Novelty Assessment Report

Paper: Synergizing Understanding and Generation with Interleaved Analyzing-Drafting Thinking

PDF URL: https://openreview.net/pdf?id=GtqmPJf00A

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-30

Abstract

Unified Vision–Language Models (UVLMs) aim to advance multimodal learning by supporting both understanding and generation within a single framework. However, existing approaches largely focus on architectural unification while overlooking the need for explicit interaction between the two capabilities during task solving. As a result, current models treat understanding and generation as parallel skills rather than synergistic processes. To achieve real synergy, we introduce the interleaved Analyzing–Drafting problem-solving loop (AD-Loop), a new think paradigm that dynamically alternates between analytic and drafting operations. By interleaving textual thoughts with visual thoughts, AD-Loop enables models to iteratively refine both comprehension and outputs, fostering genuine synergy. To train this mechanism, we design a two-stage strategy: supervised learning on interleaved thought data to initialize alternation, followed by reinforcement learning to promote adaptive and autonomous control. Extensive experiments demonstrate that AD-Loop consistently improves performance across standard benchmarks for both understanding and generation, with strong transferability to various UVLMs architectures. Visual analyses further validate the effectiveness of implicit visual thoughts. These results highlight AD-Loop as a principled and broadly applicable strategy for synergizing comprehension and creation. Code and model will be available.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Synergizing Multimodal Understanding and Generation through Interleaved Reasoning**

A total of **50 papers** were analyzed and organized into a taxonomy with **26 categories**.

Taxonomy Overview

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

- **Unified Multimodal Model Architectures**
- **Interleaved Reasoning Paradigms**
- **Training Strategies and Data Construction**
- **Evaluation Benchmarks and Metrics**
- **Domain-Specific Applications**
- **Specialized Technical Components**
- **Analysis and Robustness Studies**
- **Instruction Following and Interactive Systems**

Complete Taxonomy Tree

- Synergizing Multimodal Understanding and Generation through Interleaved Reasoning Survey Taxonomy
- Unified Multimodal Model Architectures
  - Native Autoregressive Unified Models (4 papers)
  - [1] Dreamllm: Synergistic multimodal comprehension and creation (Dong, 2023) View paper
  - [2] Emerging properties in unified multimodal pretraining (Deng, 2025) View paper
  - [11] ANOLE: An Open, Autoregressive, Native Large Multimodal Models for Interleaved Image-Text Generation (Ethan Chern, 2024) View paper
  - [15] Mogao: An omni foundation model for interleaved multi-modal generation (Liao Chao, 2025) View paper
  - Diffusion-Based Unified Models (2 papers)
  - [19] Lavida-O: Elastic Large Masked Diffusion Models for Unified Multimodal Understanding and Generation (Li Shu-Fan, 2025) View paper
  - [44] OneFlow: Concurrent Mixed-Modal and Interleaved Generation with Edit Flows (Nguyen John, 2025) View paper
  - Adapter-Based Extension Frameworks (3 papers)
  - [20] ARMOR: Empowering Multimodal Understanding Model with Interleaved Multimodal Generation Capability (Sun Jian-wen, 2025) View paper
  - [28] MChat: Empowering VLM for Multimodal LLM Interleaved Text-Image Generation (X Chi, 2023) View paper
  - [43] IUT-Plug: A Plug-in tool for Interleaved Image-Text Generation (Li Xingxing, 2025) View paper
  - Modality-Specialized Expert Architectures (2 papers)
  - [18] Lateralization lora: Interleaved instruction tuning with modality-specialized adaptations (Xu Zhiyang, 2024) View paper
  - [21] Modality-Specialized Synergizers for Interleaved Vision-Language Generalists (Xu Zhiyang, 2025) View paper
- Interleaved Reasoning Paradigms
  - Text-Image Interleaved Chain-of-Thought ★ (4 papers)
  - [0] Synergizing Understanding and Generation with Interleaved Analyzing-Drafting Thinking (Anon et al., 2026) View paper
  - [7] Interleaved-modal chain-of-thought (Jun Gao, 2025) View paper
  - [16] ThinkMorph: Emergent Properties in Multimodal Interleaved Chain-of-Thought Reasoning (Gu Jiawei, 2025) View paper
  - [35] Interleaving Reasoning for Better Text-to-Image Generation (Huang Wenxuan, 2025) View paper

