Novelty Assessment Report

Paper: Generative Universal Verifier as Multimodal Meta-Reasoner

PDF URL: https://openreview.net/pdf?id=DM0Y0oL33T

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-30

Abstract

We introduce Generative Universal Verifier, a novel concept and plugin designed for next-generation multimodal reasoning in vision-language models and unified multimodal models, providing the fundamental capability of reflection and refinement on visual outcomes during the reasoning and generation process. This work makes three main contributions: (1) We build **ViVerBench**, a comprehensive benchmark spanning 16 categories of critical tasks for evaluating visual outcomes in multimodal reasoning. Results show that existing VLMs consistently underperform across these tasks, underscoring a substantial gap from human-level capability in reliable visual verification. (2) We design two automated pipelines to construct large-scale visual verification data and train **OmniVerifier-7B**, the first omni-capable generative verifier trained for universal visual verification and achieves notable gains on ViVerBench(+8.3). Through training, we identify three atomic capabilities in visual verification and demonstrate how they generalize and interact synergistically. (3) We propose **OmniVerifier-TTS**, a sequential test-time scaling paradigm that leverages the universal verifier to bridge image generation and editing within unified models, enhancing the upper bound of generative ability through iterative fine-grained optimization. Beyond generation, we extend universal verifier to broader world-modeling interleaved reasoning scenarios. Empirically, OmniVerifier-TTS achieves improvements on T2I-ReasonBench(+3.7), and GenEval++(+4.3), outperforming existing parallel test-time scaling methods, such as Best-of-N. By endowing multimodal reasoning with reliable visual verification, OmniVerifier advances both reliable reflection during generation and scalable test-time refinement, marking a step toward more trustworthy and controllable next-generation reasoning systems.

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: **Visual Outcome Verification in Multimodal Reasoning**

A total of **50 papers** were analyzed and organized into a taxonomy with **31 categories**.

Taxonomy Overview

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

- **Chain-of-Thought Reasoning Enhancement**
- **Reinforcement Learning-Based Reasoning**
- **Verification Mechanisms**
- **Tool-Integrated Reasoning**
- **Benchmark Development**
- **Analysis and Robustness Studies**
- **Survey and Theoretical Foundations**
- **Architectural Innovations**

Complete Taxonomy Tree

- Visual Outcome Verification in Multimodal Reasoning Survey Taxonomy
- Chain-of-Thought Reasoning Enhancement
  - Autonomous Multistage Reasoning (2 papers)
  - [1] Llava-cot: Let vision language models reason step-by-step (Xu Guowei, 2025) View paper
  - [14] Insight-v: Exploring long-chain visual reasoning with multimodal large language models (Yuhao Dong, 2025) View paper
  - Iterative Self-Improvement (2 papers)
  - [2] Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement (Deng Yihe, 2025) View paper
  - [18] Vision-language models can self-improve reasoning via reflection (Li Yantao, 2025) View paper
  - Interleaved Vision-Language Reasoning (2 papers)
  - [3] Zebra-cot: A dataset for interleaved vision language reasoning (Li Ang, 2025) View paper
  - [12] Latent visual reasoning (Li, 2025) View paper
  - Decomposition-Based Reasoning (1 papers)
  - [6] Idealgpt: Iteratively decomposing vision and language reasoning via large language models (You, 2023) View paper
  - Spatial and Geometric Reasoning (2 papers)
  - [9] Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing (Wu, 2025) View paper
  - [19] ChainV: Atomic Visual Hints Make Multimodal Reasoning Shorter and Better (Yuan Zhang, 2025) View paper
  - Compression and Efficiency (1 papers)
  - [32] Reasoningtrack: Chain-of-thought reasoning for long-term vision-language tracking (Wang Xiao, 2025) View paper
- Reinforcement Learning-Based Reasoning
  - General-Purpose RL Training (2 papers)
  - [10] WeThink: Toward General-purpose Vision-Language Reasoning via Reinforcement Learning (Yang, 2025) View paper

