Novelty Assessment Report

Paper: MaskInversion: Localized Embeddings via Optimization of Explainability Maps
PDF URL: https://openreview.net/pdf?id=3xyx2ncRln
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-07

Abstract

Vision-language foundation models such as CLIP have achieved tremendous results in global vision-language alignment, but still show some limitations in creating representations for specific image regions. To address this problem, we propose MaskInversion, a method that leverages the feature representations of pre-trained foundation models, such as CLIP, to generate a context-aware embedding for a query image region specified by a mask at test time. MaskInversion starts with initializing an embedding token and compares its explainability map, derived from the pretrained model, to the query mask. The embedding token is then subsequently refined to approximate the query region by minimizing the discrepancy between its explainability map and the query mask. During this process, only the embedding vector is updated, while the underlying foundation model is kept frozen allowing to use MaskInversion with any pre-trained model. As deriving the explainability map involves computing its gradient, which can be expensive, we propose a gradient decomposition strategy that simplifies this computation. The learned region representation can be used for a broad range of tasks, including open-vocabulary class retrieval, referring expression comprehension, as well as for localized captioning and image generation. We evaluate the proposed method on all those tasks on several datasets such as PascalVOC, MSCOCO, RefCOCO, and OpenImagesV7 and show its capabilities compared to other SOTA approaches.

Core Task Landscape

This paper addresses: **Generating localized embeddings for specific image regions from vision-language models**

A total of **50 papers** were analyzed and organized into a taxonomy with **14 categories**.

Taxonomy Overview

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

- **Region-Aware Vision-Language Pre-training and Alignment**
- **Spatial and 3D-Aware Region Representation**
- **Region-Based Inference and Prompting**
- **Multi-Modal Task Integration with Region Features**
- **Specialized Region-Based Applications**
- **Analysis and Understanding of Region Representations**

Complete Taxonomy Tree

- Generating localized embeddings for specific image regions from vision-language models Survey Taxonomy
- Region-Aware Vision-Language Pre-training and Alignment
  - Region-Text Contrastive Learning (5 papers)
  - [6] Glipv2: Unifying localization and vision-language understanding (Zhang, 2022) View paper
  - [20] Regionclip: Region-based language-image pretraining (Yiwu Zhong, 2022) View paper
  - [32] Aligning Bag of Regions for Open-Vocabulary Object Detection (Size Wu, 2023) View paper
  - [34] Contrastive Localized Language-Image Pre-Training (Chen, 2024) View paper
  - [39] Region-based cluster discrimination for visual representation learning (Xie Yin, 2025) View paper
  - Region Description Generation for Alignment (3 papers)
  - [8] Regiongpt: Towards region understanding vision language model (Qiushan Guo, 2024) View paper
  - [13] Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks (Zhe Chen, 2024) View paper
  - [21] Learning Visual Grounding from Generative Vision and Language Model (Shijie Wang, 2024) View paper
  - Domain-Specific Region-Aware Pre-training (5 papers)
  - [2] Geochat: Grounded large vision-language model for remote sensing (Kartik Kuckreja, 2024) View paper
  - [11] FarmSeg_VLM: A farmland remote sensing image segmentation method considering vision-language alignment (Haiyang Wu, 2025) View paper
  - [31] RegionMed-CLIP: A Region-Aware Multimodal Contrastive Learning Pre-trained Model for Medical Image Understanding (Guiru, 2025) View paper
  - [36] Fsvlm: A vision-language model for remote sensing farmland segmentation (Haiyang Wu, 2025) View paper
  - [41] GeoMag: A Vision-Language Model for Pixel-level Fine-Grained Remote Sensing Image Parsing (Ma Xianzhi, 2025) View paper
- Spatial and 3D-Aware Region Representation
  - Depth and 3D Scene Graph Integration (5 papers)
  - [1] Spatialrgpt: Grounded spatial reasoning in vision-language models (An-Chieh Cheng, 2024) View paper
  - [3] SpatialRGPT: Grounded Spatial Reasoning in Vision Language Model (Cheng, 2024) View paper
  - [4] SpatialBot: Precise Spatial Understanding with Vision Language Models (Wenxiao Cai, 2025) View paper
  - [5] Spatial 3d-llm: Exploring spatial awareness in 3d vision-language models (Xiaoyan Wang, 2025) View paper

