Novelty Assessment Report

Paper: Beginning with You: Perceptual-Initialization Improves Vision-Language Representation and Alignment
PDF URL: https://openreview.net/pdf?id=AbmOOdWwYD
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

We introduce Perceptual-Initialization (PI), a paradigm shift in visual representation learning that incorporates human perceptual structure during the initialization phase rather than as a downstream fine-tuning step. By integrating human-derived triplet embeddings from the NIGHTS dataset to initialize a CLIP vision encoder, followed by self-supervised learning on YFCC15M, our approach demonstrates significant zero-shot performance improvements without any task-specific fine-tuning across 29 zero shot classification and 2 retrieval benchmarks. On ImageNet-1K, zero-shot gains emerge after approximately 15 epochs of pretraining. Benefits are observed across datasets of various scales, with improvements manifesting at different stages of the pretraining process depending on dataset characteristics. Our approach consistently enhances zero-shot top-1 accuracy, top-5 accuracy, and retrieval recall (e.g., R@1, R@5) across these diverse evaluation tasks, without requiring any adaptation to target domains. These findings challenge the conventional wisdom of using human-perceptual data primarily for fine-tuning and demonstrate that embedding human perceptual structure during early representation learning yields more capable and vision-language aligned systems that generalize immediately to unseen tasks. Our work shows that "beginning with you", starting with human perception, provides a stronger foundation for general-purpose vision-language intelligence.

Core Task Landscape

This paper addresses: **Incorporating Human Perceptual Structure into Vision-Language Model Initialization**

A total of **25 papers** were analyzed and organized into a taxonomy with **9 categories**.

Taxonomy Overview

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

- **Perceptual Structure Integration Methods**
- **Vision-Language Model Architectures and Capabilities**
- **Perceptual Alignment and Quality Assessment**
- **Application Domains and Use Cases**

Complete Taxonomy Tree

- Incorporating Human Perceptual Structure into Vision-Language Model Initialization Survey Taxonomy
- Perceptual Structure Integration Methods
  - Initialization-Phase Perceptual Embedding ★ (2 papers)
  - [0] Beginning with You: Perceptual-Initialization Improves Vision-Language Representation and Alignment (Anon et al., 2026) View paper
  - [2] Perceptual grouping in vision-language models (K Ranasinghe, 2022) View paper
  - Contrastive Pretraining with Perceptual Data (3 papers)
  - [4] Human-CLAP: Human-perception-based contrastive language-audio pretraining (Okamoto, 2025) View paper
  - [13] Versatile multi-modal pre-training for human-centric perception (Fangzhou Hong, 2022) View paper
  - [23] MENTOR: Human Perception-Guided Pretraining for Increased Generalization (Colton R. Crum, 2023) View paper
  - Multimodal Fusion and Cross-Modal Learning (4 papers)
  - [11] CM-ASAP: Cross-Modality Adaptive Sensing and Perception for Efficient Hand Gesture Recognition (Soheil Hor, 2024) View paper
  - [14] AllSpark: A Multimodal Spatiotemporal General Intelligence Model With Ten Modalities via Language as a Reference Framework (Run Shao, 2023) View paper
  - [15] Supervised Learning With Perceptual Similarity for Multimodal Gene Expression Registration of a Mouse Brain Atlas (Jan Krepl, 2021) View paper
  - [25] Modeling Development of Multimodal Emotion Perception Guided by Tactile Dominance and Perceptual Improvement (Takato Horii, 2018) View paper
- Vision-Language Model Architectures and Capabilities
  - Multimodal Large Language Models (2 papers)
  - [1] Language is not all you need: Aligning perception with language models (Huang, 2023) View paper
  - [3] Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution (Wang Peng, 2024) View paper
  - Task-Specific Architectures (3 papers)
  - [5] Draw with Thought: Unleashing Multimodal Reasoning for Scientific Diagram Generation (Zhiqing Cui, 2025) View paper
  - [8] MemVLT: Vision-Language Tracking with Adaptive Memory-based Prompts (Xiaotang Chen, 2024) View paper
  - [9] Language-aware Visual Semantic Distillation for Video Question Answering (Bo Zou, 2024) View paper
- Perceptual Alignment and Quality Assessment
  - Visual Quality Perception (3 papers)
  - [7] VisualCritic: Making LMMs Perceive Visual Quality Like Humans (Huang Zhi-peng, 2024) View paper