  - [20] Open vision reasoner: Transferring linguistic cognitive behavior for visual reasoning (Wei Yana, 2025) View paper
  - Process Reward Models (1 papers)
  - [31] Training Vision-Language Process Reward Models for Test-Time Scaling in Multimodal Reasoning: Key Insights and Lessons Learned (Pala Tej Deep, 2025) View paper
  - Domain-Specific RL Applications (2 papers)
  - [23] Reasoning-VLA: A Fast and General Vision-Language-Action Reasoning Model for Autonomous Driving (Dapeng Zhang, 2025) View paper
  - [39] Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning (Meng Fanqing, 2025) View paper
  - Two-Stage RL Frameworks (2 papers)
  - [44] Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models (Tan Huajie, 2025) View paper
  - [45] Perception before reasoning: Two-stage reinforcement learning for visual reasoning in vision-language models (Chen Yan, 2025) View paper
  - Visual Anchoring in RL (1 papers)
  - [42] SATORI-R1: Incentivizing Multimodal Reasoning through Explicit Visual Anchoring (Chuming Shen, 2025) View paper
- Verification Mechanisms
  - Universal Visual Verification ★ (1 papers)
  - [0] Generative Universal Verifier as Multimodal Meta-Reasoner (Anon et al., 2026) View paper
  - Chain-of-Thought Verification (2 papers)
  - [25] Mm-verify: Enhancing multimodal reasoning with chain-of-thought verification (Liang Hao, 2025) View paper
  - [30] MM-CoT: A Benchmark for Probing Visual Chain-of-Thought Reasoning in Multimodal Models (Jusheng Zhang, 2025) View paper
  - Model-to-Model Deliberation (1 papers)
  - [5] Enhancing safety of vision-language reasoning through model-to-model deliberation (Sungwoo Kim, 2025) View paper
  - Task-Specific Verification (1 papers)
  - [21] Prometheus-vision: Vision-language model as a judge for fine-grained evaluation (Lee, 2024) View paper
  - Hallucination Mitigation (2 papers)
  - [28] ClearSight: Visual Signal Enhancement for Object Hallucination Mitigation in Multimodal Large Language Models (Hao Yin, 2025) View paper
  - [37] VidHalluc: Evaluating Temporal Hallucinations in Multimodal Large Language Models for Video Understanding (Chaoyu Li, 2024) View paper
- Tool-Integrated Reasoning
  - Self-Evolving Tool Integration (1 papers)
  - [29] Agent0-VL: Exploring Self-Evolving Agent for Tool-Integrated Vision-Language Reasoning (Jiaqi Liu, 2025) View paper
  - Knowledge Graph Integration (1 papers)
  - [35] Aligning Vision to Language: Text-Free Multimodal Knowledge Graph Construction for Enhanced LLMs Reasoning (Liu Junming, 2025) View paper
  - Critical Thinking Frameworks (1 papers)
  - [50] Mmctagent: Multi-modal critical thinking agent framework for complex visual reasoning (Kumar, 2024) View paper
- Benchmark Development
  - Visual Reasoning Evaluation (3 papers)
  - [4] Verify: A benchmark of visual explanation and reasoning for investigating multimodal reasoning fidelity (Bi Jing, 2025) View paper
  - [13] Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models (XU Weiye, 2025) View paper
  - [17] Zero-shot visual reasoning by vision-language models: Benchmarking and analysis (Aishik Nagar, 2024) View paper
  - Domain-Specific Benchmarks (5 papers)
  - [11] Measuring multimodal mathematical reasoning with math-vision dataset (Hongsheng Li, 2024) View paper
  - [15] Vlr-driver: Large vision-language-reasoning models for embodied autonomous driving (F Kong, 2025) View paper
  - [16] MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts (Lu Pan, 2023) View paper
  - [26] Sci-reason: A dataset with chain-of-thought rationales for complex multimodal reasoning in academic areas (MA Chenghao, 2025) View paper
  - [33] CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs (Sanjeev Arora, 2024) View paper
  - Multimodal Understanding Benchmarks (2 papers)
  - [7] All-in-one: Understanding and Generation in Multimodal Reasoning with the MAIA Benchmark (Bonetta, 2025) View paper
  - [8] Multimodal inconsistency reasoning (mmir): A new benchmark for multimodal reasoning models (Fan Yue, 2025) View paper
  - Holistic Evaluation Frameworks (2 papers)
  - [24] Unibench: Visual reasoning requires rethinking vision-language beyond scaling (Al-Tahan, 2024) View paper
  - [27] Vhelm: A holistic evaluation of vision language models (Lee, 2024) View paper
  - Agentic and Deep Reasoning Benchmarks (1 papers)
  - [47] Agent-X: Evaluating Deep Multimodal Reasoning in Vision-Centric Agentic Tasks (Tajamul Ashraf, 2025) View paper
- Analysis and Robustness Studies
  - Reasoning Consistency Analysis (2 papers)
  - [41] Measuring and improving chain-of-thought reasoning in vision-language models (Chen Yangyi, 2024) View paper
  - [49] Visual Reasoning Consistency and Robustness Analysis of Multimodal LLMs (N Jegham, 2025) View paper
  - Bias and Robustness Analysis (3 papers)
  - [22] Scale Can't Overcome Pragmatics: The Impact of Reporting Bias on Vision-Language Reasoning (A Kamath, 2025) View paper
  - [36] Order Matters: Exploring Order Sensitivity in Multimodal Large Language Models (Tan Zhijie, 2024) View paper
  - [40] Verite: a robust benchmark for multimodal misinformation detection accounting for unimodal bias (Stefanos-Iordanis Papadopoulos, 2024) View paper
  - Visual Comprehension Training (1 papers)
  - [38] Describe-then-reason: Improving multimodal mathematical reasoning through visual comprehension training (Jia, 2024) View paper
  - Interpretability and Explainability (1 papers)