    - [7] 3d aware region prompted vision language model (Cheng, 2025) View paper
    - 2D Spatial Relationship Modeling (4 papers)
    - [19] Locality alignment improves vision-language models (Covert, 2024) View paper
    - [24] SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities (Boyuan Chen, 2024) View paper
    - [30] Why Is Spatial Reasoning Hard for VLMs? An Attention Mechanism Perspective on Focus Areas (Chen Shi-qi, 2025) View paper
    - [33] Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding (Li, 2025) View paper
- Region-Based Inference and Prompting
    - Visual Prompting and Attention Guidance (5 papers)
    - [12] Introducing Visual Perception Token into Multimodal Large Language Model (Yu, 2025) View paper
    - [14] Contrastive Region Guidance: Improving Grounding in Vision-Language Models without Training (David Wan, 2024) View paper
    - [44] Guiding Medical Vision-Language Models with Explicit Visual Prompts: Framework Design and Comprehensive Exploration of Prompt Variations (Qin, 2025) View paper
    - [46] Chain-of-Focus: Adaptive Visual Search and Zooming for Multimodal Reasoning via RL (Zhang Xintong, 2025) View paper
    - [47] Guiding Medical Vision-Language Models with Diverse Visual Prompts: Framework Design and Comprehensive Exploration of Prompt Variations (Jiang Zekun, 2025) View paper
    - Embedding Optimization for Localized Representations ★ (2 papers)
    - [0] MaskInversion: Localized Embeddings via Optimization of Explainability Maps (Anon et al., 2026) View paper
    - [10] Lare: Latent augmentation using regional embedding with vision-language model (Kosuke Sakurai, 2025) View paper
    - Region-Conditioned Generation and Grounding (5 papers)
    - [9] Groma: Localized visual tokenization for grounding multimodal large language models (Ma, 2024) View paper
    - [18] Generate to Ground: Multimodal Text Conditioning Boosts Phrase Grounding in Medical Vision-Language Models (Dombrowski, 2025) View paper
    - [22] AffordanceLLM: Grounding Affordance from Vision Language Models (Shengyi Qian, 2024) View paper
    - [23] Grounding everything: Emerging localization properties in vision-language transformers (Walid Bousselham, 2024) View paper
    - [40] MedGround-R1: Advancing Medical Image Grounding via Spatial-Semantic Rewarded Group Relative Policy Optimization (Hui-Hui Xu, 2025) View paper
- Multi-Modal Task Integration with Region Features
    - Unified Multi-Granularity Architectures (4 papers)
    - [27] Pixel-aligned language model (J Xu, 2024) View paper
    - [42] u-llava: Unifying multi-modal tasks via large language model (Jinjin Xu, 2023) View paper
    - [43] Toward Interactive Regional Understanding in Vision-Large Language Models (Chun, 2024) View paper
    - [45] Finecaption: Compositional image captioning focusing on wherever you want at any granularity (Hang Hua, 2025) View paper
    - Global-Local Feature Fusion (4 papers)
    - [17] Visual-text cross alignment: Refining the similarity score in vision-language models (Li Jinhao, 2024) View paper
    - [28] GalLoP: Learning Global and Local Prompts for Vision-Language Models (Lafon, 2024) View paper
    - [29] Seeing Out of tHe bOx: End-to-End Pre-training for Vision-Language Representation Learning (Zhicheng Huang, 2021) View paper
    - [35] Overcoming the Pitfalls of Vision-Language Model for Image-Text Retrieval (Feifei Zhang, 2024) View paper
    - Temporal and Spatial-Temporal Region Modeling (2 papers)
    - [25] Self-chained image-language model for video localization and question answering (YU Shoubin, 2023) View paper
    - [50] STOP: Integrated Spatial-Temporal Dynamic Prompting for Video Understanding (Liu Zichen, 2025) View paper
- Specialized Region-Based Applications
    - Region-Based Retrieval and Detection (2 papers)
    - [26] HOICLIP: Efficient Knowledge Transfer for HOI Detection with Vision-Language Models (Shan Ning, 2023) View paper
    - [37] TextRegion: Text-Aligned Region Tokens from Frozen Image-Text Models (Xiao Yao, 2025) View paper
    - Region-Based Captioning and Question Answering (3 papers)
    - [38] Chain-of-region: Visual language models need details for diagram analysis (X Li, 2025) View paper
    - [48] Describe anything model for visual question answering on text-rich images (Duong-Dinh Thang, 2025) View paper
    - [49] Multi-modal Self-perception Enhanced Large Language Model for 3D Region-of-Interest Captioning with Limited Data (Lu Shi, 2025) View paper
- Analysis and Understanding of Region Representations (2 papers)
    - [15] Perceptual Grouping in Contrastive Vision-Language Models (Kanchana Ranasinghe, 2023) View paper
    - [16] What's in the Image? A Deep-Dive into the Vision of Vision Language Models (Bagon, 2025) View paper