    - Latent Visual Reasoning (2 papers)
    - [30] Reasoning Within the Mind: Dynamic Multimodal Interleaving in Latent Space (Chengzhi Liu, 2025) View paper
    - [33] Interleaved Latent Visual Reasoning with Selective Perceptual Modeling (Shuai Dong, 2025) View paper
    - Video-Based Interleaved Reasoning (2 papers)
    - [36] Thinking with Video: Video Generation as a Promising Multimodal Reasoning Paradigm (Jingqi Tong, 2025) View paper
    - [46] TV2TV: A Unified Framework for Interleaved Language and Video Generation (Xiaochuang Han, 2025) View paper
    - Tool-Augmented Reasoning Frameworks (2 papers)
    - [25] Skywork-R1V4: Toward Agentic Multimodal Intelligence through Interleaved Thinking with Images and DeepResearch (Yifan Zhang, 2025) View paper
    - [41] LLM-I: LLMs are Naturally Interleaved Multimodal Creators (Zhang Feng, 2025) View paper
    - Draft-and-Refine Reasoning (2 papers)
    - [9] Generative universal verifier as multimodal meta-reasoner (Zhang, 2025) View paper
    - [47] DraCo: Draft as CoT for Text-to-Image Preview and Rare Concept Generation (Dongzhi Jiang, 2025) View paper
- Training Strategies and Data Construction
    - Reinforcement Learning for Interleaved Tasks (3 papers)
    - [26] CX-Mind: A Pioneering Multimodal Large Language Model for Interleaved Reasoning in Chest X-ray via Curriculum-Guided Reinforcement Learning (Li, 2025) View paper
    - [32] Deep thinking in text-to-image generation using unified model with reinforcement learning (Hu, 2025) View paper
    - [38] IRG-MotionLLM: Interleaving Motion Generation, Assessment and Refinement for Text-to-Motion Generation (Yuan-Ming Li, 2025) View paper
    - High-Quality Interleaved Dataset Construction (3 papers)
    - [17] RealSyn: An Effective and Scalable Multimodal Interleaved Document Transformation Paradigm (Tiancheng Gu, 2025) View paper
    - [23] CoMM: A Coherent Interleaved Image-Text Dataset for Multimodal Understanding and Generation (Wei Chen, 2024) View paper
    - [37] A High-Quality Dataset and Reliable Evaluation for Interleaved Image-Text Generation (Feng Yu-kang, 2025) View paper
    - Multi-Stage and Curriculum Training (2 papers)
    - [14] Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer (Tian, 2024) View paper
    - [24] LVLM-MIR: Large Vision-Language Model with Parameter-Efficient Fine-Tuning for Multimodal Interleaved Reasoning (Jun Yu, 2025) View paper
- Evaluation Benchmarks and Metrics
    - Comprehensive Interleaved Benchmarks (3 papers)
    - [12] Mmie: Massive multimodal interleaved comprehension benchmark for large vision-language models (Xia, 2024) View paper
    - [27] MIRAGE25: ACM MM25 Multimodal Interleaved Reasoning and Generation Challenge (Dong Chen, 2025) View paper
    - [39] WEAVE: Unleashing and Benchmarking the In-context Interleaved Comprehension and Generation (Wei Chow, 2025) View paper
    - Cross-Modal Reasoning Evaluation (1 papers)
    - [45] ROVER: Benchmarking Reciprocal Cross-Modal Reasoning for Omnimodal Generation (Liang Yongyuan, 2025) View paper
- Domain-Specific Applications
    - Medical Imaging and Clinical Reasoning (1 papers)
    - [49] Multimodal Clinical Reasoning through Knowledge-augmented Rationale Generation (Niu Shuai, 2024) View paper
    - Graph-Based Multimodal Reasoning (1 papers)
    - [4] Graphgpt-o: Synergistic multimodal comprehension and generation on graphs (Yi Fang, 2025) View paper
    - Video Understanding and Narrative Reasoning (2 papers)
    - [29] Contextual AD Narration with Interleaved Multimodal Sequence (Hanlin Wang, 2025) View paper
    - [42] Learning Video Context as Interleaved Multimodal Sequences (Lin, 2024) View paper
    - Remote Sensing and Geospatial Analysis (1 papers)
    - [34] VICoT-Agent: A Vision-Interleaved Chain-of-Thought Framework for Interpretable Multimodal Reasoning and Scalable Remote Sensing Analysis (Chujie Wang, 2025) View paper
    - Speech and Gesture Synthesis (2 papers)
    - [13] VITA-Audio: Fast Interleaved Cross-Modal Token Generation for Efficient Large Speech-Language Model (Shen, 2025) View paper
    - [22] Gelina: Unified Speech and Gesture Synthesis via Interleaved Token Prediction (TÃ©o Guichoux, 2025) View paper
    - Creative Content Generation (1 papers)
    - [40] HUMORCHAIN: Theory-Guided Multi-Stage Reasoning for Interpretable Multimodal Humor Generation (Jiajun Zhang, 2025) View paper
- Specialized Technical Components
    - Multi-Image and Multi-Frame Processing (1 papers)
    - [5] Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models (Li Feng, 2024) View paper
    - Representation Alignment and Fusion (1 papers)
    - [6] Multimodal representation alignment for image generation: Text-image interleaved control is easier than you think (Chen Liang, 2025) View paper
    - Information Retrieval for Interleaved Documents (1 papers)
    - [8] Interleaved multi-modal document representations for large-scale information retrieval using large language models (Dominic Rixewa, 2024) View paper
- Analysis and Robustness Studies
    - Adversarial Robustness and Security (1 papers)
    - [31] STaR-Attack: A Spatio-Temporal and Narrative Reasoning Attack Framework for Unified Multimodal Understanding and Generation Models (Guo Shao-xiong, 2025) View paper
    - Survey and Taxonomy Studies (1 papers)
    - [3] A Survey of Unified Multimodal Understanding and Generation: Advances and Challenges (Yan Yang, 2025) View paper
- Instruction Following and Interactive Systems (3 papers)
    - [10] Codi-2: In-context interleaved and interactive any-to-any generation (Zineng Tang, 2024) View paper

