Multimodal Test-Time Scaling via Visual Augmentation

Swadesh Sistla
University of Washington
swad@uw.edu

Jeffrey Lee
University of Washington
jlee622@uw.edu

Jack Xiao
University of Washington
xiaohl@uw.edu

Abstract

Vision-Language Models (VLMs) have demonstrated remarkable capabilities by integrating Large Language Model backbones with vision encoding capabilities. While test-time scaling techniques have improved VLM performance, current methods predominantly focus on enhancing textual processing, leaving the scaling of visual processing abilities largely underexplored. This paper proposes a novel, vision-centric approach to test-time scaling for VLMs via multiple augmentation experiments (systematic, manual, and hybrid). We aim to explicitly enable these models to reason over augmented visual contexts by applying systematic visual augmentations inspired by the Visual Sketchpad approach, including depth maps, segmentation masks, crops, and bounding boxes. We hypothesize that by introducing these diverse visual augmentations of an original image into the model's context at test time, VLMs will achieve richer multimodal reasoning and improved performance on visual and multimodal tasks. Our findings from experiments performed an open-source VLM, LLaVA-OneVision, indicate that strategic augmentation in a Sketchpad-inspired fashion outperforms both randomly applied and hand-selected augmentations, which mildly degraded or stabilized performance. Unlike this degradation, the 'lightweight Sketchpad' strategy improves on the LLaVA-OneVision baseline despite lower-cost and lower-capability substitutions for each element of the Sketchpad pipeline.

1. Introduction

The landscape of artificial intelligence research has been significantly reshaped by the advent of large VLMs: multimodal systems that can process both images and text. By combining Large Language Model (LLM) backbones with vision encoding capabilities, VLMs have achieved high performance across a diverse array of tasks, including object detection [1], visual grounding [2], and multimodal reasoning [3].

In both LLMs and VLMs, recent breakthroughs have been made in test-time scaling: post-training approaches that improve model performance at inference time. Current approaches to improve test-time scaling in VLMs mirror successful methods for LLMs, which focus predominantly on textual processing capabilities. These methods include structured inference techniques, reinforcement learning, and hybrid scaling strategies, applied primarily to language capabilities through the production of text tokens [4]. While VLMs demonstrate strong performance across diverse downstream tasks, significant opportunities remain to improve multimodal reasoning capabilities beyond the scope of existing methods [1].

As current approaches to VLM test-time scaling primarily enhance language processing capabilities, scaling the visual processing abilities of these models at inference remains largely underexplored. We propose a new, vision-centric approach to test-time scaling for VLMs that explicitly enables these models to reason over augmented visual contexts, inspired by the Visual Sketchpad approach [5]. By applying Sketchpad-inspired augmentation techniques and using these as visual inputs at test time — with augmentations such as depth maps, segmentation masks, crops, and bounding boxes — we intend to enable richer multimodal reasoning compared to traditional language-centric approaches.

Although we expect this strategy to improve VLM performance, there are several challenges it may introduce. One such challenge is the fact that models are not explicitly pre-trained to handle multiple variations of the same image. Though VLMs can process multi-image inputs, they may not inherently recognize that different augmentations represent the same underlying content—this distinction from typical multi-image scenarios may require further investigation. Another possible challenge is the increase in computational load: though traditional test-time approaches focus on language tokens, our approach introduces additional complete image representations. Thus, each forward pass may become significantly more compute-intensive as multiple augmented images increase sequence length and require simultaneous processing.

Despite these possible challenges, we believe that enabling VLMs to reason with images through test-time scaling represents a promising research direction. To test this hypothesis, we plan to perform the following evaluations on LLaVA-OneVision, an open-source VLM with strong multimodal reasoning capabilities [6]. We will first reproduce LLaVA-OneVision's results on a testbed of reasoning tasks using single images to establish baseline metrics. Next, we will assess performance improvements by applying various augmentations (e.g. image pyramids, regional crops, and visual annotations) purely at inference time. Based on results in this inference-only setting, we will explore fine-tuning LLaVA-OneVision on augmented visual inputs created using Sketchpad-inspired techniques, explicitly encouraging the model to reason across multiple transformed views.