Narrative

Core task: Generating localized embeddings for specific image regions from vision-language models. The field has organized itself into several major branches that reflect different emphases in how region-level representations are learned and applied. Region-Aware Vision-Language Pre-training and Alignment focuses on foundational training strategies that align visual regions with language at scale, often through contrastive or grounding objectives (e.g., GLIPv2[6], RegionCLIP[20]). Spatial and 3D-Aware Region Representation extends these ideas to capture geometric and depth information, enabling richer spatial reasoning (e.g., SpatialRGPT[1], Spatial 3D LLM[5]). Region-Based Inference and Prompting explores how to optimize or manipulate region embeddings at inference time, while Multi-Modal Task Integration with Region Features and Specialized Region-Based Applications address downstream uses ranging from interactive understanding to domain-specific tasks like medical imaging (e.g., RegionMed CLIP[31]). Finally, Analysis and Understanding of Region Representations investigates what these embeddings capture and how they can be improved.

A particularly active line of work centers on embedding optimization for localized representations, where methods seek to refine region features beyond what pre-training alone provides. MaskInversion[0] sits squarely in this space, focusing on inverting or optimizing embeddings to better capture fine-grained region semantics. This contrasts with approaches like LARE[10], which may emphasize different optimization strategies or prompting mechanisms to achieve localization. Meanwhile, works such as Groma[9] and RegionGPT[8] illustrate how region features can be integrated into large language models for grounded reasoning, highlighting a trade-off between optimization-centric methods and those that rely on architectural innovations or richer pre-training. The central question across these directions is how to balance computational efficiency, generalization across diverse region types, and the fidelity of localized embeddings—challenges that MaskInversion[0] addresses through its inversion-based framework.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Lare: Latent augmentation using regional embedding with vision-language model

Authors: Kosuke Sakurai, Tatsuya Ishii, Ryotaro Shimizu, Linxin Song, Masayuki Goto | **Year/Venue**: 2025 | **URL**: View paper

Abstract

In recent years, considerable research has been conducted on vision-language models (VLMs) that handle both image and text data; these models are being applied to diverse downstream tasks, such as â□□image-related chat,â□□ â□□image recognition by instruction,â□□ and â□□answering visual questions.â□□ Vision-language models, such as Contrastive Languageâ□□Image Pre-training (CLIP), are also high-performance image classifiers, and are being developed into domain adaptation methods that can utilize ...

Relationship Analysis

Both papers belong to the 'Embedding Optimization for Localized Representations' category, focusing on optimizing embeddings at test time to align with specific image regions. While MaskInversion optimizes embedding tokens by minimizing discrepancy between explainability maps and query masks for precise region localization, LARE (Latent Augmentation using Regional Embedding) embeds images as regions in the unified embedding space and samples augmented embeddings from within these latent regions for domain-robust classification. The key difference is that MaskInversion targets precise spatial localization using binary masks and explainability-driven optimization, whereas LARE focuses on domain generalization through latent region sampling for data augmentation.