    - [48] Visqa: X-raying vision and language reasoning in transformers (Jaunet, 2021) View paper
    - Fact-Checking and Misinformation (1 papers)
    - [43] End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models (Barry Menglong Yao, 2023) View paper
- Survey and Theoretical Foundations (1 papers)
    - [46] Why reasoning matters? a survey of advancements in multimodal reasoning (v1) (Bi Jing, 2025) View paper
- Architectural Innovations (1 papers)
    - [34] Corvid: Improving multimodal large language models towards chain-of-thought reasoning (Jiang JingJing, 2025) View paper

Narrative

Core task: visual outcome verification in multimodal reasoning. The field has organized itself around several complementary directions that address how models can produce, evaluate, and improve reasoning over visual and textual inputs. Chain-of-Thought Reasoning Enhancement explores structured prompting and intermediate step generation (e.g., Llava CoT[1], Zebra CoT[3]), while Reinforcement Learning-Based Reasoning applies policy optimization and reward signals to refine reasoning trajectories. Verification Mechanisms form a central pillar, encompassing approaches that explicitly check the correctness or consistency of generated outputs, including universal verifiers that operate across diverse reasoning tasks. Tool-Integrated Reasoning investigates how models can leverage external modules —such as code interpreters or symbolic solvers—to ground their predictions, and Benchmark Development provides standardized testbeds (e.g., Verify Benchmark[4], MAIA Benchmark[7]) for measuring progress. Analysis and Robustness Studies examine failure modes and biases, Survey and Theoretical Foundations synthesize emerging principles, and Architectural Innovations propose novel model designs to better fuse vision and language.

Within this landscape, a particularly active line of work focuses on building verifiers that can assess reasoning quality without task-specific training, contrasting with methods that rely on heavy supervision or domain-tailored reward models. Generative Universal Verifier[0] sits squarely in this Universal Visual Verification cluster, emphasizing a flexible verification strategy that generalizes across problem types and modalities. This approach differs from works like Model Deliberation Safety[5], which targets safety-oriented verification in high-stakes settings, and from MM Verify[25], which may incorporate more specialized checks for particular reasoning patterns. The trade-off centers on breadth versus depth: universal verifiers aim for wide applicability but must balance that generality against the precision achievable by narrower, task-tuned methods. Open questions remain about how to scale verification signals efficiently and how to integrate them into iterative reasoning loops without prohibitive computational cost.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

Universal Visual Verification focuses on generative verifiers that provide broad, cross-task visual outcome verification for multimodal reasoning and generation. Its siblings represent more specialized verification approaches: Chain-of-Thought Verification targets reasoning chain validation, Hallucination Mitigation addresses specific error types through visual enhancement, Model-to-Model Deliberation employs multi-agent frameworks, and Task-Specific Verification customizes evaluation for particular applications. The original leaf distinguishes itself by emphasizing universality and generality across diverse tasks rather than specialization.

Similarities: - All subtopics address verification or validation in multimodal or visual reasoning contexts - All aim to improve reliability, accuracy, or robustness of model outputs - Multiple subtopics involve external mechanisms or models rather than pure self-evaluation - All exclude certain verification approaches that belong in other categories, indicating clear boundary definitions

Differences: - Universal Visual Verification emphasizes cross-task generality, while Task-Specific Verification explicitly focuses on customized, task-dependent evaluation - Chain-of-Thought Verification targets reasoning process validation, whereas Universal Visual Verification focuses on outcome verification - Hallucination Mitigation addresses a specific failure mode (hallucinations), while Universal Visual Verification provides broader outcome assessment - Model-to-Model Deliberation requires multi-agent interaction, while Universal Visual Verification can operate with single generative verifiers - Universal Visual Verification explicitly excludes task-specific approaches, positioning itself as a general-purpose solution

Suggested Search Directions: - Investigate how universal verifiers handle domain transfer compared to task-specific verification methods - Explore whether universal visual verification can incorporate hallucination detection as a component capability - Examine the trade-offs between universal generative verifiers and specialized multi-agent deliberation frameworks - Study how universal verification relates to or could integrate with chain-of-thought reasoning validation

