for supervised model learning. Different from them, human brains can well match images with texts using their stored multimodal knowledge. Inspired by that, this paper studies a new scenario as unpaired image-text matching, in which paired images and texts are assumed to be unavailable during model learning. To deal with it, we ...

Relationship Analysis

Both papers belong to the Prototype and Conceptual Knowledge Alignment category, utilizing prototypical region representations and conceptual knowledge to bridge unpaired image-text modalities. They share the core approach of constructing multimodal aligned knowledge through prototypes extracted from Visual Genome and applying this knowledge for unpaired matching. The key differences are that the original paper (MASK) introduces word embeddings as explicit semantic bridges for OOD word handling, employs prototype consistency contrastive learning for variance regularization, and uses a modality transfer model to preserve semantic relationships, while the candidate paper (MACK) focuses on self-supervised refinement of domain knowledge and bidirectional similarity pooling without the explicit semantic alignment mechanisms.

Contributions Analysis

Overall novelty summary. The paper proposes MASK, a method that uses word embeddings as bridges to align image and text modalities through prototype-based semantic knowledge. It sits within the 'Prototype and Conceptual Knowledge Alignment' leaf of the taxonomy, which contains only three papers total, including this one. This is a relatively sparse research direction compared to more crowded areas like 'General Vision-Language Contrastive Learning' (six papers) or 'Pseudo-Pair Generation for Captioning' (four papers), suggesting the prototype-driven semantic alignment approach represents a less explored path within unpaired image-text matching.

The taxonomy reveals that MASK's closest neighbors are MACK and Unpaired Conceptual Knowledge, both sharing the focus on leveraging structured semantic knowledge for unpaired matching. The broader 'Semantic Knowledge and Prototype-Based Methods' branch also includes scene graph approaches and prompt-based methods, which pursue structural or language-guided alignment rather than prototype-centric strategies. Adjacent branches like 'Contrastive Learning and Alignment Frameworks' emphasize metric learning without explicit semantic structures, while 'Generation-Based Unpaired Matching' synthesizes pseudo-pairs rather than directly aligning conceptual representations. MASK's position suggests it bridges semantic knowledge exploitation with contrastive alignment objectives.

Among 26 candidates examined across three contributions, the analysis reveals mixed novelty signals. The core MASK method (Contribution 1) examined 10 candidates with zero refutations, suggesting relative novelty in its specific multimodal alignment mechanism. However, the prototype consistency contrastive loss (Contribution 2) found 2 refutable candidates among 10 examined, indicating substantial prior work on prototype-based contrastive objectives. The relation-preserving equivariant mapping (Contribution 3) identified 1 refutable candidate among 6 examined, suggesting moderate overlap with existing approaches using external word embeddings for semantic alignment. The limited search scope means these findings reflect top-30 semantic matches rather than exhaustive coverage.

Based on the limited literature search, MASK appears to offer moderate novelty in its integrated approach to multimodal prototype alignment, though individual components show varying degrees of prior exploration. The sparse population of its taxonomy leaf and the absence of refutations for its core method suggest potential distinctiveness, but the prototype consistency loss and word embedding mapping show clearer connections to existing work. The analysis covers top-K semantic matches and does not claim exhaustive field coverage.

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

Contribution 1: Multimodal Aligned Semantic Knowledge (MASK) method

Description: The authors introduce MASK, a method that uses word embeddings to bridge words and visual prototypes, enabling semantic alignment between image and text modalities. For out-of-distribution words, representative prototypes are constructed by exploiting semantic relationships in word embeddings.

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. Consensus-aware visual-semantic embedding for image-text matching

URL: View paper

Brief Assessment

Consensus-aware Embedding[60] focuses on exploiting consensus knowledge through concept correlation graphs and graph convolutional networks for image-text matching, rather than using word embeddings to bridge words and visual prototypes for handling out-of-distribution words as in the original paper.

---

2. MKVSE: Multimodal knowledge enhanced visual-semantic embedding for image-text retrieval

URL: View paper

Brief Assessment

MKVSE[66] focuses on building a multimodal knowledge graph with intra-modal semantic relations and inter-modal co-occurrence relations for image-text retrieval, rather than using word embeddings to bridge words and visual prototypes for handling out-of-distribution words as in the original paper's MASK method.