Contributions Analysis

Overall novelty summary. The paper proposes MaskInversion, a test-time optimization method that refines embedding tokens to align with query image regions by minimizing discrepancies between explainability maps and input masks. Within the taxonomy, it resides in the 'Embedding Optimization for Localized Representations' leaf under 'Region-Based Inference and Prompting'. This leaf contains only two papers total, indicating a relatively sparse research direction compared to more crowded areas like 'Region-Text Contrastive Learning' (five papers) or 'Visual Prompting and Attention Guidance' (five papers). The work thus occupies a niche focused on iterative embedding refinement rather than fixed prompting or pre-training strategies.

The taxonomy reveals that neighboring leaves emphasize different inference-time strategies: 'Visual Prompting and Attention Guidance' uses markers or bounding boxes to guide attention without embedding optimization, while 'Region-Conditioned Generation and Grounding' focuses on generative outputs conditioned on regions. Broader branches like 'Region-Aware Vision-Language Pre-training' address foundational training with region supervision, which MaskInversion explicitly avoids by keeping models frozen. The scope notes clarify that this leaf excludes fixed visual prompts and pre-training modifications, positioning MaskInversion as a post-hoc optimization approach distinct from both training-time alignment and static prompting methods.

Among thirty candidates examined across three contributions, none were flagged as clearly refuting the proposed methods. The core MaskInversion contribution examined ten candidates with zero refutable overlaps, as did the gradient decomposition strategy and regularization loss components. This suggests that within the limited search scope—primarily top-K semantic matches and citation expansion—no prior work directly anticipates the combination of explainability-map-driven inversion with gradient decomposition for region embedding. However, the small candidate pool and sparse leaf population mean the analysis captures a snapshot rather than exhaustive coverage of potential prior art.

Given the limited search scope and the sparse taxonomy leaf, the work appears to introduce a distinct optimization-centric approach to region embeddings. The absence of refutable candidates among thirty examined papers, combined with only one sibling paper in the taxonomy, suggests the method occupies a relatively unexplored niche. However, the analysis does not cover broader embedding optimization literature outside vision-language models or alternative explainability-based techniques, leaving open questions about connections to related optimization paradigms in other domains.

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

Contribution 1: MaskInversion method for localized embeddings via explainability map optimization

Description: The authors introduce MaskInversion, a test-time optimization method that learns localized embedding tokens for specific image regions by iteratively refining an embedding to match its explainability map to a query mask, while keeping the foundation model frozen.

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. Artxai: Explainable artificial intelligence curates deep representation learning for artistic images using fuzzy techniques

URL: View paper

Brief Assessment

ArtXAI[77] focuses on explainable AI for artistic image classification using fuzzy techniques and does not address learning localized embeddings from image regions via explainability map optimization. The candidate paper's methodology centers on fuzzy rule-based classification and clustering for art analysis tasks, which is technically distinct from the original paper's test-time optimization approach for region-specific embeddings.

---

2. Finding Regions of Counterfactual Explanations via Robust Optimization

URL: View paper

Brief Assessment

Counterfactual Regions[74] focuses on generating robust counterfactual explanations for machine learning classifiers through optimization, not on learning localized embeddings from image regions using explainability maps for vision-language models.

---

3. CEIR: Concept-based Explainable Image Representation Learning

URL: View paper

Brief Assessment

CEIR[75] focuses on concept-based representation learning using VAE and concept bottleneck models for interpretability, not on learning localized embeddings from image regions via explainability map optimization as in MaskInversion.

---

4. Interpreting CLIP's Image Representation via Text-Based Decomposition

URL: View paper

Brief Assessment

Interpreting CLIP[71] focuses on decomposing CLIP's image representation across layers, heads, and tokens to interpret what information is encoded, not on learning localized embeddings for specific image regions via test-time optimization of explainability maps.