Sibling Subtopics

- **Chain-of-Thought Verification** (leaves: 1, papers: 2)
- Scope: Methods verifying reasoning chains through external verification models or multi-round verification mechanisms.
- Exclude: Excludes self-reflection without external verification; those belong in Iterative Self-Improvement.
- **Hallucination Mitigation** (leaves: 1, papers: 2)
- Scope: Techniques reducing object or temporal hallucinations through visual signal enhancement or contrastive decoding strategies.
- Exclude: Excludes general verification without hallucination focus; those belong in Chain-of-Thought Verification.
- **Model-to-Model Deliberation** (leaves: 1, papers: 1)
- Scope: Frameworks using multi-agent deliberation or model-to-model interaction for safety and robustness through structured verification.
- Exclude: Excludes single-model verification; those belong in Chain-of-Thought Verification.
- **Task-Specific Verification** (leaves: 1, papers: 1)
- Scope: Verification methods targeting specific tasks like long-form response assessment or fine-grained evaluation with customized rubrics.
- Exclude: Excludes universal verification; those belong in Universal Visual Verification.

Contributions Analysis

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

Contribution 1: ViVerBench: comprehensive benchmark for visual verification

Description: The authors construct ViVerBench, a benchmark with 3,594 manually annotated questions across 16 subtasks in 6 categories to systematically evaluate multimodal models' ability to verify visual outcomes. The benchmark reveals substantial gaps between current VLMs and human-level visual verification capability.

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. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi

URL: View paper

Brief Assessment

MMMU[54] focuses on college-level multimodal understanding across 30 subjects with diverse question types, not specifically on visual verification tasks for evaluating model-generated visual outcomes during reasoning processes.

2. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning

URL: View paper

Brief Assessment

Visual CoT[55] focuses on visual chain-of-thought reasoning with bounding box annotations for question-answering tasks, not on visual verification of generated outcomes. The benchmarks serve fundamentally different purposes: ViVerBench evaluates verification of visual outcomes (e.g., image-prompt alignment, physics plausibility), while Visual CoT evaluates region identification for VQA tasks.

3. Visualtrans: A benchmark for real-world visual transformation reasoning

URL: View paper

Brief Assessment

VisualTrans[51] focuses on visual transformation reasoning in human-object interaction scenarios, evaluating spatial, procedural, and quantitative reasoning across manipulation tasks. This differs fundamentally from ViVerBench's focus on verifying visual outcomes in multimodal reasoning (e.g., image-prompt alignment, world dynamics, state value evaluation). The two benchmarks address distinct aspects of visual understanding.

4. Beyond seeing: Evaluating multimodal llms on tool-enabled image perception, transformation, and reasoning

URL: View paper

Brief Assessment

Tool Enabled Perception[57] focuses on tool-enabled image perception and transformation for multimodal reasoning tasks, not visual verification of generated outcomes. The benchmarks serve fundamentally different purposes.

5. Grounded Reinforcement Learning for Visual Reasoning

URL: View paper

Brief Assessment

Grounded Reinforcement Learning[56] focuses on training vision-language models for visual reasoning through reinforcement learning with spatial grounding, not on constructing benchmarks for evaluating visual verification capabilities across diverse multimodal reasoning tasks.

6. Fakebench: Probing explainable fake image detection via large multimodal models

URL: View paper

Brief Assessment

FakeBench[52] focuses specifically on fake image detection with explainability for AI-generated images, while ViVerBench evaluates visual verification across diverse multimodal reasoning tasks including image generation, editing, and world-modeling scenarios. The scopes and task definitions differ fundamentally.

7. Multimodal inconsistency reasoning (mmir): A new benchmark for multimodal reasoning models

URL: View paper

Brief Assessment

Multimodal Inconsistency Reasoning[8] focuses on detecting semantic mismatches and inconsistencies in layout-rich artifacts (webpages, slides, posters), not on visual verification of generated outcomes across diverse reasoning tasks as in ViVerBench.

8. MM-CoT: A Benchmark for Probing Visual Chain-of-Thought Reasoning in Multimodal Models

URL: View paper

Brief Assessment

MM CoT[30] focuses on verifying chain-of-thought reasoning chains for logical coherence and visual grounding in multimodal models, not on visual outcome verification during generation/reasoning processes as in ViVerBench.

9. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models

URL: View paper

Brief Assessment

VisuLogic[13] focuses on visual logical reasoning tasks (e.g., quantitative shifts, spatial relations) rather than visual verification of outcomes during multimodal reasoning. The two benchmarks serve fundamentally different purposes and assess distinct capabilities.

10. Benchlmm: Benchmarking cross-style visual capability of large multimodal models

URL: View paper

Brief Assessment

BenchLMM[53] focuses on evaluating cross-style visual capabilities (artistic, sensor, application styles) rather than visual verification of outcomes in multimodal reasoning tasks.