- ◦ [48] Multimodal linguistic inference (Michael Moortgat, 1995) View paper
- ◦ [50] TextBind: Multi-turn Interleaved Multimodal Instruction-following in the Wild (Li, 2023) View paper

Narrative

Core task: Synergizing multimodal understanding and generation through interleaved reasoning. The field has evolved around the challenge of building models that can seamlessly integrate perception and creation across modalities—text, images, video, and audio—while maintaining coherent reasoning throughout. The taxonomy reveals several complementary directions: Unified Multimodal Model Architectures (e.g., Dreamllm[1], ANOLE[11]) focus on designing end-to-end systems that handle multiple modalities within a single framework, while Interleaved Reasoning Paradigms explore how to structure step-by-step multimodal thought processes, including text-image chain-of-thought approaches. Training Strategies and Data Construction address the practical challenges of curating interleaved datasets (MM Interleaved[14], CoMM Dataset[23]) and developing effective learning objectives. Evaluation Benchmarks and Metrics (MMIE Benchmark[12], MIRAGE Challenge[27]) provide standardized assessments, while Domain-Specific Applications and Instruction Following systems demonstrate how these capabilities translate to real-world interactive scenarios. Specialized Technical Components and Analysis studies examine architectural details and robustness properties that underpin reliable multimodal reasoning.

A particularly active line of work centers on interleaved chain-of-thought methods that alternate between textual reasoning steps and visual generation or analysis. Interleaved Analyzing Drafting[0] exemplifies this paradigm by proposing a framework where understanding and generation phases are tightly coupled through intermediate reasoning traces. This approach contrasts with works like Interleaved Modal CoT[7] and ThinkMorph[16], which emphasize different granularities of modal interleaving—some focusing on fine-grained step-by-step transitions, others on coarser analyze-then-generate pipelines. Interleaving Reasoning Generation[35] explores similar territory but with distinct emphases on how reasoning tokens guide subsequent generative steps. The original paper sits squarely within this text-image interleaved reasoning cluster, sharing with neighbors like ThinkMorph[16] a commitment to explicit intermediate reasoning, while differing in how tightly the analyzing and drafting phases are synchronized and whether generation occurs incrementally or in discrete bursts.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. Interleaved-modal chain-of-thought

Authors: Jun Gao, Yongqi Li, Ziqiang Cao, Yongqing Li, Wenjie Li | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Chain-of-Thought (CoT) prompting elicits large language models (LLMs) to produce a series of intermediate reasoning steps before arriving at the final answer. However, when transitioning to vision-language models (VLMs), their text-only rationales struggle to express the fine-grained associations with the original image. In this paper, we propose an image-incorporated multimodal Chain-of-Thought, named Interleaved-modal Chain-of-Thought (ICoT), which generates sequential reasoning steps consisti...

Relationship Analysis

Both papers belong to the Text-Image Interleaved Chain-of-Thought category, generating sequential reasoning steps with paired visual and textual rationales. They overlap in their core approach of interleaving visual and textual reasoning to enhance multimodal understanding and generation, with both proposing mechanisms to alternate between analysis and visual drafting. The key difference is that the original paper introduces the AD-Loop paradigm with a two-stage training strategy (supervised learning followed by reinforcement learning) for unified vision-language models, while the candidate paper (ICoT) focuses on an attention-driven selection mechanism that is plug-and-play and training-free, selecting patches from input images rather than generating new visual content.