5. Interpretable representations in explainable AI: from theory to practice

URL: View paper

Brief Assessment

Interpretable Representations XAI[72] focuses on interpretable representations for explainability (tabular, image, text data) using methods like discretization and segmentation, not on learning localized embeddings via explainability map optimization as proposed in the original paper.

6. Explainable self-supervised learning for medical image diagnosis based on DINO V2 model and semantic search

URL: View paper

Brief Assessment

Explainable DINO[78] focuses on self-supervised learning (DINO V2) for medical image diagnosis with semantic search capabilities, not on learning localized embeddings via explainability map optimization for vision-language models.

7. Local Concept Embeddings for Analysis of Concept Distributions in DNN Feature Spaces

URL: View paper

Brief Assessment

Local Concept Embeddings[80] focuses on supervised concept-based explainability with dataset-level optimization for concept vectors, not test-time optimization of localized embeddings for specific image regions using explainability maps.

8. Explainable artificial intelligence (XAI) for deep learning based medical imaging classification

URL: View paper

Brief Assessment

Explainable Medical Imaging[79] focuses on medical image classification using segmentation-based explainability for COVID-19 diagnosis, not on learning localized embeddings through explainability map optimization for vision-language models.

9. Explainable AI enhanced transformer based UNet for medical images segmentation using gradient weighted class activation map

URL: View paper

Brief Assessment

Explainable Transformer UNet[73] focuses on medical image segmentation using gradient-weighted class activation maps (CAM) for CNNs, not on learning localized embeddings through explainability map optimization for vision-language models.

10. ICEv2: Interpretability, Comprehensiveness, and Explainability in Vision Transformer.

URL: View paper

Brief Assessment

ICEv2[76] focuses on explainability visualization for vision transformers using patch embeddings to predict image classes, not on learning localized embeddings for specific image regions through test-time optimization of explainability maps.

Contribution 2: Gradient decomposition strategy for efficient explainability map computation

Description: The authors propose a gradient decomposition technique that eliminates the need to compute second-order derivatives at each iteration by decomposing the gradient computation, thereby improving computational efficiency especially when processing multiple masks.

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. Deeply Explain CNN Via Hierarchical Decomposition

URL: View paper

Brief Assessment

Deeply Explain CNN[69] focuses on hierarchical decomposition of CNN decisions through layers for interpretability, not on eliminating second-order derivatives for computational efficiency when processing multiple masks as in the original paper.

2. A survey of post-hoc xai methods from a visualization perspective: Challenges and opportunities

URL: View paper

Brief Assessment

Post Hoc XAI[67] is a survey paper that categorizes existing XAI methods including gradient-based approaches, but does not propose novel gradient decomposition techniques for computational efficiency in explainability map generation.

3. Gradient based feature attribution in explainable ai: A technical review

URL: View paper

Brief Assessment

Gradient Feature Attribution[62] is a survey paper reviewing existing gradient-based explanation methods in XAI. It does not propose novel gradient decomposition techniques for efficiency improvements in explainability map computation, but rather categorizes and reviews existing methods in the field.

4. Decomposition and completion network for salient object detection

URL: View paper

Brief Assessment

Decomposition Completion Network[63] focuses on salient object detection using edge and skeleton decomposition for segmentation tasks, not gradient decomposition techniques for explainability map computation in vision-language models.

5. Interpretable basis decomposition for visual explanation

URL: View paper

Brief Assessment

Interpretable Basis Decomposition[61] focuses on decomposing neural activations into semantic components for visual explanation, not on gradient decomposition techniques for computational efficiency in explainability map generation as proposed in the original paper.

6. An Effective Infrared and Visible Image Fusion Approach via Rolling Guidance Filtering and Gradient Saliency Map

URL: View paper

Brief Assessment

Rolling Guidance Filtering[70] focuses on image fusion using gradient saliency maps for combining infrared and visible images, not on gradient decomposition for explainability map computation in vision-language models.