We expect that test-time image augmentation will result in stronger VLM performance by enabling richer multimodal reasoning capabilities over existing approaches. By introducing augmentations of the original image into the model's context, we hypothesize that models will acquire stronger visual reasoning capabilities, leading to improved performance on visual and multimodal reasoning tasks.

2. Related Work

Vision-language models have rapidly evolved to become powerful multimodal systems capable of understanding and reasoning across visual and textual modalities. While significant progress has been made in developing these models, current approaches to improving their performance at inference time have primarily focused on language-based scaling methods, leaving visual processing enhancement largely underexplored.

In this section, we review the current landscape of test-time scaling approaches, examining both the established methods in language models and their adaptation to vision-language contexts. We then explore existing work in visual prompting and image augmentation techniques that have shown promise for enhancing VLM capabilities. Finally, we discuss recent advances in multi-image understanding, highlighting both the potential and limitations of current approaches. This review reveals a significant gap: while various visual enhancement techniques exist, there is limited work on systematic visual test-time scaling that leverages multiple complementary augmentations to improve reasoning performance without requiring model retraining.

2.1. Test-time Scaling in Vision-Language Models

Test-time scaling has emerged as a powerful paradigm for improving model performance without additional training. The foundational work by Wei et al. introduced chain-of-thought (CoT) prompting, showing significant success in language models [7]. This approach showed large improvements on arithmetic reasoning benchmarks, such as GSM8k.

Recent advances have further expanded test-time scaling capabilities. Most notably, OpenAI's o1-preview model demonstrates significant improvement in areas such as coding, scientific reasoning, language processing, and creative problem-solving. Models are given more time to ”think” before responding, leading to significant improvements on reasoning tasks [9].

Despite significant progress in LLM test-time scaling, analogous approaches in VLM models remain somewhat unexplored. Current VLM scaling methods primarily mirror LLM approaches, focusing on generating more language tokens during inference time to produce longer chains of thought. Zhang et al. introduced multimodal CoT reasoning, extending CoT to incorporate both text and visual inputs [8]. Their approach is not about scaling the visual processing itself, but rather generating better language-based reasoning similar to LLM approaches.

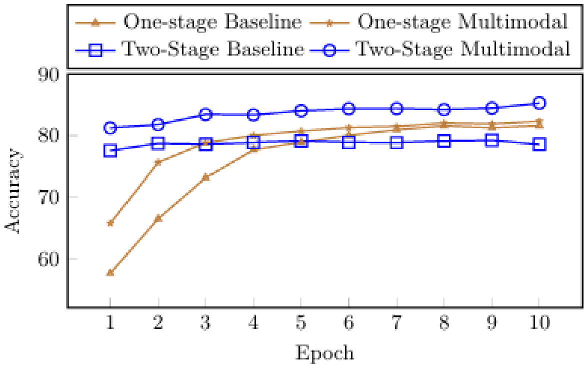

Figure 1. Validation accuracy curves of the No-CoT baseline and Multimodal-CoT variants from ”Multimodal Chain-of-Thought Reasoning in Language Models.” [8].

2.2. Visual Prompting and Image Augmentation

Visual prompting has emerged as an effective technique for enhancing VLM performance through strategic input modifications. The Visual Sketchpad work by Hu et al., which inspires our image augmentation approach, demonstrates how models can benefit from intermediate visual artifacts, such as auxiliary lines, depth maps, and segmentation masks, as reasoning aids during problem-solving [5]. This iterative approach achieved substantial gains of 12.7% on mathematical tasks and 8.6% on vision tasks, demonstrating the potential of visual-centric scaling.

However, existing visual prompting methods often apply single augmentations rather than multiple complementary perspectives. We plan to address these limitations by developing a systematic approach to visual test-time scaling that applies multiple complementary augmentations.