---

3. AdaFV: Rethinking of Visual-Language alignment for VLM acceleration

URL: View paper

Brief Assessment

AdaFV[57] focuses on visual token pruning for VLM acceleration using text-to-image similarity and visual saliency, not on constructing semantic knowledge bridges between word embeddings and visual prototypes for unpaired image-text matching.

---

4. Align2Concept: Language Guided Interpretable Image Recognition by Visual Prototype and Textual Concept Alignment

URL: View paper

Brief Assessment

Align2Concept[63] focuses on aligning visual prototypes with textual concepts for interpretable image recognition using manifold alignment and CLIP embeddings, not on unpaired image-text matching or constructing prototypes for out-of-distribution words through word embedding semantic relationships.

---

5. Alignvlm: Bridging vision and language latent spaces for multimodal understanding

URL: View paper

Brief Assessment

AlignVLM[62] focuses on aligning vision encoder outputs to LLM text embeddings via weighted averages for document understanding tasks, not on constructing cross-modal semantic knowledge using word embeddings to bridge visual prototypes with text for unpaired image-text matching.

---

6. Aligning Information Capacity Between Vision and Language via Dense-to-Sparse Feature Distillation for Image-Text Matching

URL: View paper

Brief Assessment

Dense-to-Sparse Distillation[61] focuses on distilling dense text features into sparse text embeddings to enhance information capacity for image-text matching, not on using word embeddings to bridge visual prototypes with text for semantic alignment as in MASK.

---

7. Kernel-based unsupervised embedding alignment for enhanced visual representation in vision-language models

URL: View paper

Brief Assessment

Kernel-based Embedding Alignment[58] focuses on aligning visual representations between CLIP and DINOv2 in kernel space for vision-language models, not on using word embeddings to bridge words and visual prototypes for semantic alignment as in MASK.

---

8. Dpa: Dual prototypes alignment for unsupervised adaptation of vision-language models

URL: View paper

Brief Assessment

Dual Prototypes Alignment[64] focuses on unsupervised domain adaptation for vision-language models using dual prototypes (image and textual) for pseudo-labeling, not on bridging words and visual prototypes through word embeddings for semantic alignment as in MASK.

---

9. Asymmetric Visual Semantic Embedding Framework for Efficient Vision-Language Alignment

URL: View paper

Brief Assessment

Asymmetric Visual Semantic[59] focuses on asymmetric visual-text embeddings with meta-semantic units for image-text matching, not on using word embeddings to bridge words and visual prototypes for semantic alignment as in MASK.

---

10. Semantic prompt for few-shot image recognition

URL: View paper

Brief Assessment

Semantic Prompt[65] focuses on few-shot image recognition by using text embeddings as prompts to tune visual feature extraction networks, whereas MASK addresses unpaired image-text matching by constructing cross-modal aligned knowledge with prototypical region representations.

Contribution 2: Prototype consistency contrastive learning loss

Description: A novel contrastive loss is proposed that uses prototypes as class centers to maximize similarity between region representations and their corresponding prototypes while minimizing similarity with other prototypes. This regularizes the feature space and reduces the impact of distributional variance across different words.

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. Prototypical contrastive learning through alignment and uniformity for recommendation

URL: View paper

Brief Assessment

Prototypical Alignment Uniformity[69] focuses on recommendation systems using graph collaborative filtering, not visual-linguistic feature space regularization for image-text matching. The technical contexts are fundamentally different.

---

2. Prototypical graph contrastive learning

URL: View paper

Prior Art Analysis

Prototypical Graph Contrastive[72] demonstrates that using prototypes as class centers in contrastive learning was already proposed in the graph domain. The candidate paper introduces a prototype consistency contrastive learning approach that maximizes similarity between representations and their corresponding prototypes while minimizing similarity with other prototypes, which is fundamentally the same mechanism described in the original paper's contribution. Both papers employ prototypes to regularize the feature space and reduce distributional variance, with the candidate explicitly stating this objective through their clustering consistency and reweighted contrastive objectives.

Evidence

Evidence 1 - **Rationale**: Both approaches use prototype-based mechanisms to regularize the feature space and handle distributional variance through contrastive objectives. - **Original**: we introduce a prototype consistency contrastive learning loss to structurally regularize the feature space, which explicitly encourages region representations associated with the same word to align closely with their prototype, thereby mitigating the adverse impact of distributional variance. - **Candidate**: we design a reweighted contrastive objective, which reweights the negative samples based on the distance between their prototypes and the query prototype. in this way, those negative pairs having moderate prototype distance enjoy relatively large weights, which ensures the semantic difference betwee...