---

7. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition

URL: View paper

Brief Assessment

Detail Preserved Fusion[66] focuses on image fusion using gradient-based decomposition for saliency extraction in visible/infrared images, not on optimizing explainability maps for vision-language models or eliminating second-order derivatives in iterative mask embedding optimization.

---

8. Full-gradient representation for neural network visualization

URL: View paper

Brief Assessment

Full Gradient Representation[64] proposes gradient decomposition for neural network visualization by separating input-gradients and bias-gradients to satisfy completeness and weak dependence properties. The original paper's gradient decomposition focuses on eliminating second-order derivatives during iterative mask optimization for localized embeddings, which is a different technical context and application.

---

9. DecompX: Explaining Transformers Decisions by Propagating Token Decomposition

URL: View paper

Brief Assessment

DecompX[68] focuses on decomposing token representations through transformer layers for explainability, not on optimizing explainability map computation efficiency. The original paper's gradient decomposition eliminates second-order derivatives for mask optimization, while DecompX[68] propagates decomposed vectors through model components for attribution analysis.

---

10. DecomCAM: Advancing Beyond Saliency Maps through Decomposition and Integration

URL: View paper

Brief Assessment

DecomCAM[65] focuses on decomposition and integration techniques for saliency maps in general explainability contexts. The candidate paper's full text is not available (marked as 'n/a'), making it impossible to assess whether it addresses the specific gradient decomposition strategy proposed in the original paper for eliminating second-order derivatives during iterative mask processing.

Contribution 3: Regularization loss for balancing global and local representations

Description: The authors introduce an auxiliary regularization loss that forces the localized embedding token to remain close to the original global image embedding, enabling control over the trade-off between regional specificity and global context.

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. Conformer: Local Features Coupling Global Representations for Visual Recognition

URL: View paper

Brief Assessment

Conformer[58] addresses the balance between local and global features through architectural design (Feature Coupling Unit) rather than through a regularization loss during optimization. The approaches are fundamentally different in methodology.

---

2. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization

URL: View paper

Brief Assessment

DNGaussian[51] focuses on depth regularization for 3D Gaussian radiance fields in novel view synthesis, not on balancing global-local image representations through regularization losses. The technical domains and objectives are fundamentally different.

---

3. Learning Robust Global Representations by Penalizing Local Predictive Power

URL: View paper

Brief Assessment

Robust Global Representations[53] penalizes local predictive power to force models to learn global concepts, while the original paper uses regularization to keep localized embeddings close to global embeddings for context preservation. These are fundamentally different objectives and mechanisms.

---

4. SLN-RED: Regularization by Simultaneous Local and Nonlocal Denoising for Image Restoration

URL: View paper

Brief Assessment

SLN-RED[57] addresses image restoration using local and nonlocal denoisers in a regularization framework, which is fundamentally different from the original paper's vision-language embedding optimization that balances regional and global image context through an auxiliary loss.

---

5. Global-local spatially aware preserving projection for dimensionality reduction of hyperspectral images

URL: View paper

Brief Assessment

Global Local Spatially[59] focuses on dimensionality reduction for hyperspectral images using spatial preservation techniques, not on vision-language models or embedding optimization for image regions.

---

6. VICRegL: Self-Supervised Learning of Local Visual Features

URL: View paper

Brief Assessment

VICRegL[52] focuses on self-supervised learning with spatial matching of local features in vision tasks, not on balancing localized embeddings with global context through regularization as in the original paper's mask-based optimization approach.

---

7. Polyp Segmentation Using a Hybrid Vision Transformer and a Hybrid Loss Function

URL: View paper

Brief Assessment

Polyp Segmentation Hybrid[60] focuses on medical image segmentation using Vision Transformers for polyp detection, not on balancing global and local representations in vision-language models through regularization losses.

---

8. GaussEdit: Adaptive 3D Scene Editing With Text and Image Prompts

URL: View paper

Brief Assessment

GaussEdit[56] uses regularization to balance global scene coherence with local edits in 3D scene editing, not for vision-language embedding optimization. The technical contexts differ fundamentally: GaussEdit operates on 3D Gaussian representations for scene editing, while the original paper optimizes embedding tokens for vision-language alignment.