- [12] Exploring Human Perception-Aligned Perceptual Hashing (Patrick De Smet, 2026) View paper
- [21] VLIC: Vision-Language Models As Perceptual Judges for Human-Aligned Image Compression (Kyle Sargent, 2025) View paper
- Semantic Alignment and Interpretability (3 papers)
- [6] Large Language Models estimate fine-grained human color-concept associations (Mukherjee, 2024) View paper
- [16] Aligning Multi-Modal Object Representations to Human Cognition (Yu, 2025) View paper
- [18] Interpretable Zero-Shot Learning with Locally-Aligned Vision-Language Model (Chen Shi-ming, 2025) View paper

• Application Domains and Use Cases
- Autonomous Systems and Spatial Understanding (2 papers)
- [10] OmniScene: Attention-Augmented Multimodal 4D Scene Understanding for Autonomous Driving (Liu Pei, 2025) View paper
- [17] Urban Street-Scene Perception and Renewal Strategies Powered by Visionâ□□Language Models (Y Yao, 2025) View paper
- Human-Centered Interaction and Assessment (4 papers)
- [19] Multi-modal anchoring for humanâ□□robot interaction (J. Fritsch, 2003) View paper
- [20] RecompGPT: Generative Pre-Trained Transformers-Assisted Interactive Human Gaze Pattern Learning and Distribution Modeling for Scene Recomposition (Wang Shang, 2025) View paper
- [22] Human-AI Alignment of Multimodal Large Language Models with Speech-Language Pathologists in Parent-Child Interactions (Shi, 2025) View paper
- [24] Joint Visual and Text Prompting for Zero-Shot Object-Oriented Perception with Multimodal Large Language Models (Songtao Jiang, 2024) View paper

Narrative

Core task: Incorporating human perceptual structure into vision-language model initialization. The field organizes around four main branches that reflect distinct stages and concerns in building perceptually grounded multimodal systems. Perceptual Structure Integration Methods explore how human-like perceptual cues—ranging from grouping principles (Perceptual Grouping VLM[2]) to similarity metrics (Perceptual Similarity Registration[15])—can be embedded during model initialization or training. Vision-Language Model Architectures and Capabilities address the design of multimodal backbones and their representational power, spanning general-purpose frameworks (Qwen2-VL[3], Versatile Multimodal Pretraining[13]) and specialized mechanisms for memory or reasoning (MemVLT[8], RecompGPT[20]). Perceptual Alignment and Quality Assessment focuses on measuring and enforcing correspondence between model outputs and human judgments, including alignment strategies (Aligning Perception Language[1], Aligning Human Cognition[16]) and quality evaluation (VisualCritic[7]). Application Domains and Use Cases demonstrate how perceptual grounding benefits downstream tasks such as urban scene understanding (Urban Scene Perception[17]), creative generation (Draw with Thought[5]), and cross-modal retrieval (Human-CLAP[4]).

Several active lines of work highlight trade-offs between early-stage perceptual embedding and post-hoc alignment. Some studies inject perceptual priors directly at initialization (Perceptual Initialization[0], Perceptual Grouping VLM[2]), aiming to shape the learned representation space from the outset, while others refine alignment through distillation or feedback loops after pretraining (Language Visual Distillation[9], MENTOR[23]). Perceptual Initialization[0] sits squarely within the Initialization-Phase Perceptual Embedding cluster, emphasizing the integration of human perceptual structure before large-scale training begins. This contrasts with approaches like Aligning Perception Language[1] or Aligning Human Cognition[16], which typically adjust pretrained models to better match human judgments. By anchoring perceptual cues early, Perceptual Initialization[0] seeks to reduce the gap between machine and human vision from the ground up, complementing neighbor work (Perceptual Grouping VLM[2]) that similarly leverages Gestalt-like grouping at the model's foundation.

Related Works in Same Category

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

1. Perceptual grouping in vision-language models

Authors: K Ranasinghe, B McKinzie, S Ravi, Y Yang, AT Toshev | **Year/Venue**: 2022 | **URL**: View paper

Abstract

â□¦ The topic of perceptual grouping has a long, rich history in human visual â□¦ investigate initializing the image model with several methods. First, we investigate initializing the image model â□¦

Relationship Analysis

Both papers belong to the Initialization-Phase Perceptual Embedding category, focusing on incorporating human perceptual structure during model initialization before primary training. The original paper initializes a CLIP vision encoder with human triplet judgments from NIGHTS before contrastive learning on YFCC15M, while the candidate paper modifies vision-language models through architectural changes (aggregation methods and pretraining strategies) to achieve perceptual grouping for segmentation tasks. The key difference is that the original paper explicitly uses human perceptual data (NIGHTS triplets) for initialization, whereas the candidate paper achieves perceptual alignment through architectural modifications and self-supervised pretraining without direct human perceptual data injection at initialization.