Evidence 2 - **Rationale**: Both papers maximize similarity to corresponding prototypes while minimizing similarity to other prototypes, achieving the same intra-class aggregation and inter-class separation objective. - **Original**: the loss lcl employs prototypes as class centers, maximizing the similarity between region representations and their corresponding prototypes while minimizing similarity with other prototypes, thereby achieving intra-class aggregation and inter-class separation. - **Candidate**: to encourage the clustering consistency between two correlated views gi and g′ i, we predict the cluster assignments of g′ i with the representation zi (rather than z′ i) from the correlated view and vice versa. formally, we define the clustering consistency objective via minimizing the average cross...

---

3. Prototype Enhancement-Based Incremental Evolution Learning for Urban Garbage Classification

URL: View paper

Brief Assessment

Prototype Enhancement Incremental[75] focuses on incremental learning for garbage classification using prototype enhancement via Gaussian kernel density estimation and contrastive features for task consistency. The original paper's prototype consistency contrastive learning loss maximizes similarity between region representations and prototypes in image-text matching, which is a fundamentally different application domain and technical approach.

---

4. Prototype-driven multi-view attribute-missing graph clustering

URL: View paper

Brief Assessment

Prototype Multi-view Clustering[73] focuses on multi-view attribute-missing graph clustering with prototype consistency loss for feature interpolation, while the original paper addresses unpaired image-text matching with prototype consistency contrastive learning for regularizing distributional variance across words. These are fundamentally different problem domains and technical approaches.

---

5. Multi-Label Prototype Visual Spatial Search for Weakly Supervised Semantic Segmentation

URL: View paper

Brief Assessment

Multi-Label Prototype Search[76] focuses on weakly supervised semantic segmentation using multi-label prototypes with different loss objectives (cross-class, cross-image, patch-to-prototype), while the original paper addresses unpaired image-text matching with a different prototype consistency mechanism for cross-modal alignment.

---

6. Prototypical contrastive learning of unsupervised representations

URL: View paper

Prior Art Analysis

Prototypical Contrastive Learning[67] demonstrates that a prototype-based contrastive learning approach was proposed prior to the original paper. The candidate paper introduces prototypes as cluster centroids and optimizes a contrastive loss (ProtoNCE) that maximizes similarity between embeddings and their assigned prototypes while minimizing similarity with other prototypes. This directly addresses the same objective as the original paper's contribution: using prototypes to regularize the feature space and reduce distributional variance. The candidate's formulation in Equation 10 and 11 shows the mathematical framework for prototype-based contrastive learning that predates the original work.

Evidence

Evidence 1 - **Rationale**: Both papers propose a contrastive loss that uses prototypes to regularize feature distributions. The candidate's ProtoNCE loss serves the same purpose as the original's prototype consistency contrastive learning loss. - **Original**: we introduce a prototype consistency contrastive learning loss to structurally regularize the feature space, which explicitly encourages region representations associated with the same word to align closely with their prototype, thereby mitigating the adverse impact of distributional variance. - **Candidate**: we propose protonce, a new contrastive loss which improves the widely used infonce by dynamically estimating the concentration for the feature distribution around each prototype. protonce also includes an infonce term in which the instance embeddings can be interpreted as instance-based prototypes.

Evidence 2 - **Rationale**: The candidate's Equation 10 shows the mathematical formulation of a prototype-based contrastive loss that maximizes similarity with assigned prototypes (cs) while minimizing with others (cj), with concentration estimation to handle distributional variance - the exact mechanism described in the original contribution. - **Original**: inspired by clustering theory and contrastive learning, we design a prototype consistencycontrastive learning loss lcl to reduce the influence of distributional variance between prototypes and their related region representations. the loss lcl employs prototypes as class centers, maximizing the simil... - **Candidate**: θ∗= arg min θ n∑ i=1 -log exp(vi ·cs/φs) ∑k j=1 exp(vi ·cj/φj) , (10) where φ∝σ2 denotes the concentration level of the feature distribution around a prototype

---

7. Semi-supervised semantic segmentation with prototype-based consistency regularization

URL: View paper

Brief Assessment

Prototype-based Consistency[68] focuses on semi-supervised semantic segmentation with pixel-level predictions, using prototypes to regularize feature distributions within classes. The original paper addresses unpaired image-text matching with cross-modal alignment, a fundamentally different task and technical approach.