---

9. LPTMono: monocular depth estimation for underwater images using local perception transformer and globalâ□□local context fusion: D. Liu et al.

URL: View paper

Brief Assessment

LPTMono[54] focuses on monocular depth estimation for underwater images using transformer architectures. The candidate's loss function methodology differs fundamentally from the original paper's approach of aligning localized embeddings with global image representations through cosine similarity-based regularization.

---

10. Globalâ□□local consistent semi-supervised segmentation of histopathological image with different perturbations

URL: View paper

Brief Assessment

Global Local Histopathological[55] focuses on semi-supervised histopathological image segmentation with regularization for consistency across perturbations, not on balancing global and local image representations in vision-language models.

Appendix: Text Similarity Detection

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

References

- [0] MaskInversion: Localized Embeddings via Optimization of Explainability Maps View paper
- [1] Spatialrgpt: Grounded spatial reasoning in vision-language models View paper
- [2] Geochat: Grounded large vision-language model for remote sensing View paper
- [3] SpatialRGPT: Grounded Spatial Reasoning in Vision Language Model View paper
- [4] SpatialBot: Precise Spatial Understanding with Vision Language Models View paper
- [5] Spatial 3d-llm: Exploring spatial awareness in 3d vision-language models View paper
- [6] Glipv2: Unifying localization and vision-language understanding View paper
- [7] 3d aware region prompted vision language model View paper
- [8] Regiongpt: Towards region understanding vision language model View paper
- [9] Groma: Localized visual tokenization for grounding multimodal large language models View paper
- [10] Lare: Latent augmentation using regional embedding with vision-language model View paper
- [11] FarmSeg_VLM: A farmland remote sensing image segmentation method considering vision-language alignment View paper
- [12] Introducing Visual Perception Token into Multimodal Large Language Model View paper
- [13] Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks View paper
- [14] Contrastive Region Guidance: Improving Grounding in Vision-Language Models without Training View paper
- [15] Perceptual Grouping in Contrastive Vision-Language Models View paper
- [16] What's in the Image? A Deep-Dive into the Vision of Vision Language Models View paper
- [17] Visual-text cross alignment: Refining the similarity score in vision-language models View paper
- [18] Generate to Ground: Multimodal Text Conditioning Boosts Phrase Grounding in Medical Vision-Language Models View paper
- [19] Locality alignment improves vision-language models View paper
- [20] Regionclip: Region-based language-image pretraining View paper
- [21] Learning Visual Grounding from Generative Vision and Language Model View paper
- [22] AffordanceLLM: Grounding Affordance from Vision Language Models View paper
- [23] Grounding everything: Emerging localization properties in vision-language transformers View paper
- [24] SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities View paper
- [25] Self-chained image-language model for video localization and question answering View paper
- [26] HOICLIP: Efficient Knowledge Transfer for HOI Detection with Vision-Language Models View paper
- [27] Pixel-aligned language model View paper
- [28] GalLoP: Learning Global and Local Prompts for Vision-Language Models View paper
- [29] Seeing Out of tHe bOx: End-to-End Pre-training for Vision-Language Representation Learning View paper
- [30] Why Is Spatial Reasoning Hard for VLMs? An Attention Mechanism Perspective on Focus Areas View paper
- [31] RegionMed-CLIP: A Region-Aware Multimodal Contrastive Learning Pre-trained Model for Medical Image Understanding View paper
- [32] Aligning Bag of Regions for Open-Vocabulary Object Detection View paper
- [33] Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding View paper
- [34] Contrastive Localized Language-Image Pre-Training View paper
- [35] Overcoming the Pitfalls of Vision-Language Model for Image-Text Retrieval View paper
- [36] Fsvlm: A vision-language model for remote sensing farmland segmentation View paper