2.3. Multi-Image Understanding and Reasoning

Current multi-image approaches primarily focus on architectural improvements and specialized training on interleaved image-text data, such as MM-Interleaved [10] and LLaVA-NeXT-Interleave [11]. However, these methods assume multiple distinct images as input rather than leveraging multiple perspectives of the same scene.

Our work explores a different paradigm: using multiple complementary augmentations of a single image to enhance reasoning capabilities. Drawing inspiration from Visual Sketchpad's multi-perspective approach, we investigate both inference-time application of augmentations and training methods to help models better utilize these augmented inputs for improved reasoning performance.

3. Methodology

Our methodology investigates whether visual augmentations can improve VLM performance when applied in experiments that to varying degrees simulate Visual Sketchpad-style iterative reasoning. We conduct three distinct experimental approaches to isolate the impact of visual augmentation from semantic reasoning guidance, and to attempt to find a maximally general (but experimentally sound) conclusion.

3.1. Design Overview

We design our experiments to systematically evaluate the contribution of visual augmentations to VLM performance across three paradigms:

- Systematic Augmentation Approach: Applying predefined sets of visual augmentations without model-guided selection
- Manual Augmentation Approach: Hand-crafted augmentations targeting specific failure cases
- Hybrid Approach: Using GPT-4o-mini for augmentation selection but LLaVA-OneVision for final inference

3.2. Base Model and Data

Vison-Language Model. We use LLaVA-OneVision as our primary evaluation model, chosen for its strong multimodal reasoning capabilities and publicly available weights.

Data. The primary dataset for our experiments is Multimodal Bench (MMBench) [13]. This dataset provides comprehensive evaluation frameworks for multimodal large language models across diverse tasks and domains. It offers an all-around assessment framework testing model versatility and robustness in multimodal understanding across diverse domains, evaluating whether models can perform consistently across various task types.

3.3. Core Techniques

Our experimental approach leverages a combination of established models, augmentation strategies, and varied input presentation methods.

3.3.1 Visual Augmentation Strategy

An essential element of the Visual Sketchpad approach is the breadth of tools available. We provide the full set of Visual Sketchpad augmentations, with a modification: we use lightweight, off-the-shelf HuggingFace models in lieu of the compute-intensive, specialized 'vision experts' used in the original work. For each original input image, augmentations may include:

- **Segmentation and Marking (`segment_and_mark`) :** Generating images where segments are identified, highlighted with colored masks, and labeled with numerical identifiers. This also provides bounding boxes for each segment, aiding in spatial reasoning and object relationship understanding.

- **Object Detection (`detection`):** Producing annotated images where specified objects are detected and enclosed in bounding boxes, labeled with the object name and an index. This allows the model to focus on specific entities of interest.

- **Depth Estimation (`depth`):** Creating depth maps of the input image, represented using a colormap (e.g., Inferno) to indicate the relative distance of objects from the camera. This can enhance understanding of 3D structure and spatial relationships.

- **Cropping and Zooming (`crop_image`, `zoom_in_image_by_bbox`):**
  - `crop_image`: Generating new images by cropping the original image based on normalized coordinates, effectively zooming into specific regions.
  - `zoom_in_image_by_bbox`: Cropping the image based on a given bounding box (often from detection results) with optional padding, allowing for a focused view of detected objects.

- **Sliding Window Detection (`sliding_window_detection`):** For scenarios where objects might be too small for standard detection, this technique involves creating smaller, zoomed-in patches of the image and running object detection on these patches. This can yield annotated patches containing the objects of interest.

- **Image Overlay (`overlay_images`):** Combining two images with transparency, for example, overlaying an annotated image (like a segmentation mask or bounding boxes) onto a depth map or the original image. This can help in visualizing multiple layers of information simultaneously.

These specific augmentations, derived directly from the Visual Sketchpad's toolkit, form the basis of the visual information provided to the VLM in addition to the original image.

