AMVICC: A Novel Benchmark for Cross-Modal Failure Mode Profiling for VLMs and IGMs

Anonymous ACL submission

Abstract

We investigated visual reasoning limitations of both multimodal large language models (MLLMs) and image generation models (IGMs) by creating a novel benchmark to systematically compare failure modes across image-to-text and text-to-image tasks, enabling cross-modal evaluation of visual understanding. Despite rapid growth in machine learning, vision language models (VLMs) still fail to understand or generate basic visual concepts such as object orientation, quantity, or spatial relationships, which highlighted gaps in elementary visual reasoning. By adapting MMVP benchmark questions into explicit and implicit prompts, we create *AMVICC*, a novel benchmark for profiling failure modes across various modalities. After testing 11 MLLMs and 3 IGMs in nine categories of visual reasoning, our results show that failure modes are often shared between models and modalities, but certain failures are model-specific and modality-specific, and this can potentially be attributed to various factors. IGMs consistently struggled to manipulate specific visual components in response to prompts, especially in explicit prompts, suggesting poor control over fine-grained visual attributes. Our findings apply most directly to the evaluation of existing state-of-the-art models on structured visual reasoning tasks. This work lays the foundation for future cross-modal alignment studies, offering a framework to probe whether generation and interpretation failures stem from shared limitations to guide future improvements in unified vision-language modeling.

1 Introduction

Recently, multi-modal models have improved significantly and have shown proficiency in several fields with emergent capabilities (Stability AI, 2024). However, recent work has highlighted that despite their strength in visual reasoning, instruction following, and image understanding proficiencies, many fail to consistently and accurately answer straightforward visual understanding questions that most humans find trivial (Anis et al., 2025). The extensive visual shortcomings of MLLMs and VLMs have been defined and tested in benchmarks such as MediConfusion, GMAI-MMBench, and the MMVP Benchmark (Sepehri et al., 2024; Chen et al., 2024; Tong et al., 2024).

Compared to other generative model modalities, IGMs are steadily improving: Gemini 2.5 Flash Image and OpenAI's DALL·E 3 revolutionized instruction following and realism within image generation (Fortin et al., 2025; OpenAI, 2024b). However, despite their drastic growth, IGMs demonstrate similar elementary failures in generating images that align with given prompts, especially those with a complex combination of entities, attributes, and spatial relationships (Marioriyad et al., 2025; Gokhale et al., 2023). Several benchmarks and metrics have attempted to classify failure modes (for example, quantitative shifts, attribute comparisons, spatial relationships) to identify prospective points of improvement such as VisuLogic, VISOR, T2I-CompBench, and SAGE (Xu et al., n.d.; Huang et al., 2021).

However, there is a notable lack of research comparing visual reasoning and generations between models in IGMs and MLLMs. In this paper, we extend the work done by Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs to profile the cross-modal failure modes in visual reasoning and recognition of MLLMs and image generation models (Tong et al., 2024). We created matching generation prompts based on the MMVP-Benchmark to evaluate failure mode similarities between the two modalities. Through these tests, we hoped to uncover insights into the elementary visual shortcomings of image generation models and multi-modal LLMs. In this paper, we introduce a novel benchmark, **Assessment** of **Modality-Specific Visual Intelligence Comprehension** and

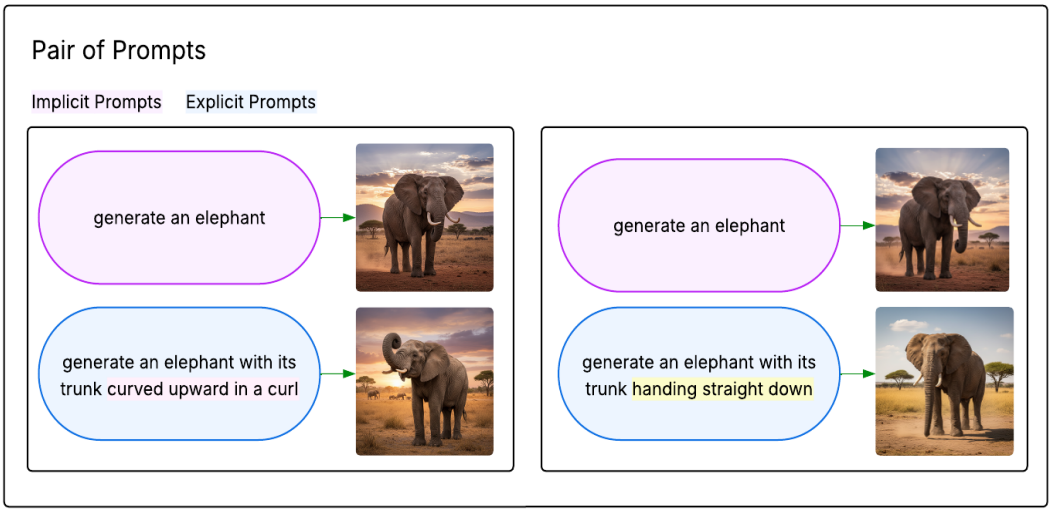

Figure 1: Comparison of implicit prompts to explicit prompts for the 2 pictures in a pair