- [37] TextRegion: Text-Aligned Region Tokens from Frozen Image-Text Models View paper
- [38] Chain-of-region: Visual language models need details for diagram analysis View paper
- [39] Region-based cluster discrimination for visual representation learning View paper
- [40] MedGround-R1: Advancing Medical Image Grounding via Spatial-Semantic Rewarded Group Relative Policy Optimization View paper
- [41] GeoMag: A Vision-Language Model for Pixel-level Fine-Grained Remote Sensing Image Parsing View paper
- [42] u-llava: Unifying multi-modal tasks via large language model View paper
- [43] Toward Interactive Regional Understanding in Vision-Large Language Models View paper
- [44] Guiding Medical Vision-Language Models with Explicit Visual Prompts: Framework Design and Comprehensive Exploration of Prompt Variations View paper
- [45] Finecaption: Compositional image captioning focusing on wherever you want at any granularity View paper
- [46] Chain-of-Focus: Adaptive Visual Search and Zooming for Multimodal Reasoning via RL View paper
- [47] Guiding Medical Vision-Language Models with Diverse Visual Prompts: Framework Design and Comprehensive Exploration of Prompt Variations View paper
- [48] Describe anything model for visual question answering on text-rich images View paper
- [49] Multi-modal Self-perception Enhanced Large Language Model for 3D Region-of-Interest Captioning with Limited Data View paper
- [50] STOP: Integrated Spatial-Temporal Dynamic Prompting for Video Understanding View paper
- [51] Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization View paper
- [52] VICRegL: Self-Supervised Learning of Local Visual Features View paper
- [53] Learning Robust Global Representations by Penalizing Local Predictive Power View paper
- [54] LPTMono: monocular depth estimation for underwater images using local perception transformer and globalâ□□local context fusion: D. Liu et al. View paper
- [55] Globalâ□□local consistent semi-supervised segmentation of histopathological image with different perturbations View paper
- [56] GaussEdit: Adaptive 3D Scene Editing With Text and Image Prompts View paper
- [57] SLN-RED: Regularization by Simultaneous Local and Nonlocal Denoising for Image Restoration View paper
- [58] Conformer: Local Features Coupling Global Representations for Visual Recognition View paper
- [59] Global-local spatially aware preserving projection for dimensionality reduction of hyperspectral images View paper
- [60] Polyp Segmentation Using a Hybrid Vision Transformer and a Hybrid Loss Function View paper
- [61] Interpretable basis decomposition for visual explanation View paper
- [62] Gradient based feature attribution in explainable ai: A technical review View paper
- [63] Decomposition and completion network for salient object detection View paper
- [64] Full-gradient representation for neural network visualization View paper
- [65] DecomCAM: Advancing Beyond Saliency Maps through Decomposition and Integration View paper
- [66] Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition View paper
- [67] A survey of post-hoc xai methods from a visualization perspective: Challenges and opportunities View paper
- [68] DecompX: Explaining Transformers Decisions by Propagating Token Decomposition View paper
- [69] Deeply Explain CNN Via Hierarchical Decomposition View paper
- [70] An Effective Infrared and Visible Image Fusion Approach via Rolling Guidance Filtering and Gradient Saliency Map View paper
- [71] Interpreting CLIP's Image Representation via Text-Based Decomposition View paper
- [72] Interpretable representations in explainable AI: from theory to practice View paper
- [73] Explainable AI enhanced transformer based UNet for medical images segmentation using gradient weighted class activation map View paper
- [74] Finding Regions of Counterfactual Explanations via Robust Optimization View paper
- [75] CEIR: Concept-based Explainable Image Representation Learning View paper
- [76] ICEv2: Interpretability, Comprehensiveness, and Explainability in Vision Transformer. View paper
- [77] Artxai: Explainable artificial intelligence curates deep representation learning for artistic images using fuzzy techniques View paper
- [78] Explainable self-supervised learning for medical image diagnosis based on DINO V2 model and semantic search View paper
- [79] Explainable artificial intelligence (XAI) for deep learning based medical imaging classification View paper
- [80] Local Concept Embeddings for Analysis of Concept Distributions in DNN Feature Spaces View paper