Contribution 2: OmniVerifier-7B: first omni-capable generative verifier

Description: The authors develop two automated data construction pipelines and train OmniVerifier-7B, achieving notable gains on ViVerBench (+8.3). They identify three atomic capabilities in visual verification (explicit alignment, relational verification, and integrative reasoning) and demonstrate their generalization and synergistic interaction.

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. An Efficient Rubric-based Generative Verifier for Search-Augmented LLMs

URL: View paper

Brief Assessment

Rubric Generative Verifier[59] focuses on verifying search-augmented LLM outputs using nugget-based rubrics for factual tasks, not universal visual verification across 16 diverse task categories.

2. MedVLSynther: Synthesizing High-Quality Visual Question Answering from Medical Documents with Generator-Verifier LMMs

URL: View paper

Brief Assessment

MedVLSynther[60] focuses on medical VQA data synthesis using a generator-verifier framework for quality control, not on training universal visual verifiers for general multimodal reasoning tasks.

3. Generative hierarchical features from synthesizing images

URL: View paper

Brief Assessment

Generative Hierarchical Features[58] focuses on learning hierarchical visual features from GAN generators for image synthesis and editing tasks, not on training generative verifiers for visual verification using automated data pipelines.

Contribution 3: OmniVerifier-TTS: sequential test-time scaling paradigm

Description: The authors propose OmniVerifier-TTS, a sequential test-time scaling method that uses the universal verifier to iteratively refine generated images through verification and editing. This approach achieves improvements on T2I-ReasonBench (+3.7) and GenEval++ (+4.3), outperforming parallel test-time scaling methods like Best-of-N.

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. Scalingnoise: Scaling inference-time search for generating infinite videos

URL: View paper

Brief Assessment

ScalingNoise[64] focuses on video generation through inference-time search for optimal initial noises in diffusion models, not on sequential test-time scaling for image generation and editing using verifiers as proposed in the original paper.

2. From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning

URL: View paper

Prior Art Analysis

Reflection Tuning[65] demonstrates that sequential test-time scaling using verifiers to refine image generation was explored prior to the original paper's submission. The candidate paper presents reflectionflow, which implements sequential refinement through iterative reflection-based correction of generated images. Both papers employ a sequential paradigm where a verifier evaluates generated images and provides feedback for iterative refinement, with the candidate showing this approach achieves superior performance compared to parallel methods like Best-of-N sampling. The candidate's framework explicitly uses textual reflections from verifiers to guide iterative corrections, similar to the original paper's use of OmniVerifier for sequential refinement.

Evidence

Evidence 1 - **Rationale**: Both papers propose sequential test-time scaling frameworks that use iterative refinement. The candidate's 'reflection-level scaling' with 'actionable reflections to iteratively assess and correct previous generations' directly parallels the original's 'sequential test-time scaling paradigm' using 'universal verifier' for 'iterative fine-grained optimization.' - **Original**: we proposeomniverifier-tts, a sequential test-time scaling paradigm that leverages the universal verifier to bridge image generation and editing within unified models, enhancing the upper bound of generative ability through iterative fine-grained optimization. - **Candidate**: we propose reflectionflow, an inference-time framework enabling diffusion models to iteratively reflect upon and refine their outputs. reflectionflow introduces three complementary inference-time scaling axes: (1) noise-level scaling to optimize latent initialization; (2) prompt-level scaling for pr...

Evidence 2 - **Rationale**: Both papers describe sequential test-time scaling that progressively refines images through multiple rounds. The candidate's 'reflection-level scaling' involves iterative refinement at inference time, similar to the original's 'multiple rounds of verification and editing.' - **Original**: we proposeomniverifier-tts, a sequential test-time scaling method designed for enhancing the generation of unified multimodal models (deng et al., 2025; wu et al., 2025a) with omniverifier-7b. starting from a generated image, it progressively refines images through multiple rounds of verification an... - **Candidate**: leveraging the trained corrector model, we aim to maximize the generative capability of the diffusion model at inference time. in this section, we propose revisiting test-time scaling for t2i diffusion models along three distinct yet complementary dimensions: noise-level scaling, reflection-level sc...

Evidence 3 - **Rationale**: Both papers use a verifier to evaluate generated images and produce textual guidance for refinement. The original's 'edit prompt' that 'offers instructive guidance on how the image should be modified' parallels the candidate's 'textual reflections aimed at correcting identified errors.' - **Original**: omniverifier then analyzes this image and outputs a binary judgment (true/false) along with an explanation, following the same procedure as in its rl training. if the judgment is false, indicating misalignment between the prompt and the image, omniverifier additionally outputs an edit prompt, a reph... - **Candidate**: at iteration i, we utilize an mllm verifier to comprehensively evaluate and rank then images generated in the previous iteration across multiple dimensions. based on these evaluation scores and previously generated images, the mllm generates textual reflections aimed at correcting identified errors ...