2. ThinkMorph: Emergent Properties in Multimodal Interleaved Chain-of-Thought Reasoning

Authors: Gu Jiawei, Yunzhuo Hao, Li, Linjie, Huichen Will Wang, et al. (15 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Multimodal reasoning requires iterative coordination between language and vision, yet it remains unclear what constitutes a meaningful interleaved chain of thought. We posit that text and image thoughts should function as complementary rather than isomorphic modalities that mutually advance reasoning. Guided by this principle, we build ThinkMorph, a unified model fine-tuned on approximately 24K high-quality interleaved reasoning traces spanning tasks with varying visual engagement. ThinkMorph le...

Relationship Analysis

Both papers belong to the Text-Image Interleaved Chain-of-Thought category, focusing on methods that generate sequential reasoning steps with paired visual and textual rationales. They overlap in their core approach of interleaving textual and visual thoughts to enhance multimodal reasoning, with both employing two-stage training strategies (supervised learning followed by reinforcement learning) to enable models to alternate between understanding and generation. The key difference is that the original paper (AD-Loop) emphasizes the synergy between understanding and generation as complementary processes through an analyzing-drafting loop, while ThinkMorph focuses on emergent properties and adaptive switching between reasoning modes, treating text and image thoughts as complementary modalities that mutually advance reasoning rather than isomorphic representations.

3. Interleaving Reasoning for Better Text-to-Image Generation

Authors: Huang Wenxuan, Chen Shuang, Wenxuan Huang, Shuang Chen, Cao Shaosheng, et al. (41 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Unified multimodal understanding and generation models recently have achieve significant improvement in image generation capability, yet a large gap remains in instruction following and detail preservation compared to systems that tightly couple comprehension with generation such as GPT-4o. Motivated by recent advances in interleaving reasoning, we explore whether such reasoning can further improve Text-to-Image (T2I) generation. We introduce Interleaving Reasoning Generation (IRG), a framework ...

Relationship Analysis

Both papers belong to the Text-Image Interleaved Chain-of-Thought category, employing sequential reasoning steps with paired visual and textual rationales to enhance multimodal generation. They overlap in using interleaved text-image thinking processes to improve generation quality through iterative refinement. However, the original paper (AD-Loop) focuses on synergizing understanding and generation through dynamic alternation between analyzing and drafting operations with reinforcement learning for adaptive control, while the candidate paper (IRG) specifically targets text-to-image generation improvement through a two-stage reflection process (initial generation followed by quality refinement) with decomposed learning modes and specialized CFG conditioning strategies.

Contributions Analysis

Overall novelty summary. The paper proposes an Analyzing-Drafting loop (AD-Loop) that alternates between analytic and drafting operations to synergize understanding and generation in unified vision-language models. It resides in the 'Text-Image Interleaved Chain-of-Thought' leaf, which contains four papers including the original work. This leaf sits within the broader 'Interleaved Reasoning Paradigms' branch, indicating a moderately populated research direction focused on sequential reasoning mechanisms. The taxonomy shows this is an active but not overcrowded area, with sibling papers exploring similar interleaving strategies but differing in synchronization granularity and reasoning structure.

The taxonomy reveals neighboring leaves addressing related but distinct approaches: 'Latent Visual Reasoning' performs reasoning in feature space to avoid pixel-level encoding, 'Draft-and-Refine Reasoning' generates low-resolution previews for iterative refinement, and 'Tool-Augmented Reasoning Frameworks' orchestrate external tools during reasoning. The original paper's AD-Loop differs by emphasizing explicit alternation between understanding and generation phases within a single framework, rather than latent processing, external tool calls, or preview-based refinement. The broader 'Unified Multimodal Model Architectures' branch contains architectural designs that could host such reasoning mechanisms, suggesting the paper's framework is complementary to rather than overlapping with architectural innovations.

Among thirty candidates examined across three contributions, none were identified as clearly refuting the proposed work. The AD-Loop mechanism examined ten candidates with zero refutable matches, suggesting limited direct prior work on this specific alternating paradigm within the search scope. The two-stage training strategy (supervised initialization followed by reinforcement learning) also examined ten candidates without refutation, though the taxonomy shows related work in 'Reinforcement Learning for Interleaved Tasks' and 'Multi-Stage and Curriculum Training' leaves. The architecture-agnostic framework claim examined ten candidates with no refutations, aligning with the taxonomy's distinction between reasoning paradigms and architectural designs. These statistics reflect a focused semantic search rather than exhaustive coverage.

The analysis suggests the paper occupies a relatively novel position within its immediate research neighborhood, particularly in the explicit synchronization of analyzing and drafting phases. However, the limited search scope (thirty candidates from semantic matching) means the assessment is provisional. The taxonomy structure indicates the paper builds on established foundations in interleaved reasoning while proposing a distinct control mechanism, but comprehensive novelty assessment would require broader examination of the 'Training Strategies' and 'Unified Architectures' branches where overlapping ideas might exist.