---

8. MVPCL: multi-view prototype consistency learning for semi-supervised medical image segmentation

URL: View paper

Brief Assessment

MVPCL[71] focuses on medical image segmentation with multi-view consistency and uncertainty-based prototype regularization, which differs from the ORIGINAL paper's cross-modal image-text matching framework. The candidate's approach is domain-specific to medical imaging rather than general multimodal alignment.

---

9. Decoupled Prototype Learning for Reliable Test-Time Adaptation

URL: View paper

Brief Assessment

Decoupled Prototype Learning[74] focuses on test-time adaptation with a decoupled prototype learning approach for handling noisy pseudo-labels during domain shift, not on regularizing feature space for cross-modal image-text matching as in the original paper.

---

10. Calibration-based multi-prototype contrastive learning for domain generalization semantic segmentation in traffic scenes

URL: View paper

Brief Assessment

Multi-prototype Contrastive[70] focuses on domain generalization for semantic segmentation in traffic scenes, using multi-prototype learning to handle domain shift. The original paper addresses unpaired image-text matching with a different objective of cross-modal alignment.

Contribution 3: Relation-preserving equivariant mapping using external word embeddings

Description: The authors integrate pre-trained word vectors as supervision to create a mapping that preserves semantic relationships between the visual and linguistic modalities. This enables region representations to effectively capture semantic correlations among words.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Unsupervised object representation learning using translation and rotation group equivariant vae

URL: View paper

Brief Assessment

Equivariant VAE[55] focuses on learning rotation and translation equivariant representations for visual objects in images, not on creating mappings between visual regions and word embeddings for cross-modal matching tasks.

---

2. Equivact: Sim (3)-equivariant visuomotor policies beyond rigid object manipulation

URL: View paper

Brief Assessment

Equivact[52] focuses on sim(3)-equivariant visuomotor policies for robot manipulation tasks using point cloud networks. The original paper addresses cross-modal semantic alignment between visual regions and word embeddings for image-text matching. These are fundamentally different technical domains with no overlap in methodology or application.

---

3. Understanding the role of equivariance in self-supervised learning

URL: View paper

Brief Assessment

Equivariance Self-supervised[51] focuses on self-supervised learning with equivariance to data augmentations (e.g., rotations) for visual representation learning, not on mapping visual regions to word embeddings for cross-modal matching.

---

4. Equivariant Open-vocabulary Pick and Place via Language Kernels and Patch-level Semantic Maps

URL: View paper

Brief Assessment

Equivariant Pick Place[56] focuses on robotic manipulation with SE(2) equivariance for pick-and-place tasks, not on cross-modal image-text matching or semantic alignment between visual regions and word embeddings.

---

5. Visual Relationship Transformation

URL: View paper

Brief Assessment

Visual Relationship Transformation[54] focuses on predicting visual relationships across different camera views in 3D scenes, not on mapping visual regions to word embeddings for image-text matching. The equivariance in the candidate refers to spatial transformations (rotation/translation) between views, not semantic relationship preservation between visual and linguistic modalities.

---

6. Equivariant similarity for vision-language foundation models

URL: View paper

Prior Art Analysis

Equivariant Similarity[53] demonstrates prior work on relation-preserving equivariant mappings between visual and linguistic modalities. The candidate paper explicitly defines and implements equivariant feature maps that preserve semantic relationships between image and text spaces, using external supervision to ensure that similarity changes correspond to semantic changes. This directly challenges the novelty claim of the original paper's contribution, as both works establish mappings that preserve semantic relationships between modalities using external knowledge as supervision.

Evidence

Evidence 1 - **Rationale**: This pair shows that Equivariant Similarity[53] formally defines equivariant mappings between image and text spaces, establishing the theoretical foundation for relation-preserving transformations between modalities before the original paper's submission. - **Original**: we incorporate external knowledge from pre-trained word vectors as auxiliary supervision signals, which establishes a relation-preserving equivariant mapping between region representations and word embeddings, enabling the region representations to effectively capture semantic relationships among wo... - **Candidate**: definition 2 let i and t be two continuous feature spaces. let g be a group whose group action on i is defined by g : i → i, and that on t is defined by g′ : t → t. then, ϕ : i → tis an equivariant feature map if and only if g′ ·ϕ(i) =ϕ(g ·i) for all the group actions andi ∈ i.