Creation (AMVICC), to evaluate the failure modes of image generation models and MLLMs with the same contextual input and provide analysis of tests completed on current state-of-the-art models.

2 Methods

In this section, we explain our evaluation of the following Vision Language Models: Meta: Llama 3.2 90B Vision Instruct (90 billion parameters), Meta: Llama 4 Maverick (17 billion active parameters and 128 experts), Meta: Llama 4 Scout (17 billion active parameters and 16 experts), xAI: Grok 4, Google: Gemma 3 27B (27 billion parameters), Google: Gemini 2.5 Pro, OpenAI: GPT-4o, Qwen: Qwen2.5 VL 72B Instruct (72 billion parameters), Mistral: Pixtral Large 2411 (124 billion parameters), Anthropic: Claude Opus 4.1, and Anthropic: Claude Sonnet 4, on a modified version of the MMVP Benchmark (Meta, 2024, 2025; xAI, 2025; Gemma Team, 2025; Gemini Team, 2025; OpenAI, 2024a; Qwen Team, 2025; Mistral AI, 2024; Anthropic, 2025a,b). To our modified version, we add categories to the MMVP Benchmark to match the tasks of the visual understanding questions. We also evaluated the following Image Generation Models: OpenAI: DALL·E 3, Google: Gemini 2.5 Flash Image, & Stability AI: Stable Diffusion 3.5 Large (8.1 billion parameters) on *AMVICC*, which contains prompts constructed to mirror the questions on the MMVP Benchmark with corresponding categories (OpenAI, 2024b; Fortin et al., 2025; Stability AI, 2024). VLM models were chosen to provide a variance across open-source and closed-source models while also providing variability across model size, architecture, and training methods. Due to a smaller selection of state-of-the-art image generation models due to access and availability, we were only able to choose 3 models with variance across providers, training data, architecture, and size.

2.1 Prompting Procedure

To evaluate model performance in both directions (image → text and text → image), we used 300 original MMVP benchmark questions, and 600 additional prompts were created [found in Appendix A] to probe specific failure modes.

- **For VLMs:** For VLMs, the benchmark questions were paired with MMVP images, and resulting answers were graded by GPT-4o to determine model accuracy.
- **For Image Generation Models:** For image generation, we (4 authors) designed handcrafted explicit and implicit prompts derived from those questions in order to test corresponding tasks in image generation models with 2 consequent checks for correct structure and prompting style.

These mixed evaluation methods were utilized due to the known inaccuracy of VLMs when evaluating images for positioning and elementary understanding, while proven to be accurate with text. This methodology also mirrors the evaluation methodology of the MMVP benchmark evaluation to summarize outputs from the VLMs into the final multiple choice answer (for example, (a) or (b)). Our

implicit prompts were created by defining the generalized situation between a pair of images in order to establish the foundation of a model's ability to generate the background. Afterwards, each explicit prompt, correlating to an MMVP question, added the element required by the correct answer choice of the corresponding MMVP question. They clearly define the required visual concept while implicit prompts used more natural, generalized phrasing to create a prompt relevant to both questions and correct answer choices. There are a total of 600 prompts with 300 implicit prompts and 300 explicit prompts. Each implicit prompt tests an image generation model's ability to generate a scenario while each explicit prompt tests an image generation model's ability to change the generated image to satisfy a specific newly-added component (for example, dog in grass is implicit and dog in grass looking to the right is explicit) similar to the way the MMVP benchmark tests a model's ability to visually understand a specific component.