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

Contribution 1: Interleaved Analyzing-Drafting problem-solving loop (AD-Loop)

Description: A novel thinking paradigm that enables unified vision-language models to dynamically alternate between understanding (analyzing) and generation (drafting) operations. By interleaving textual thoughts with visual thoughts, AD-Loop fosters genuine synergy between comprehension and creation during task solving.

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. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model

URL: View paper

Brief Assessment

Internlm Xcomposer[51] focuses on text-image composition and comprehension capabilities through architectural design (partial LoRA) rather than interleaved analyzing-drafting operations during problem-solving. The candidate does not demonstrate prior work on dynamic alternation between understanding and generation operations as a thinking paradigm.

2. DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models

URL: View paper

Brief Assessment

DriveVLM[55] focuses on autonomous driving with hierarchical planning modules for scene understanding, not on interleaved analyzing-drafting loops that alternate between textual and visual thoughts during general problem-solving.

3. Evaluating text-to-visual generation with image-to-text generation

URL: View paper

Brief Assessment

Text Visual Evaluation[56] focuses on evaluating text-to-visual generation using VQA-based metrics, not on designing interleaved analyzing-drafting mechanisms for unified vision-language models during task solving.

4. Mini-Gemini: Mining the Potential of Multi-modality Vision Language Models

URL: View paper

Brief Assessment

Mini Gemini[53] focuses on architectural improvements for VLMs (high-resolution visual tokens, data quality) rather than interleaved reasoning loops. It does not propose dynamic alternation between understanding and generation during task solving.

5. Learning interleaved image-text comprehension in vision-language large models

URL: View paper

Brief Assessment

Interleaved Comprehension Learning[52] focuses on comprehension tasks where models must locate relevant text-image pairs within long documents to answer questions. The original paper's AD-Loop addresses a different problem: dynamically alternating between understanding (analyzing) and generation (drafting) operations during task solving to create synergy between comprehension and creation.

6. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models

URL: View paper

Brief Assessment

Visual Program Distillation[54] focuses on distilling LLM-generated programs and tool invocations into VLMs through chain-of-thought training, not on interleaved analyzing-drafting operations during task solving. The candidate's approach converts program execution traces into natural language rationales for distillation, rather than enabling dynamic alternation between understanding and generation modes within the model itself.

7. Fine-tuning large vision-language models as decision-making agents via reinforcement learning

URL: View paper

Brief Assessment

Vision Language RL[60] focuses on fine-tuning VLMs with reinforcement learning for decision-making tasks using chain-of-thought reasoning, not on interleaved analyzing-drafting operations that alternate between understanding and generation capabilities within unified vision-language models.

8. Zebra-cot: A dataset for interleaved vision language reasoning

URL: View paper

Brief Assessment

Zebra CoT[57] focuses on creating a dataset for training models to perform visual chain-of-thought reasoning with interleaved text-image traces. It does not propose a dynamic problem-solving loop that alternates between understanding and generation operations during task execution, which is the core novelty of AD-Loop.

9. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control

URL: View paper

Brief Assessment

RT[59] focuses on vision-language-action models for robotic control, not on interleaved analyzing-drafting operations in vision-language models. The candidate addresses action prediction for robots, not the dynamic alternation between understanding and generation operations described in the original contribution.

10. Source-free domain adaptation with frozen multimodal foundation model

URL: View paper

Brief Assessment

Source Free Adaptation[58] focuses on domain adaptation using frozen multimodal models for transfer learning across domains, not on interleaved analyzing-drafting operations during problem-solving within vision-language models.

Contribution 2: Two-stage training strategy for AD-Loop

Description: A training framework consisting of supervised learning on interleaved thought data to initialize the alternation mechanism, followed by reinforcement learning with hybrid feedback to enable the model to intelligently and autonomously decide when to invoke understanding versus generation.

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. Automatic berthing using supervised learning and reinforcement learning

URL: View paper

Brief Assessment

Automatic Berthing[73] applies two-stage training (SL then RL) to ship berthing control, not to adaptive alternation between understanding and generation in vision-language models. The technical domains and objectives are fundamentally different.

2. ARMOR: Robust Reinforcement Learning-based Control for UAVs under Physical Attacks

URL: View paper

Brief Assessment

ARMOR UAV[80] addresses UAV control under physical sensor attacks using a two-stage framework (teacher-student encoders with privileged information), while the original paper focuses on vision-language model reasoning with supervised learning followed by RL for adaptive understanding-generation alternation. The domains, objectives, and technical mechanisms are fundamentally different.