3. Scaling Inference Time Compute for Diffusion Models

URL: View paper

Brief Assessment

Scaling Inference Compute[69] focuses on diffusion models for image generation through noise search optimization, not on sequential verification-editing refinement for unified multimodal models. The candidate's search framework operates on initial noise selection, while OmniVerifier-TTS iteratively refines generated images through verification and editing cycles.

4. Let's Verify and Reinforce Image Generation Step by Step

URL: View paper

Prior Art Analysis

Verify Reinforce Generation[68] demonstrates that sequential test-time scaling using verifiers to refine image generation was explored prior to the original paper's submission. The candidate paper presents a comprehensive investigation of chain-of-thought reasoning strategies applied to autoregressive image generation, including test-time verification with reward models (ORM and PRM) and iterative refinement approaches. Both papers employ similar sequential refinement paradigms where verifiers assess intermediate or final generation outputs and guide iterative improvements, though they differ in implementation details (the candidate uses autoregressive models while the original uses unified multimodal models).

Evidence

Evidence 1 - **Rationale**: Both papers propose sequential test-time scaling paradigms using verifiers. The candidate explicitly investigates 'scaling test-time computation with outcome/process reward model (orm/prm) as verifiers', which directly parallels the original's sequential test-time scaling approach. - **Original**: we proposeomniverifier-tts, a sequential test-time scaling paradigm that leverages the universal verifier to bridge image generation and editing within unified models, enhancing the upper bound of generative ability through iterative fine-grained optimization. - **Candidate**: we conduct a systematic investigation into the potential of cot reasoning for autoregressive image generation. we adopt show-o [53], a latest discrete generative model, as our baseline, and evaluate on a challenging text-to-image generation benchmark: geneval [12]. specifically, we focus on examinin...

Evidence 2 - **Rationale**: Both papers use verifiers to assess alignment between generated images and text prompts during test-time, providing feedback for iterative refinement. This demonstrates prior exploration of the core verification mechanism. - **Original**: as illustrated in fig. 5, we employ omniverifier as a 'misalignment-finder' due to its strong capability in explicit alignment and relational verification. the process begins with umm generating an image from a given prompt. omniverifier then analyzes this image and outputs a binary judgment (true/f... - **Candidate**: orm vs prm as test-time verifiers.as the top-1 result may not always be reliable, reward models are employed to score sampled candidates and perform outcome selection, where orm is instance-level and prm is processlevel. in our settings, the score assesses whether each candidate image is inherently ...

Evidence 3 - **Rationale**: Both papers describe iterative refinement processes where verifiers assess generation quality at multiple steps and determine whether to continue or terminate refinement, demonstrating similar sequential verification architectures. - **Original**: this iterative refinement loop continues omniverifier returns a true judgment or the maximum number of refinement steps is reached. - **Candidate**: potential assessment. for each clear step that passes the clarity judgment, parm assesses the potential of the current step to determine whether it can lead to a highquality final image, again using a binary label. if labeled 'no', the generation path is truncated immediately. if labeled 'yes', the ...

5. Video-t1: Test-time scaling for video generation

URL: View paper

Brief Assessment

Video T1[62] focuses on test-time scaling for video generation using tree-of-frames search and verifiers, not on sequential refinement of image generation through verification and editing as in OmniVerifier-TTS.

6. Can We Generate Images with CoT? Let's Verify and Reinforce Image Generation Step by Step

URL: View paper

Prior Art Analysis

CoT Image Generation[63] demonstrates prior work on sequential test-time scaling for image generation using verifiers. The candidate paper proposes PARM (Potential Assessment Reward Model) that performs step-by-step verification and refinement through iterative processes, similar to the original paper's OmniVerifier-TTS. Both papers employ reward models to iteratively verify and refine generated images through multiple rounds, achieving improvements on similar benchmarks (GenEval). The candidate's approach of using reward models for sequential verification and the original's use of universal verifiers for iterative refinement represent substantially similar sequential test-time scaling paradigms.

Evidence

Evidence 1 - **Rationale**: Both papers investigate sequential test-time scaling using verifiers for image generation, demonstrating that this approach was explored prior to the original paper's submission. - **Original**: we proposeomniverifier-tts, a sequential test-time scaling paradigm that leverages the universal verifier to bridge image generation and editing within unified models, enhancing the upper bound of generative ability through iterative fine-grained optimization. - **Candidate**: we conduct a systematic investigation into the potential of cot reasoning for autoregressive image generation. we adopt showo [29], a latest discrete generative model, as our baseline, and evaluate on a challenging textto-image generation benchmark: geneval [38]. specifically, we focus on examining ...