3.3.2 Approach 1: Systematic Augmentation

Methodology. For each input image, we systematically apply 3-7 augmentations from our toolkit, creating a comprehensive set of visual perspectives. The augmented images are concatenated and fed to LLaVA-OneVision alongside the original image and task query. Rather than using Visual Sketchpad's iterative approach where a model decides which tools to apply, we systematically apply these proven augmentation techniques to create multiple perspectives of each input image. This approach maintains the benefits of diverse visual representations while avoiding the computational overhead of iterative model conversations [5]. This allows for the creation of a rich set of visual inputs beyond the original image.

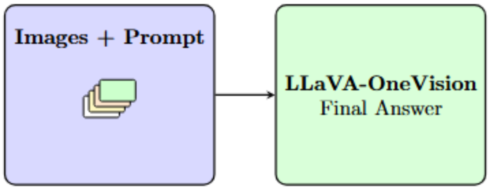

Figure 2. LLaVA-OneVision processing stage. The model receives multiple augmented image views (represented by stacked layers) along with question prompts to generate the final answer.

Hypothesis. Multiple complementary visual perspectives will enhance the model's reasoning capabilities through richer visual context.

3.3.3 Approach 2: Manual Augmentation for Failure Cases

Methodology. We identify specific examples where the baseline LLaVA-OneVision model fails on benchmark tasks. For these failure cases, we manually create targeted augmentations designed to highlight the visual information necessary for correct answers. This includes strategic cropping of key regions, enhanced object detection overlays, and focused zoom-ins on relevant image areas. **Hypothesis:** With optimal augmentations specifically designed to address model failures, performance improvements will be much higher.

3.3.4 Approach 3: Hybrid GPT-4o-mini Planning with LLaVA-OneVision Inference

Methodology. We leverage GPT-4o-mini's planning capabilities to determine appropriate augmentations for each task, following the Visual Sketchpad approach. However, instead of using GPT-4o-mini for final inference, we pass the GPT-4o-mini-selected augmentations to LLaVA-OneVision for answer generation.

1. GPT-4o-mini receives the original image and task query
2. GPT-4o-mini generates a plan and creates appropriate visual augmentations
3. The original image and GPT-4o-mini-generated augmentations are fed to LLaVA-OneVision
4. LLaVA-OneVision provides the final answer

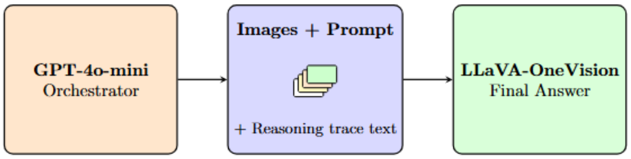

Figure 3. Two-model cascade architecture. GPT-4o-mini orchestrates visual analysis tools to generate augmented image views and reasoning traces, which are then processed by LLaVA-OneVision for final answer generation.

Hypothesis. Strategic augmentation selection will improve performance, but the benefits may be limited by the target model's ability to effectively utilize the augmented visual information.

4. Experiments

We first establish baseline performance metrics for LLaVA-OneVision on the MME benchmark using standard single-image inputs. This provides our reference point for measuring the impact of various augmentation strategies.

4.1. Systematic Augmentation Results

Setup. We evaluate LLaVA-OneVision performance when provided with systematically generated augmentations across all the benchmarks.

Results. Contrary to our hypothesis, systematic augmentation consistently degraded model performance. The additional visual information appeared to confuse rather than assist the model's reasoning process. Performance decreased by an average of -1.57% across benchmark tasks, suggesting that without guided selection, multiple visual perspectives create a 'cognitive overload', adding additional noise rather than enhanced understanding.

Failure on this experiment prompted us to analyze the failure cases thoroughly, thus, we conducted our second experiment.

4.2. Manual Augmentation Analysis

Setup. We manually analyze 30 failure cases from the baseline MMBench evaluation, creating optimal augmentations for each case designed to highlight the visual information needed for correct answers.