3. Flexible resource management in high-throughput satellite communication systems: A two-stage machine learning framework

URL: View paper

Brief Assessment

Satellite Resource Management[75] addresses satellite communication resource allocation using self-supervised learning followed by reinforcement learning, not adaptive control of understanding versus generation capabilities in vision-language models.

4. A fuzzy controller with supervised learning assisted reinforcement learning algorithm for obstacle avoidance

URL: View paper

Brief Assessment

Fuzzy Obstacle Avoidance[78] addresses obstacle avoidance in mobile robotics using fuzzy controllers, not vision-language models or multimodal understanding/generation tasks. The two-stage approach (supervised then reinforcement learning) serves a fundamentally different purpose—tuning fuzzy membership functions for robot navigation rather than enabling adaptive alternation between understanding and generation in unified multimodal models.

5. Design and experimental validation of a cooperative adaptive cruise control system based on supervised reinforcement learning

URL: View paper

Brief Assessment

Cooperative Cruise Control[79] focuses on vehicle dynamics control using supervised RL for adaptive cruise control, not on vision-language models with interleaved understanding-generation loops. The domains and technical objectives are fundamentally different.

6. Formation control with collision avoidance through deep reinforcement learning using model-guided demonstration

URL: View paper

Brief Assessment

Formation Control Collision[74] addresses formation control with collision avoidance in multi-robot systems using a two-stage framework (imitation learning then RL), whereas the original paper proposes AD-Loop for unified vision-language models that alternate between understanding and generation tasks. The domains, objectives, and technical implementations are fundamentally different.

7. ADP: Adaptive Diffusion Policy Energizes Robots Thinking in Both Learning and Practice

URL: View paper

Brief Assessment

ADP[71] proposes a two-stage framework for robot control (offline imitation learning followed by online RL adaptation), while the original paper focuses on unified vision-language models with interleaved textual-visual reasoning. The domains and objectives are fundamentally different.

8. Towards Adaptive Humanoid Control via Multi-Behavior Distillation and Reinforced Fine-Tuning

URL: View paper

Brief Assessment

Adaptive Humanoid Control[77] focuses on humanoid locomotion control with multi-behavior distillation followed by reinforced fine-tuning for terrain adaptation, not on vision-language models with understanding-generation alternation mechanisms.

9. Pre-training with asynchronous supervised learning for reinforcement learning based autonomous driving

URL: View paper

Brief Assessment

Autonomous Driving Pretraining[76] focuses on pre-training RL-based autonomous driving models with supervised learning on driving demonstrations, then deploying for real-world RL training. This is domain-specific (autonomous driving) and does not address the interleaved analyzing-drafting mechanism or adaptive control between understanding and generation capabilities in unified vision-language models.

10. AdaCtrl: Towards Adaptive and Controllable Reasoning via Difficulty-Aware Budgeting

URL: View paper

Brief Assessment

AdaCtrl[72] uses a two-stage training approach (cold-start fine-tuning followed by difficulty-aware RL) for adaptive reasoning budget allocation in mathematical problem-solving, not for alternating between understanding and generation capabilities in unified vision-language models.

Contribution 3: Architecture-agnostic framework for UVLMs

Description: The proposed AD-Loop thinking mechanism and training strategy are designed to be broadly applicable across different unified vision-language model architectures, enabling seamless integration and performance improvements on diverse models.

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. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training

URL: View paper

Brief Assessment

Unicoder[63] focuses on cross-modal pre-training for vision-language tasks using a fixed transformer architecture, not on designing architecture-agnostic training frameworks that can be applied across different UVLM architectures.

2. Unified-io: A unified model for vision, language, and multi-modal tasks

URL: View paper

Brief Assessment

Unified IO[64] focuses on a unified transformer architecture for diverse vision-language tasks but does not present an architecture-agnostic training framework that can be applied across different UVLM architectures. The original paper's AD-Loop is designed to work with various existing UVLM architectures, while Unified IO proposes a single specific architecture.

3. Univl: A unified video and language pre-training model for multimodal understanding and generation

URL: View paper

Brief Assessment

Univl[67] focuses on video-language pre-training with a fixed encoder-decoder architecture for understanding and generation tasks, not on providing an architecture-agnostic training framework that can be applied across different unified vision-language model architectures.

4. Uncertainty-o: One Model-agnostic Framework for Unveiling Uncertainty in Large Multimodal Models

URL: View paper

Brief Assessment

Uncertainty[69] focuses on uncertainty estimation across different LMM architectures for hallucination detection, not on training frameworks for unified vision-language models that synergize understanding and generation capabilities.

5. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks

URL: View paper

Brief Assessment

Visionllm[68] focuses on a unified multimodal framework with task-specific decoders for hundreds of vision-language tasks, rather than a training framework for synergizing understanding and generation capabilities through interleaved thinking mechanisms.