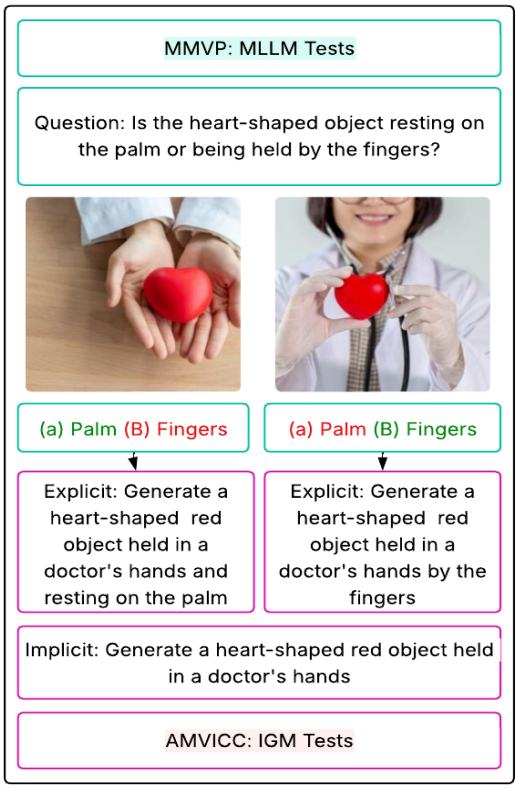

Figure 2: Diagram of the AMVICC Creation Pipeline: We created implicit prompts based off of the general scenario introduced by the question and created explicit prompts by adding specifics in line with the specific answer choice for that image ID.

2.2 Evaluation

To systematically evaluate the performance of both vision language models and image generation models, we determined the success and failure of each question-answering task. We designed a rubric that defines success and failure based on the intended visual understanding goals of the MMVP benchmark. We began by testing various VLMs [detailed in the beginning of section 3] based on the 300 questions that were present in the original MMVP dataset, and then moved to testing various IGMs with the additional prompts we created based on those questions [detailed in 3.1]. By performing this, we were able to highlight certain aspects that automated benchmarks might not have been able to catch. This was intended to measure the visual understanding goals of the MMVP benchmark and of each prompt based on the visual understanding of the *AMVICC* benchmark.

Images generated by IGMs are evaluated differently depending on if the prompt is implicit or explicit. An implicit image is considered correct if it satisfies all components of the provided prompt, regardless of whether it matches the corresponding question's distinction. An explicit image is considered correct if it generates the specific feature that the prompt asks for based on the corresponding visual understanding question and category from our modified MMVP benchmark. Each image is scored by human evaluators and double-checked for accuracy in order to check for bias and ensure correct grading for each image.

3 Results

We evaluated the accuracy of 11 multi-modal LLMs in visual reasoning and understanding tasks through the use of the MMVP benchmark (Tong et al., 2024). Accuracy scores depict the model's accuracy across the 9 categories of visual reasoning questions. After our evaluation of these 11 models, we extended our experiments to 3 image generation models using our AMVICC benchmark to evaluate each model's proficiency in generating images across the 9 categories. The thresholds for failure modes are 80% or below for individual accuracies and 70% or below for pair accuracies. This applies to both MLLMs and IGMs. Each pair of the questions are only considered correct if both questions are answered correctly or both images generated are aligned with their respective explicit prompts.

3.1 MLLM Score Analysis

Many of the models shared the same failure modes; however, some of the models had failure modes that served as outliers. For example, in both the Orientation and Direction and the Quantity and Count categories, the individual VLM accuracies for xAI: Grok 4 were 40.00% and 50.00%, respectively (see Table 1). In the Viewpoint and Perspective category, xAI: Grok 4 and Anthropic: Claude Sonnet 4 were outliers, both attaining an accuracy of 55.56%, a notable 16.66% difference from the next highest accuracy. That being said, an opposite trend is evident in Table 2, which displays the pair VLM accuracies. Instead of the outliers being a failure mode, they are the highest accuracy for models such as Meta: Llama 3.2 90B Vision Instruct and Meta: Llama 4 Maverick. This is exhibited in the Positional and Relational Context as well as the Viewpoint and Perspective categories for Meta: Llama 3.2 90B Vision Instruct. This trend is apparent in the Quantity and Count category for Meta: Llama 4 Maverick. Consequently, this trend highlights the fact that certain models succeeded where either all or most of the other models failed. Furthermore, most MLLMs fail in similar contexts, particularly in Positional and Relational Context and Quantity and Count. Additional common failure modes include Viewpoint and Perspective as well as Orientation and Direction. However, model-specific failure modes occurred as well, with only Grok 4 failing on Color and Appearance, and four models out of eleven (Google: Gemini 2.5 Pro, Grok 4, Google: Gemma 3 27B, and Anthropic: Claude Opus 4.1) failing on visual reasoning within the category of Structural and Physical Characteristics (see Table 1). This suggests a variance of failure modes for certain models in addition to the common failure modes.