Results. Even with carefully crafted augmentations targeting specific visual reasoning challenges, improvements were sporadic and inconsistent. While we occasionally observed correct answer flips, these successes showed no discernible pattern and were not consistently related to the quality or relevance of our augmentations.

As we could not find a more general approach to generating the visual augmentations that effectively increased task performance, we performed Experiment 3 to simulate a 'lightweight' approximation of Visual Sketchpad [5].

4.3. Hybrid Approach Evaluation

Setup. We implement the hybrid approach where GPT-4o-mini generates augmentations following Visual Sketchpad methodology, but LLaVA-OneVision performs the final reasoning and answer generation.

Results. This approach yielded the most promising results, though improvements remained modest. The strategic selection of augmentations by GPT-4o-mini appeared to reduce the cognitive load observed in systematic augmentation, while the targeted visual enhancements provided some benefit to LLaVA-OneVision's reasoning process.

5. Results

Our experiments across three distinct approaches reveal the complex relationship between visual augmentation and VLM performance. We present results for each experimental paradigm, demonstrating the challenges and opportunities in vision-centric test-time scaling.

5.1. Baseline Performance

LLaVA-OneVision achieved baseline performance of 85.57% on MMBench using standard single-image inputs. This establishes our reference point for measuring augmentation impact across perception and cognition tasks.

5.2. Systematic Augmentation Results

Contrary to our initial hypothesis, systematic application of 3-7 augmentations consistently degraded model performance. LLaVA-OneVision's accuracy decreased by an average of 1.57% on all benchmarks.

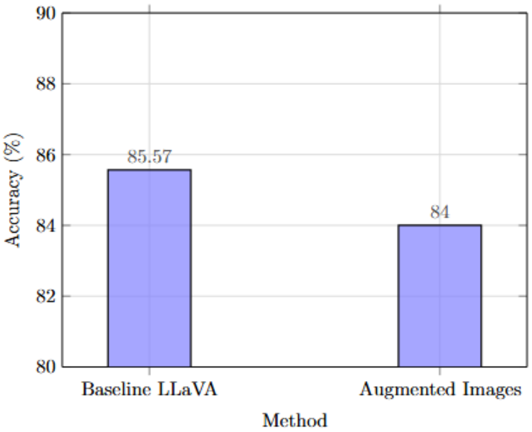

Figure 4. Performance comparison on MMBench dataset. Baseline LLaVA achieves 85.57% accuracy while the cascade approach with augmented images achieves 84.0% accuracy, showing a modest decrease of 1.5 percentage points.

Unguided visual augmentation harms rather than helps VLM performance, indicating that the quantity of visual information is less important than its strategic selection and presentation.

5.3. Manual Augmentation Analysis

Analysis of 30 carefully selected failure cases revealed limited and inconsistent benefits from hand-crafted augmentations. Despite targeting specific visual reasoning challenges with optimal cropping, object detection overlays, and strategic zoom-ins, only 3 samples of failure cases were successfully converted to correct answers.

Importantly, these successes showed no discernible pattern related to augmentation quality or relevance. This suggests that even optimal augmentations designed specifically to address model failures provide minimal and unpredictable performance gains.

5.4. Hybrid Approach Evaluation

The hybrid approach, utilizing GPT-4o-mini for augmentation planning and LLaVA-OneVision for inference, yielded the most promising results. Performance improved to 86.94% accuracy, representing a 1.37 percentage point gain over baseline.