Evidence 2 - **Rationale**: Both papers use external supervision to create mappings that preserve semantic relationships. The candidate demonstrates that equivariant mappings preserve distance metrics between modalities, which is the core mechanism claimed as novel in the original paper. - **Original**: the loss function lcm enforces the predicted word embeddings v′to gradually converge toward the word embeddingsv, while simultaneously ensuring that the region representations effectively capture the semantic relationships between

words. - **Candidate**: this implies that the equivariant map establishes an isometry for the measure μ(i) in image space and ϕ(μ(i)) in text space. thus, they only differ by a constant scale c > 0, i.e., ϕ(μ(i)) =cμ(i)

Evidence 3 - **Rationale**: This pair demonstrates that Equivariant Similarity[53] implements a mathematically rigorous framework for preserving semantic relationships across modalities through equivariant constraints, which is the same fundamental approach claimed in the original paper. - **Original**: to enable region representations to effectively capture semantic correlations between words, we utilize a modality transfer model (mtm) f with three self-attention layers and three fc layers that can map the mean μoutput by the pae model hinto the word embedding space - **Candidate**: by combining eq. (1), (2), (3), and (4), we have the following ratio equality as our eqsim constraint: s11 - s12 s11 - s21 = s22 - s21 s22 - s12 = c = 1.

Appendix: Text Similarity Detection

Textual similarity detection checked 28 papers and found 2 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Unpaired Image-Text Matching via Multimodal Aligned Conceptual Knowledge

Detected in: Core Task (sibling)

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Multimodal Aligned Semantic Knowledge for Unpaired Image-text Matching View paper
- [1] LITA: lmm-guided image-text alignment for art assessment View paper
- [2] Revamping cross-modal recipe retrieval with hierarchical transformers and self-supervised learning View paper
- [3] UniAlign: Scaling Multimodal Alignment within One Unified Model View paper
- [4] Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval View paper
- [5] Categorical SchrÃ¶dinger Bridge Matching View paper
- [6] MedCLIP: Contrastive Learning from Unpaired Medical Images and Text View paper
- [7] Improving cross-modal alignment with synthetic pairs for text-only image captioning View paper
- [8] Prompt-based learning for unpaired image captioning View paper
- [9] PathDiff: Histopathology Image Synthesis with Unpaired Text and Mask Conditions View paper
- [10] Zero-Shot Cross-Modal Retrieval for Remote Sensing Images With Minimal Supervision View paper
- [11] Quality-Aware Image-Text Alignment for Opinion-Unaware Image Quality Assessment View paper
- [12] Unpaired image captioning via scene graph alignments View paper
- [13] The "Law" of the Unconscious Contrastive Learner: Probabilistic Alignment of Unpaired Modalities View paper
- [14] MACK: Multimodal Aligned Conceptual Knowledge for Unpaired Image-text Matching View paper
- [15] Enhanced Partially Relevant Video Retrieval through Inter- and Intra-Sample Analysis with Coherence Prediction View paper
- [16] Revising similarity relationship hashing for unsupervised cross-modal retrieval View paper
- [17] Generating Multimodal Images with GAN: Integrating Text, Image, and Style View paper
- [18] Breaking Through the Noisy Correspondence: A Robust Model for Image-Text Matching View paper
- [19] PyramidCLIP: Hierarchical Feature Alignment for Vision-language Model Pretraining View paper
- [20] Text-based person search without parallel image-text data View paper
- [21] Backdoor Attack on Un-paired Medical Image-Text Pretrained Models: A Pilot Study on MedCLIP View paper
- [22] Better Together: Leveraging Unpaired Multimodal Data for Stronger Unimodal Models View paper
- [23] Unpaired image captioning with semantic-constrained self-learning View paper
- [24] Twoâ□□stage partial imageâ□□text clustering (TPITâ□□C) View paper
- [25] CL2CM: Improving Cross-Lingual Cross-Modal Retrieval via Cross-Lingual Knowledge Transfer View paper
- [26] Cross-modal generation and alignment via attribute-guided prompt for unsupervised text-based person retrieval View paper
- [27] Relaxing Binary Constraints in Contrastive Vision-Language Medical Representation Learning View paper
- [28] From coarse to fine: a two-stage common semantic space construction for unpaired cross modal retrieval View paper
- [29] Doracycle: Domain-oriented adaptation of unified generative model in multimodal cycles View paper
- [30] Can Language-Guided Unsupervised Adaptation Improve Medical Image Classification Using Unpaired Images and Texts? View paper
- [31] Self-supervised image-text pre-training with mixed data in chest x-rays View paper
- [32] Ternary Adversarial Networks With Self-Supervision for Zero-Shot Cross-Modal Retrieval View paper
- [33] Dual alignment unsupervised domain adaptation for video-text retrieval View paper
- [34] Multimodal llms as customized reward models for text-to-image generation View paper
- [35] Start from Video-Music Retrieval: An Inter-Intra Modal Loss for Cross Modal Retrieval View paper
- [36] MadCLIP: Few-shot Medical Anomaly Detection with CLIP View paper
- [37] Multimodal LLM Enhanced Cross-lingual Cross-modal Retrieval View paper
- [38] Surface material retrieval using weakly paired cross-modal learning View paper
- [39] Language quantized autoencoders: Towards unsupervised text-image alignment View paper
- [40] Leveraging unpaired data for vision-language generative models via cycle consistency View paper
- [41] UCMH: Unpaired cross-modal hashing with matrix factorization View paper
- [42] S-CLIP: Semi-supervised Vision-Language Learning using Few Specialist Captions View paper
- [43] Unsupervised Dual Deep Hashing With Semantic-Index and Content-Code for Cross-Modal Retrieval View paper
- [44] MAtch, eXpand and Improve: Unsupervised Finetuning for Zero-Shot Action Recognition with Language Knowledge View paper
- [45] Improving Joint Speech-Text Representations Without Alignment View paper
- [46] Unsupervised Dual Hashing Coding (UDC) on Semantic Tagging and Sample Content for Cross-Modal Retrieval View paper
- [47] Unpaired Image-Text Matching via Multimodal Aligned Conceptual Knowledge View paper
- [48] Self-supervised adversarial hashing networks for cross-modal retrieval View paper
- [49] Deep multimodal transfer learning for cross-modal retrieval View paper
- [50] Descriptive Image-Text Matching with Graded Contextual Similarity View paper