Contributions Analysis

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

Contribution 1: Perceptual-Initialization paradigm for vision-language models

Description: The authors propose a new training paradigm that integrates human perceptual structure at the initialization stage of model training, rather than applying it as a post-hoc fine-tuning step. This approach uses human-derived triplet embeddings from the NIGHTS dataset to initialize a CLIP vision encoder before self-supervised learning.

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. Towards model-based recognition of human movements in image sequences

URL: View paper

Brief Assessment

Model Based Recognition[32] focuses on model-based recognition of human movements in image sequences using 3D descriptions, not on initializing vision-language models with human perceptual structure for contrastive learning.

2. POV Learning: Individual Alignment of Multimodal Models using Human Perception

URL: View paper

Brief Assessment

POV Learning[28] focuses on individual-level alignment using eye-tracking data during inference/evaluation, not on initializing vision encoders with human perceptual structure before pre-training. The candidate addresses a fundamentally different problem of personalizing predictions to individual users rather than improving general zero-shot performance through perceptual initialization.

---

3. ScanDMM: A Deep Markov Model of Scanpath Prediction for 360Â° Images

URL: View paper

Brief Assessment

ScanDMM[29] focuses on scanpath prediction for 360° images using a Deep Markov Model architecture, not on vision-language model initialization or incorporating human perceptual structure during model training initialization.

---

4. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution

URL: View paper

Brief Assessment

Qwen2-VL[3] focuses on dynamic resolution processing and multimodal position embeddings for vision-language models, not on incorporating human perceptual structure during initialization. The candidate does not address initialization with human-derived embeddings or perceptual priors.

---

5. Perceptual quality assessment for no-reference image via optimization-based meta-learning

URL: View paper

Brief Assessment

Meta Learning Quality[30] focuses on no-reference image quality assessment using meta-learning for weight initialization and optimization rules, not on incorporating human perceptual structure during vision-language model initialization.

---

6. Vision-Aware Text Features in Referring Image Segmentation: From Object Understanding to Context Understanding

URL: View paper

Brief Assessment

Vision Aware Text[27] focuses on referring image segmentation using CLIP priors for object localization, not on perceptual initialization during model training. The candidate applies pre-trained CLIP features as a module within a task-specific architecture, rather than initializing vision encoders with human perceptual structure before self-supervised learning.

---

7. Fixating on Attention: Integrating Human Eye Tracking into Vision Transformers

URL: View paper

Brief Assessment

Fixating Attention[31] focuses on integrating human eye-tracking data into vision transformers for driving decision tasks, not on initializing vision-language models with human perceptual structure before contrastive learning.

---

8. MENTOR: Human Perception-Guided Pretraining for Increased Generalization

URL: View paper

Brief Assessment

MENTOR[23] focuses on autoencoder-based pretraining for CNNs in anomaly detection tasks (iris PAD, synthetic face detection, chest X-ray diagnosis), not vision-language models like CLIP. The original paper initializes CLIP's vision encoder with human triplet embeddings before contrastive image-text pretraining, while MENTOR[23] trains autoencoders to reconstruct human saliency maps before classification tasks without any language component.

---

9. Perceptual Inductive Bias Is What You Need Before Contrastive Learning

URL: View paper

Brief Assessment

Perceptual Inductive Bias[26] focuses on incorporating perceptual constructs (shape prototypes, intrinsic images) as additional views during contrastive learning itself, not as an initialization step before large-scale pretraining. The original paper's novelty lies in using human triplet judgments to initialize model weights before web-scale training, which is architecturally and methodologically distinct.

---

10. GMODiff: One-Step Gain Map Refinement with Diffusion Priors for HDR Reconstruction

URL: View paper

Brief Assessment

GMODiff[33] focuses on HDR image reconstruction using diffusion models for gain map estimation, not on vision-language model initialization or incorporating human perceptual structure during model training.

Contribution 2: Two-stage training pipeline with human perceptual initialization

Description: The method consists of two sequential stages: first initializing the vision encoder by training on human similarity judgments from NIGHTS, then performing conventional large-scale contrastive pretraining on 15M image-text pairs from YFCC15M. This converts random initialization into perceptual initialization.