This shows that the inference models benefits from the context of the augmented images when seeking to reason

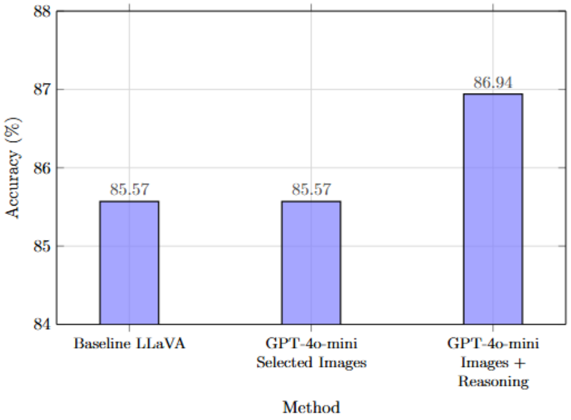

Figure 5. Performance comparison on MMBench dataset across three approaches. Baseline LLaVA achieves 85.57% accuracy. GPT-4o-mini selected augmentations alone show no improvement (85.57%), while adding reasoning traces yields a 1.37 percentage point improvement to 86.94% accuracy.

about them. Without this context, we hypothesize that the model may treat the augmented views as noise rather than meaningful input it can reason about. It also highlights that using augmentations selected by GPT-4o-mini yields at least equivalent performance to baseline, rather than the degradations witnessed in earlier experiments.

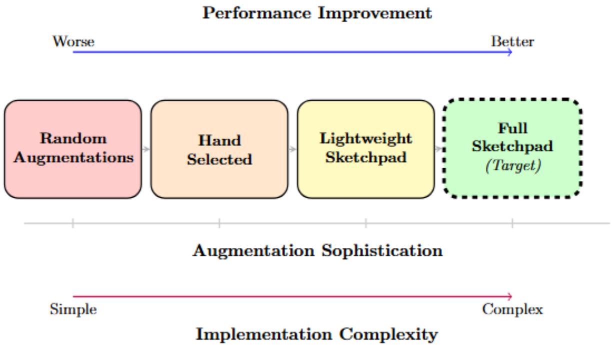

Figure 6. Spectrum of visual augmentation approaches showing the relationship between implementation sophistication and performance gains. *Note: Full Sketchpad results were not computed on this dataset, but served as inspiration for our lightweight approach.*

6. Discussion

6.1. Analysis of Results

Our experimental results reveal three key insights from our comprehensive analysis:

Strategic Selection Outperforms Systematic Application. The stark contrast between our systematic augmentation failures (-1.57%) and the promising results from our hybrid approach demonstrates that the quality and relevance of visual augmentations matter significantly more than their quantity. Furthermore, our ability to hand-select augmentations was outperformed by GPT-4o-mini, suggesting that with the proper augmentations, models can perform at least at baseline.

Contextual Reasoning. Our hybrid approach's performance compared to manual augmentation suggests the importance of providing reasoning context alongside visual enhancements. When GPT-4o-mini generates both augmentations and reasoning traces, LLaVA-OneVision can more effectively leverage the visual information. This suggests that visual augmentations alone may be insufficient, in violation of our hypothesis — models may require semantic guidance to interpret and integrate multiple visual perspectives effectively.

Lightweight Architecture Maintains Effectiveness. Our hybrid approach demonstrates that a simplified version of the Visual Sketchpad architecture can achieve meaningful performance improvements while reducing computational complexity. By separating the orchestration (GPT-4o-mini) and execution (LLaVA-OneVision) stages and using off-the-shelf tools for augmentations, we maintain the core benefits of guided visual augmentation without requiring the resource base of the original Sketchpad. Furthermore, because the models themselves have reduced parameter sizes, total computation is greatly reduced. This architectural choice proves that strategic visual enhancement can be effective across different model configurations, suggesting that the fundamental principles of contextual visual reasoning are transferable to more efficient implementations.

6.2. Implications for VLM Test-Time Scaling

Our findings have important implications for the broader field of VLM enhancement and test-time scaling research:

Reconsidering Vision-Centric Scaling Paradigms. While our initial hypothesis suggested that enhancing visual processing could complement language-centric scaling approaches, our results indicate a more complex relationship. The gains from our hybrid approach suggest that vision-centric scaling is viable but requires sophisticated integration mechanisms rather than simple input augmentation.