- [51] Understanding the role of equivariance in self-supervised learning View paper
- [52] Equivact: Sim (3)-equivariant visuomotor policies beyond rigid object manipulation View paper
- [53] Equivariant similarity for vision-language foundation models View paper
- [54] Visual Relationship Transformation View paper
- [55] Unsupervised object representation learning using translation and rotation group equivariant vae View paper
- [56] Equivariant Open-vocabulary Pick and Place via Language Kernels and Patch-level Semantic Maps View paper
- [57] AdaFV: Rethinking of Visual-Language alignment for VLM acceleration View paper
- [58] Kernel-based unsupervised embedding alignment for enhanced visual representation in vision-language models View paper
- [59] Asymmetric Visual Semantic Embedding Framework for Efficient Vision-Language Alignment View paper
- [60] Consensus-aware visual-semantic embedding for image-text matching View paper
- [61] Aligning Information Capacity Between Vision and Language via Dense-to-Sparse Feature Distillation for Image-Text Matching View paper
- [62] Alignvlm: Bridging vision and language latent spaces for multimodal understanding View paper
- [63] Align2Concept: Language Guided Interpretable Image Recognition by Visual Prototype and Textual Concept Alignment View paper
- [64] Dpa: Dual prototypes alignment for unsupervised adaptation of vision-language models View paper
- [65] Semantic prompt for few-shot image recognition View paper
- [66] MKVSE: Multimodal knowledge enhanced visual-semantic embedding for image-text retrieval View paper
- [67] Prototypical contrastive learning of unsupervised representations View paper
- [68] Semi-supervised semantic segmentation with prototype-based consistency regularization View paper
- [69] Prototypical contrastive learning through alignment and uniformity for recommendation View paper
- [70] Calibration-based multi-prototype contrastive learning for domain generalization semantic segmentation in traffic scenes View paper
- [71] MVPCL: multi-view prototype consistency learning for semi-supervised medical image segmentation View paper
- [72] Prototypical graph contrastive learning View paper
- [73] Prototype-driven multi-view attribute-missing graph clustering View paper
- [74] Decoupled Prototype Learning for Reliable Test-Time Adaptation View paper
- [75] Prototype Enhancement-Based Incremental Evolution Learning for Urban Garbage Classification View paper
- [76] Multi-Label Prototype Visual Spatial Search for Weakly Supervised Semantic Segmentation View paper