Llama 3.2 90B Vision-Instruct achieved the highest performance with one pair failure mode in Quantity and Count and no defined individually-measured failure modes, indicating stronger visual understanding and reasoning for similar pictures compared to other models. Conversely, Grok 4 performed the worst with only one category above the benchmark for failure modes. Llama 4 Maverick and Llama 4 Scout are both from the same LLM family but contain key differences in architecture and structural setup. Maverick is attuned to high-performance generation and implementation with 17 billion active parameters for each of the 128 experts in the MoE (mixture-of-experts architecture outlined in (Meta, 2025)), totaling 400 billion parameters. This is larger than Scout's input-focused architecture with 17 billion active parameters and 16 experts in MoE, totaling 109 billion parameters. Mixture-of-experts utilizes gating networks, which essentially direct certain inputs to experts. Experts are smaller models meant for specific tasks that are part of the MLLM. The benefit of experts is that these smaller models can process the inputs without the entire MLLM having to be utilized, and this, in turn, would augment the MLLM's efficiency. Since the entire model isn't being used, only some of its parameters are going to be active, and this is why, for example, Maverick only has 17 billion active parameters out of its 400 billion total parameters. On this note, Maverick's MoE architecture is represented as 17Bx128E whereas Scout's MoE architecture is represented as 17Bx16E. However, both models perform relatively the same with Llama 4 Maverick performing only slightly better.

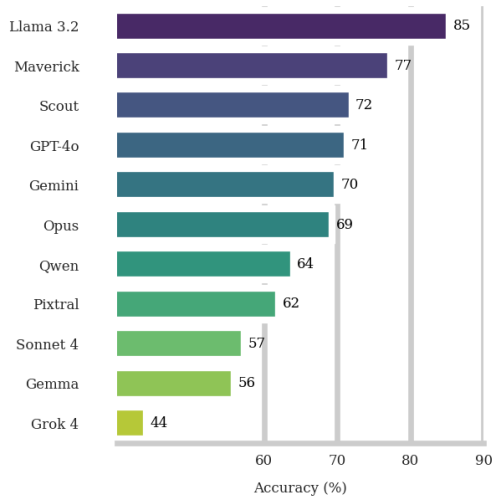

Figure 3: Benchmark results of current MLLMs: We evaluate pair accuracy across 11 models based on the questions and images from the MMVP dataset.

3.2 IGM Score Analysis

The IGMs overall shared 2 common failure modes across pair explicit and individual explicit accuracies for each of the three models: quantity and count (qc) and text (tx) (see Tables 3 and 4). Majority of the models (2/3) also exhibited failure modes for both pair and individual explicit accuracy in categories structural and physical (sh), orientation

Model	Params Size (B)	🔄	⚙	🧭	A	↕	📍	🔍	📷	🎨	Model Average
OpenAI: GPT-4o (OpenAI, 2024a)	–	77.78	83.33	83.33	85.71	75.00	78.13	91.43	94.44	96.43	**85.06**
Google: Gemini 2.5 Pro (Gemini Team, 2025)	–	79.63	76.67	86.67	92.86	79.17	78.13	88.57	88.89	89.29	**84.43**
Qwen: Qwen2.5 VL 72B Instruct (Qwen Team, 2025)	72	74.07	83.33	73.33	85.71	79.17	71.88	90.00	72.22	82.14	**79.09**
Mistral: Pixtral Large 2411 (Mistral AI, 2024)	124	83.33	86.67	66.67	71.43	79.17	71.88	88.57	72.22	85.71	**78.41**
xAI: Grok 4 (xAI, 2025)	–	62.96	73.33	40.00	64.29	50.00	50.00	82.86	55.56	67.86	**60.76**
Google: Gemma 3 27B (Gemma Team, 2025)	27	68.52	73.33	66.67	78.57	70.83	68.75	90.00	72.22	89.29	**75.35**
Meta: Llama 3.2 90B Vision Instruct (Meta, 2024)	90	87.04	96.67	90.00	85.71	83.33	90.63	97.14	100.00	96.43	**91.88**
Meta: Llama 4 Maverick (Meta, 2025)	$17Bx128E$	88.89	93.33	86.67	92.86	95.83	71.88	90.00	77.78	89.29	**87.39**
Meta: Llama 4 Scout (Meta, 2025)	$17Bx16E$	81.48	93.33	70.00	85.71	79.17	75.00	95.71	72.22	92.86	**82.83**
Anthropic: Claude Opus 4.1 (Anthropic, 2025a)	–	83.33	76.67	83.33	85.71	75.00	81.25	87.14	94.44	85.71	**83.62**
Anthropic: Claude Sonnet 4 (Anthropic, 2025b)	–	77.78	80.00	60.00	78.57	70.83	75.00	87.14	55.56	89.29	**74.91**
Category Average		**78.62**	**83.33**	**73.97**	**82.40**	**75.23**	**74.78**	**89.87**	**77.78**	**87.66**	**80.34**