Evidence 2 - **Rationale**: Both papers report improvements on GenEval benchmark using sequential test-time scaling with verifiers, showing that similar evaluation approaches and performance gains were achieved in prior work. - **Original**: omniverifier-tts achieves improvements on t2i-reasonbench(+3.7), and geneval++(+4.3), outperforming existing parallel test-time scaling methods, such as best-of-n. - **Candidate**: with parm, our baseline model (showo) is enhanced to achieve leading generation performance. compared to other image generation models in table 3, our best-performing configuration, i.e., integrating parm with iterative dpo in both post-training and test-time, achieves a score of 77%, improving the ...

Evidence 3 - **Rationale**: Both papers describe iterative verification mechanisms where a verifier model evaluates generated images step-by-step and guides refinement, demonstrating similar sequential test-time scaling architectures. - **Original**: as illustrated in fig. 5, we employ omniverifier as a 'misalignment-finder' due to its strong capability in explicit alignment and relational verification. the process begins with umm generating an image from a given prompt. omniverifier then analyzes this image and outputs a binary judgment (true/f... - **Candidate**: parm combines the best of both worlds: 1) it operates adaptively in a step-wise manner, using a potential assessment mechanism to overcome prm's evaluation challenges; and 2) it performs a best-ofn′ selection across n′ (n′ ≤ n) high-potential reasoning paths, thus inheriting orm's advantage. specifi...

Evidence 4 - **Rationale**: Both papers implement iterative refinement loops with stopping criteria based on verifier judgments or maximum iterations, showing similar sequential test-time scaling mechanisms were established in prior work. - **Original**: this iterative refinement loop continues omniverifier returns a true judgment or the maximum number of refinement steps is reached. - **Candidate**: this iterative refinement process continues until parm++ produces a 'yes' in the reflection evaluation, thereby progressively improving both the visual fidelity and the image-text correspondence. we set the maximum number of reflection iterations to 3.

7. Revise: Learning to refine at test-time via intrinsic self-verification

URL: View paper

Brief Assessment

Revise[67] focuses on text-based reasoning tasks (mathematical and coding problems) using self-verification for LLMs, not on image generation and editing with visual verification as in OmniVerifier-TTS.

8. Sdedit: Guided image synthesis and editing with stochastic differential equations

URL: View paper

Brief Assessment

SDEdit[66] focuses on image editing through iterative denoising via stochastic differential equations, not on test-time scaling using verifiers for refinement. The technical approaches differ fundamentally.

9. Inference-time scaling for diffusion models beyond scaling denoising steps

URL: View paper

Brief Assessment

Inference Time Scaling[61] focuses on search algorithms over sampling noises in diffusion models for image generation, not on sequential verification-editing refinement loops using universal verifiers as proposed in the original paper.

10. Generation as search operator for test-time scaling of diffusion-based combinatorial optimization

URL: View paper

Brief Assessment

Generation Search Operator[70] focuses on combinatorial optimization using diffusion models with search-driven generation cycles, not sequential test-time scaling for image generation and editing with verifiers.