6. MAO: Efficient Model-Agnostic Optimization of Prompt Tuning for Vision-Language Models

URL: View paper

Brief Assessment

MAO[70] focuses on prompt tuning for vision-language models (specifically CLIP-based models) rather than unified vision-language models that integrate both understanding and generation. The candidate addresses parameter-efficient fine-tuning through prompt optimization, not the architectural unification or training strategies for models that perform both comprehension and generation tasks simultaneously.

7. Hiprune: Training-free visual token pruning via hierarchical attention in vision-language models

URL: View paper

Brief Assessment

Hiprune[62] focuses on visual token pruning for vision-language models, not on training frameworks for unified vision-language models that synergize understanding and generation capabilities.

8. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model

URL: View paper

Brief Assessment

Ureader[65] focuses on OCR-free visually-situated language understanding using instruction tuning on a specific MLLM (mplug-owl), not on a broadly applicable training framework across different UVLM architectures.

9. EigenShield: Causal Subspace Filtering via Random Matrix Theory for Adversarially Robust Vision-Language Models

URL: View paper

Brief Assessment

EigenShield[66] focuses on adversarial defense for vision-language models through spectral filtering, not on training frameworks for unified vision-language models. The candidate addresses security vulnerabilities rather than architectural unification or training strategies for understanding-generation synergy.

10. E5-v: Universal embeddings with multimodal large language models

URL: View paper

Brief Assessment

E5 Universal Embeddings[61] focuses on universal multimodal embeddings using MLLMs for retrieval tasks, not on unified vision-language models that perform both understanding and generation. The architectural concerns and training strategies differ fundamentally from the AD-Loop framework.

Appendix: Text Similarity Detection

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

References

- [0] Synergizing Understanding and Generation with Interleaved Analyzing-Drafting Thinking View paper
- [1] Dreamllm: Synergistic multimodal comprehension and creation View paper
- [2] Emerging properties in unified multimodal pretraining View paper
- [3] A Survey of Unified Multimodal Understanding and Generation: Advances and Challenges View paper
- [4] Graphgpt-o: Synergistic multimodal comprehension and generation on graphs View paper
- [5] Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models View paper
- [6] Multimodal representation alignment for image generation: Text-image interleaved control is easier than you think View paper
- [7] Interleaved-modal chain-of-thought View paper
- [8] Interleaved multi-modal document representations for large-scale information retrieval using large language models View paper
- [9] Generative universal verifier as multimodal meta-reasoner View paper
- [10] Codi-2: In-context interleaved and interactive any-to-any generation View paper
- [11] ANOLE: An Open, Autoregressive, Native Large Multimodal Models for Interleaved Image-Text Generation View paper
- [12] Mmie: Massive multimodal interleaved comprehension benchmark for large vision-language models View paper
- [13] VITA-Audio: Fast Interleaved Cross-Modal Token Generation for Efficient Large Speech-Language Model View paper
- [14] Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer View paper
- [15] Mogao: An omni foundation model for interleaved multi-modal generation View paper
- [16] ThinkMorph: Emergent Properties in Multimodal Interleaved Chain-of-Thought Reasoning View paper
- [17] RealSyn: An Effective and Scalable Multimodal Interleaved Document Transformation Paradigm View paper
- [18] Lateralization lora: Interleaved instruction tuning with modality-specialized adaptations View paper
- [19] Lavida-O: Elastic Large Masked Diffusion Models for Unified Multimodal Understanding and Generation View paper
- [20] ARMOR: Empowering Multimodal Understanding Model with Interleaved Multimodal Generation Capability View paper
- [21] Modality-Specialized Synergizers for Interleaved Vision-Language Generalists View paper
- [22] Gelina: Unified Speech and Gesture Synthesis via Interleaved Token Prediction View paper
- [23] CoMM: A Coherent Interleaved Image-Text Dataset for Multimodal Understanding and Generation View paper
- [24] LVLM-MIR: Large Vision-Language Model with Parameter-Efficient Fine-Tuning for Multimodal Interleaved Reasoning View paper
- [25] Skywork-R1V4: Toward Agentic Multimodal Intelligence through Interleaved Thinking with Images and DeepResearch View paper
- [26] CX-Mind: A Pioneering Multimodal Large Language Model for Interleaved Reasoning in Chest X-ray via Curriculum-Guided Reinforcement Learning View paper
- [27] MIRAGE25: ACM MM25 Multimodal Interleaved Reasoning and Generation Challenge View paper
- [28] MChat: Empowering VLM for Multimodal LLM Interleaved Text-Image Generation View paper
- [29] Contextual AD Narration with Interleaved Multimodal Sequence View paper
- [30] Reasoning Within the Mind: Dynamic Multimodal Interleaving in Latent Space View paper
- [31] STaR-Attack: A Spatio-Temporal and Narrative Reasoning Attack Framework for Unified Multimodal Understanding and Generation Models View paper
- [32] Deep thinking in text-to-image generation using unified model with reinforcement learning View paper
- [33] Interleaved Latent Visual Reasoning with Selective Perceptual Modeling View paper
- [34] VICoT-Agent: A Vision-Interleaved Chain-of-Thought Framework for Interpretable Multimodal Reasoning and Scalable Remote Sensing Analysis View paper
- [35] Interleaving Reasoning for Better Text-to-Image Generation View paper
- [36] Thinking with Video: Video Generation as a Promising Multimodal Reasoning Paradigm View paper
- [37] A High-Quality Dataset and Reliable Evaluation for Interleaved Image-Text Generation View paper
- [38] IRG-MotionLLM: Interleaving Motion Generation, Assessment and Refinement for Text-to-Motion Generation View paper
- [39] WEAVE: Unleashing and Benchmarking the In-context Interleaved Comprehension and Generation View paper
- [40] HUMORCHAIN: Theory-Guided Multi-Stage Reasoning for Interpretable Multimodal Humor Generation View paper
- [41] LLM-I: LLMs are Naturally Interleaved Multimodal Creators View paper