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. Liftedcl: Lifting contrastive learning for human-centric perception

URL: View paper

Brief Assessment

LiftedCL[36] uses a two-stage pipeline but focuses on 3D human structure information for human-centric perception tasks (pose estimation, shape recovery, parsing), not human perceptual similarity judgments from datasets like NIGHTS for vision-language alignment.

---

2. Perceptual Inductive Bias Is What You Need Before Contrastive Learning

URL: View paper

Brief Assessment

Perceptual Inductive Bias[26] proposes a pre-pretraining stage using perceptual constructs as augmented views within contrastive learning, not a two-stage pipeline where Stage 1 initializes weights with human similarity judgments (NIGHTS triplets) followed by Stage 2 web-scale pretraining. The training paradigms differ fundamentally in structure and data sources.

---

3. JSQA: Speech Quality Assessment with Perceptually-Inspired Contrastive Pretraining Based on JND Audio Pairs

URL: View paper

Prior Art Analysis

JSQA[37] demonstrates that a two-stage training pipeline incorporating human perceptual information before large-scale training was already established in the speech quality assessment domain. The candidate paper explicitly describes a two-stage framework where an encoder is first pretrained using perceptually-guided contrastive learning on just noticeable difference (JND) pairs, followed by fine-tuning for prediction tasks. This approach directly parallels the original paper's claimed novelty of initializing with human perceptual structure before contrastive pretraining, showing that the concept of perceptual initialization followed by standard training was not novel to the original work.

Evidence

Evidence 1 - **Rationale**: Both papers describe a two-stage approach where human perceptual information (triplet embeddings vs. JND pairs) is used to initialize/pretrain an encoder before subsequent training on a larger dataset. This demonstrates that the concept of perceptual initialization before standard training existed prior to the original work. - **Original**: we introduceperceptual -initialization(pi), a paradigm shift in visual representation learning that incorporates human perceptual structureduring the initialization phaserather than as a downstream fine -tuning step. by integrating human-derived triplet embeddings from the nights dataset to initiali... - **Candidate**: we propose jsqa, a two-stage framework that pretrains an audio encoder using perceptually-guided contrastive learning on just noticeable difference (jnd) pairs, followed by fine-tuning for mos prediction. we first generate pairs of audio data within jnd levels, which are then used to pretrain an enc...

Evidence 2 - **Rationale**: Both papers explicitly describe sequential stages where perceptual information is incorporated first, followed by task-specific training. The candidate demonstrates that perceptually-inspired pretraining improves performance over training from scratch, establishing this methodology before the original paper's submission. - **Original**: this paradigm consists of two sequential stages: first, initializing the vision encoder by training it on human similarity judgments, followed by a second stage of conventional large-scale contrastive pretraining on image-text pairs from the web. - **Candidate**: the encoder is later fine-tuned with audio samples from the nisqa dataset for mos prediction. experimental results suggest that perceptually-inspired contrastive pretraining significantly improves the model performance evaluated by various metrics when compared against the same network trained from ...

Evidence 3 - **Rationale**: The original paper claims to transform random initialization into perceptual initialization as a novel contribution. However, JSQA[37] already implements this concept by pretraining with perceptually-guided contrastive learning before fine-tuning, demonstrating that replacing random initialization with perceptual pretraining was an established approach. - **Original**: by transforming random seeds into perceptual seeds, we convert an often ignored source of variance into a principled inductive bias and set the trajectory of representation learning on a more human -aligned course from the very first gradient step. - **Candidate**: to this end, we propose jsqa, a two-stage framework that pretrains an audio encoder using perceptually-guided contrastive learning on just noticeable difference (jnd) pairs, followed by fine-tuning for mos prediction.

Contribution 3: First approach using human triplet judgments for vision-language model initialization

Description: The authors claim this is the first work to directly integrate supervised human perceptual data into the initialization of vision-language models before web-scale training, distinguishing it from prior work that applied human perceptual alignment only as post-hoc fine-tuning.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Aligning machine and human visual representations across abstraction levels

URL: View paper

Brief Assessment

Machine Human Alignment[34] applies human perceptual alignment as post-hoc fine-tuning of pretrained models, not as initialization before web-scale training. The paper explicitly states 'Instead of applying human perceptual alignment as a post-hoc fine-tuning step, our approach integrates human perceptual judgments at the initial stage of representation learning' but this refers to fine-tuning pretrained models with alignet, not initializing random weights before contrastive pretraining as in the original paper.