Appendix: Text Similarity Detection

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

References

- [0] Generative Universal Verifier as Multimodal Meta-Reasoner View paper
- [1] Llava-cot: Let vision language models reason step-by-step View paper
- [2] Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement View paper
- [3] Zebra-cot: A dataset for interleaved vision language reasoning View paper
- [4] Verify: A benchmark of visual explanation and reasoning for investigating multimodal reasoning fidelity View paper
- [5] Enhancing safety of vision-language reasoning through model-to-model deliberation View paper
- [6] Idealgpt: Iteratively decomposing vision and language reasoning via large language models View paper
- [7] All-in-one: Understanding and Generation in Multimodal Reasoning with the MAIA Benchmark View paper
- [8] Multimodal inconsistency reasoning (mmir): A new benchmark for multimodal reasoning models View paper
- [9] Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing View paper
- [10] WeThink: Toward General-purpose Vision-Language Reasoning via Reinforcement Learning View paper
- [11] Measuring multimodal mathematical reasoning with math-vision dataset View paper
- [12] Latent visual reasoning View paper
- [13] Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models View paper
- [14] Insight-v: Exploring long-chain visual reasoning with multimodal large language models View paper
- [15] Vlr-driver: Large vision-language-reasoning models for embodied autonomous driving View paper
- [16] MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts View paper
- [17] Zero-shot visual reasoning by vision-language models: Benchmarking and analysis View paper
- [18] Vision-language models can self-improve reasoning via reflection View paper
- [19] ChainV: Atomic Visual Hints Make Multimodal Reasoning Shorter and Better View paper
- [20] Open vision reasoner: Transferring linguistic cognitive behavior for visual reasoning View paper
- [21] Prometheus-vision: Vision-language model as a judge for fine-grained evaluation View paper
- [22] Scale Can't Overcome Pragmatics: The Impact of Reporting Bias on Vision-Language Reasoning View paper
- [23] Reasoning-VLA: A Fast and General Vision-Language-Action Reasoning Model for Autonomous Driving View paper
- [24] Unibench: Visual reasoning requires rethinking vision-language beyond scaling View paper
- [25] Mm-verify: Enhancing multimodal reasoning with chain-of-thought verification View paper
- [26] Sci-reason: A dataset with chain-of-thought rationales for complex multimodal reasoning in academic areas View paper
- [27] Vhelm: A holistic evaluation of vision language models View paper
- [28] ClearSight: Visual Signal Enhancement for Object Hallucination Mitigation in Multimodal Large Language Models View paper
- [29] Agent0-VL: Exploring Self-Evolving Agent for Tool-Integrated Vision-Language Reasoning View paper
- [30] MM-CoT: A Benchmark for Probing Visual Chain-of-Thought Reasoning in Multimodal Models View paper
- [31] Training Vision-Language Process Reward Models for Test-Time Scaling in Multimodal Reasoning: Key Insights and Lessons Learned View paper
- [32] Reasoningtrack: Chain-of-thought reasoning for long-term vision-language tracking View paper
- [33] CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs View paper
- [34] Corvid: Improving multimodal large language models towards chain-of-thought reasoning View paper
- [35] Aligning Vision to Language: Text-Free Multimodal Knowledge Graph Construction for Enhanced LLMs Reasoning View paper
- [36] Order Matters: Exploring Order Sensitivity in Multimodal Large Language Models View paper
- [37] VidHalluc: Evaluating Temporal Hallucinations in Multimodal Large Language Models for Video Understanding View paper
- [38] Describe-then-reason: Improving multimodal mathematical reasoning through visual comprehension training View paper
- [39] Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning View paper
- [40] Verite: a robust benchmark for multimodal misinformation detection accounting for unimodal bias View paper
- [41] Measuring and improving chain-of-thought reasoning in vision-language models View paper
- [42] SATORI-R1: Incentivizing Multimodal Reasoning through Explicit Visual Anchoring View paper
- [43] End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models View paper
- [44] Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models View paper
- [45] Perception before reasoning: Two-stage reinforcement learning for visual reasoning in vision-language models View paper
- [46] Why reasoning matters? a survey of advancements in multimodal reasoning (v1) View paper
- [47] Agent-X: Evaluating Deep Multimodal Reasoning in Vision-Centric Agentic Tasks View paper
- [48] Visqa: X-raying vision and language reasoning in transformers View paper
- [49] Visual Reasoning Consistency and Robustness Analysis of Multimodal LLMs View paper
- [50] Mmctagent: Multi-modal critical thinking agent framework for complex visual reasoning View paper
- [51] Visualtrans: A benchmark for real-world visual transformation reasoning View paper
- [52] Fakebench: Probing explainable fake image detection via large multimodal models View paper
- [53] Benchlmm: Benchmarking cross-style visual capability of large multimodal models View paper
- [54] Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi View paper

- [55] Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning View paper
- [56] Grounded Reinforcement Learning for Visual Reasoning View paper
- [57] Beyond seeing: Evaluating multimodal llms on tool-enabled image perception, transformation, and reasoning View paper
- [58] Generative hierarchical features from synthesizing images View paper
- [59] An Efficient Rubric-based Generative Verifier for Search-Augmented LLMs View paper
- [60] MedVLSynther: Synthesizing High-Quality Visual Question Answering from Medical Documents with Generator-Verifier LMMs View paper
- [61] Inference-time scaling for diffusion models beyond scaling denoising steps View paper
- [62] Video-t1: Test-time scaling for video generation View paper
- [63] Can We Generate Images with CoT? Let's Verify and Reinforce Image Generation Step by Step View paper
- [64] Scalingnoise: Scaling inference-time search for generating infinite videos View paper
- [65] From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning View paper
- [66] Sdedit: Guided image synthesis and editing with stochastic differential equations View paper
- [67] Revise: Learning to refine at test-time via intrinsic self-verification View paper
- [68] Let's Verify and Reinforce Image Generation Step by Step View paper
- [69] Scaling Inference Time Compute for Diffusion Models View paper
- [70] Generation as search operator for test-time scaling of diffusion-based combinatorial optimization View paper