- [42] Learning Video Context as Interleaved Multimodal Sequences View paper
- [43] IUT-Plug: A Plug-in tool for Interleaved Image-Text Generation View paper
- [44] OneFlow: Concurrent Mixed-Modal and Interleaved Generation with Edit Flows View paper
- [45] ROVER: Benchmarking Reciprocal Cross-Modal Reasoning for Omnimodal Generation View paper
- [46] TV2TV: A Unified Framework for Interleaved Language and Video Generation View paper
- [47] DraCo: Draft as CoT for Text-to-Image Preview and Rare Concept Generation View paper
- [48] Multimodal linguistic inference View paper
- [49] Multimodal Clinical Reasoning through Knowledge-augmented Rationale Generation View paper
- [50] TextBind: Multi-turn Interleaved Multimodal Instruction-following in the Wild View paper
- [51] Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model View paper
- [52] Learning interleaved image-text comprehension in vision-language large models View paper
- [53] Mini-Gemini: Mining the Potential of Multi-modality Vision Language Models View paper
- [54] Visual program distillation: Distilling tools and programmatic reasoning into vision-language models View paper
- [55] DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models View paper
- [56] Evaluating text-to-visual generation with image-to-text generation View paper
- [57] Zebra-cot: A dataset for interleaved vision language reasoning View paper
- [58] Source-free domain adaptation with frozen multimodal foundation model View paper
- [59] RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control View paper
- [60] Fine-tuning large vision-language models as decision-making agents via reinforcement learning View paper
- [61] E5-v: Universal embeddings with multimodal large language models View paper
- [62] Hiprune: Training-free visual token pruning via hierarchical attention in vision-language models View paper
- [63] Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training View paper
- [64] Unified-io: A unified model for vision, language, and multi-modal tasks View paper
- [65] Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model View paper
- [66] EigenShield: Causal Subspace Filtering via Random Matrix Theory for Adversarially Robust Vision-Language Models View paper
- [67] Univl: A unified video and language pre-training model for multimodal understanding and generation View paper
- [68] Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks View paper
- [69] Uncertainty-o: One Model-agnostic Framework for Unveiling Uncertainty in Large Multimodal Models View paper
- [70] MAO: Efficient Model-Agnostic Optimization of Prompt Tuning for Vision-Language Models View paper
- [71] ADP: Adaptive Diffusion Policy Energizes Robots Thinking in Both Learning and Practice View paper
- [72] AdaCtrl: Towards Adaptive and Controllable Reasoning via Difficulty-Aware Budgeting View paper
- [73] Automatic berthing using supervised learning and reinforcement learning View paper
- [74] Formation control with collision avoidance through deep reinforcement learning using model-guided demonstration View paper
- [75] Flexible resource management in high-throughput satellite communication systems: A two-stage machine learning framework View paper
- [76] Pre-training with asynchronous supervised learning for reinforcement learning based autonomous driving View paper
- [77] Towards Adaptive Humanoid Control via Multi-Behavior Distillation and Reinforced Fine-Tuning View paper
- [78] A fuzzy controller with supervised learning assisted reinforcement learning algorithm for obstacle avoidance View paper
- [79] Design and experimental validation of a cooperative adaptive cruise control system based on supervised reinforcement learning View paper
- [80] ARMOR: Robust Reinforcement Learning-based Control for UAVs under Physical Attacks View paper