Training-Inference Gap in Visual Processing. The limited effectiveness of even carefully crafted manual augmentations suggests a significant gap between how VLMs are trained and how they process multiple visual inputs at inference time. Current VLMs may require explicit training on augmented visual inputs to fully leverage their potential.

Computational Trade-offs in Test-Time Enhancement. Our hybrid approach demonstrates that meaningful performance improvements in VLMs often require increased computational investment. However, the gains achieved suggest that strategic augmentation selection can provide better cost-benefit ratios than comprehensive augmentation pipelines, offering a practical path forward for real-world applications.

6.3. Limitations and Future Work

Model and Dataset Scope. Our evaluation focuses primarily on LLaVA-OneVision and MMBench, which may not fully represent the potential of vision-centric scaling across different VLM architectures and task domains. Future work should evaluate these approaches across a broader range of models, particularly those specifically designed for multi-image reasoning or with different vision-language fusion mechanisms.

Augmentation Strategy Limitations. While our augmentation toolkit draws from the proven Visual Sketchpad approach, it uses off-the-shelf models in lieu of large specialized 'vision experts'. Beyond simply replicating these experts, future research could explore learned augmentations, domain-adaptive visual modifications, or adversarial visual prompts that might be more effective for specific reasoning challenges. Additionally, investigating the optimal number and types of augmentations for different task categories could yield more targeted improvements.

Integration of Training and Inference Enhancement. Our work focuses exclusively on inference-time improvements, but our results indicate that training-time integration of visual augmentation techniques may be necessary for substantial performance improvements. Future work should investigate multi-stage training approaches that first teach models to understand individual augmentations, then progress to reasoning across multiple visual perspectives simultaneously.

Multimodal Integration Opportunities. Our findings suggest that visual augmentation may be most effective when combined with complementary enhancements in other modalities. Future work should explore integrated approaches that simultaneously enhance visual, textual, and reasoning capabilities, potentially through unified frameworks that optimize augmentation strategies across all input modalities rather than treating them independently.

7. References

1. Feng, Yongchao, et al. ”Vision-Language Model for Object Detection and Segmentation: A Review and Evaluation.” arXiv preprint arXiv:2504.09480 (2025).

2. Wang, Shijie, et al. ”Learning visual grounding from generative vision and language model.” 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2025.

3. Liu, Shuhang, et al. ”MMC: Iterative Refinement of VLM Reasoning via MCTS-based Multimodal Critique.” arXiv preprint arXiv:2504.11009 (2025).

4. Zhang, Qiyuan, et al. ”What, how, where, and how well? a survey on test-time scaling in large language models.” arXiv preprint arXiv:2503.24235 (2025).

5. Hu, Yushi, et al. ”Visual sketchpad: Sketching as a visual chain of thought for multimodal language models.” arXiv preprint arXiv:2406.09403 (2024).

6. Li, Bo, et al. ”Llava-onevision: Easy visual task transfer.” arXiv preprint arXiv:2408.03326 (2024).

7. Wei, Jason, et al. ”Chain-of-thought prompting elicits reasoning in large language models.” Advances in Neural Information Processing Systems 35 (2022): 24824-24837.

8. Zhang, Zhuosheng, et al. ”Multimodal Chain-of-Thought Reasoning in Language Models.” arXiv preprint arXiv:2302.00923 (2024).

9. Zhong, Tianyang, et al. ”Evaluation of OpenAI o1: Opportunities and Challenges of AGI.” arXiv preprint arXiv:2409.18486 (2024).

10. Tian, Changyao, et al. ”MM-Interleaved: Interleaved Image-Text Generative Modeling via Multi-modal Feature Synchronizer.” arXiv preprint arXiv:2401.10208 (2024).

11. Li, Feng, et al. ”LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models.” arXiv preprint arXiv:2407.07895 (2024).

12. Liu, Yuan, et al. ”MMBench: Is Your Multimodal Model an All-around Player?” arXiv preprint arXiv:2307.06281 (2024).