---

2. When does perceptual alignment benefit vision representations?

URL: View paper

Brief Assessment

Perceptual Alignment Benefits[35] applies human perceptual alignment as fine-tuning on pre-trained models (CLIP, DINO, DINOv2), not as initialization before web-scale training. The candidate explicitly states they 'propose to use the method described below as a "second pretraining stage"' and 'finetune state-of-the-art models on human similarity judgments', which is fundamentally different from the original paper's initialization-first paradigm.

Appendix: Text Similarity Detection

Textual similarity detection checked 15 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. When does perceptual alignment benefit vision representations?

Detected in: Contribution: contribution_3

⚠ **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] Beginning with You: Perceptual-Initialization Improves Vision-Language Representation and Alignment View paper
- [1] Language is not all you need: Aligning perception with language models View paper
- [2] Perceptual grouping in vision-language models View paper
- [3] Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution View paper

- [4] Human-CLAP: Human-perception-based contrastive language-audio pretraining View paper
- [5] Draw with Thought: Unleashing Multimodal Reasoning for Scientific Diagram Generation View paper
- [6] Large Language Models estimate fine-grained human color-concept associations View paper
- [7] VisualCritic: Making LMMs Perceive Visual Quality Like Humans View paper
- [8] MemVLT: Vision-Language Tracking with Adaptive Memory-based Prompts View paper
- [9] Language-aware Visual Semantic Distillation for Video Question Answering View paper
- [10] OmniScene: Attention-Augmented Multimodal 4D Scene Understanding for Autonomous Driving View paper
- [11] CM-ASAP: Cross-Modality Adaptive Sensing and Perception for Efficient Hand Gesture Recognition View paper
- [12] Exploring Human Perception-Aligned Perceptual Hashing View paper
- [13] Versatile multi-modal pre-training for human-centric perception View paper
- [14] AllSpark: A Multimodal Spatiotemporal General Intelligence Model With Ten Modalities via Language as a Reference Framework View paper
- [15] Supervised Learning With Perceptual Similarity for Multimodal Gene Expression Registration of a Mouse Brain Atlas View paper
- [16] Aligning Multi-Modal Object Representations to Human Cognition View paper
- [17] Urban Street-Scene Perception and Renewal Strategies Powered by Visionâ□□Language Models View paper
- [18] Interpretable Zero-Shot Learning with Locally-Aligned Vision-Language Model View paper
- [19] Multi-modal anchoring for humanâ□□robot interaction View paper
- [20] RecompGPT: Generative Pre-Trained Transformers-Assisted Interactive Human Gaze Pattern Learning and Distribution Modeling for Scene Recomposition View paper
- [21] VLIC: Vision-Language Models As Perceptual Judges for Human-Aligned Image Compression View paper
- [22] Human-AI Alignment of Multimodal Large Language Models with Speech-Language Pathologists in Parent-Child Interactions View paper
- [23] MENTOR: Human Perception-Guided Pretraining for Increased Generalization View paper
- [24] Joint Visual and Text Prompting for Zero-Shot Object-Oriented Perception with Multimodal Large Language Models View paper
- [25] Modeling Development of Multimodal Emotion Perception Guided by Tactile Dominance and Perceptual Improvement View paper
- [26] Perceptual Inductive Bias Is What You Need Before Contrastive Learning View paper
- [27] Vision-Aware Text Features in Referring Image Segmentation: From Object Understanding to Context Understanding View paper
- [28] POV Learning: Individual Alignment of Multimodal Models using Human Perception View paper
- [29] ScanDMM: A Deep Markov Model of Scanpath Prediction for 360Â° Images View paper
- [30] Perceptual quality assessment for no-reference image via optimization-based meta-learning View paper
- [31] Fixating on Attention: Integrating Human Eye Tracking into Vision Transformers View paper
- [32] Towards model-based recognition of human movements in image sequences View paper
- [33] GMODiff: One-Step Gain Map Refinement with Diffusion Priors for HDR Reconstruction View paper
- [34] Aligning machine and human visual representations across abstraction levels View paper
- [35] When does perceptual alignment benefit vision representations? View paper
- [36] Liftedcl: Lifting contrastive learning for human-centric perception View paper
- [37] JSQA: Speech Quality Assessment with Perceptually-Inspired Contrastive Pretraining Based on JND Audio Pairs View paper