Table 1: Individual VLM Accuracies: Based on images and associated questions from the MMVP dataset. Failure modes are highlighted across all models based on definitions (see 3.1). Highest non-failure mode accuracies in each category are spread across the models. We use symbols as a representation for all nine categories: 🔄: State and Condition, ⚙: Structural and Physical Characteristics, 🧭: Orientation and Direction, **A**: Text, ↕: Quantity and Count, 📍: Positional and Relational Context, 🔍: Presence of Specific Features, 📷: Viewpoint and Perspective, 🎨: Color and Appearance. Based on the accuracies, it's evident that Quantity and Count, as well as Positional and Relational Context, are the two categories the VLMs struggled the most with.

Model	Params Size (B)	🔄	⚙	🧭	A	↕	📍	🔍	📷	🎨	Model Average
OpenAI: GPT-4o (OpenAI, 2024a)	–	62.96	66.67	66.67	71.43	50.00	56.25	82.86	88.89	92.86	**70.95**
Google: Gemini 2.5 Pro (Gemini Team, 2025)	–	66.67	60.00	73.33	85.71	58.33	56.25	77.14	77.78	78.57	**70.42**
Qwen: Qwen2.5 VL 72B Instruct (Qwen Team, 2025)	72	59.26	73.33	53.33	71.43	58.33	50.00	80.00	44.44	64.29	**61.60**
Mistral: Pixtral Large 2411 (Mistral AI, 2024)	124	66.67	73.33	40.00	42.86	58.33	43.75	77.14	44.44	71.43	**57.55**
xAI: Grok 4 (xAI, 2025)	–	37.04	53.33	33.33	42.86	25.00	25.00	71.43	33.33	35.71	**39.67**
Google: Gemma 3 27B (Gemma Team, 2025)	27	44.44	46.67	33.33	71.43	41.67	43.75	80.00	44.44	78.57	**53.81**
Meta: Llama 3.2 90B Vision Instruct (Meta, 2024)	90	77.78	93.33	80.00	71.43	66.67	81.25	94.29	100.00	92.86	**84.18**
Meta: Llama 4 Maverick (Meta, 2025)	$17Bx128E$	77.78	86.67	73.33	85.71	91.67	56.25	80.00	66.67	78.57	**77.41**
Meta: Llama 4 Scout (Meta, 2025)	$17Bx16E$	66.67	86.67	53.33	71.43	66.67	50.00	91.43	44.44	85.71	**68.48**
Anthropic: Claude Opus 4.1 (Anthropic, 2025a)	–	70.37	60.00	66.67	71.43	58.33	62.50	74.29	88.89	71.43	**69.32**
Anthropic: Claude Sonnet 4 (Anthropic, 2025b)	–	62.96	60.00	33.33	57.14	41.67	50.00	74.29	11.11	78.57	**52.12**
Category Average		**62.96**	**69.09**	**56.67**	**67.01**	**56.06**	**52.27**	**80.29**	**58.00**	**75.27**	**64.18**

Table 2: Pair VLM accuracies: Based on and associated questions from the MMVP dataset. Failure modes are highlighted across all models based on definitions (see 3.1). Highest non-failure mode accuracies in each category are spread across the models.

Model	Params Size (B)	🔄	⚙	🧭	A	↕	📍	🔍	📷	🎨	Model Average
OpenAI: DALL·E 3 (OpenAI, 2024b)	–	77.78	90.00	66.67	71.43	66.67	75.00	75.71	83.33	89.29	**77.32**
Google: Gemini 2.5 Flash Image (Fortin et al., 2025)	–	94.44	96.67	96.67	78.57	75.00	90.63	85.71	100.00	96.43	**90.46**
Stability AI: Stable Diffusion 3.5 Large (Stability AI, 2024)	8.1	55.56	73.33	56.67	42.86	50.00	43.75	67.14	77.78	78.57	**60.63**
Category Average		**75.93**	**86.67**	**73.34**	**64.29**	**63.89**	**69.79**	**76.19**	**87.04**	**88.10**	**76.14**

Table 3: Individual Explicit Accuracy for Image Generation Models: Based on AMVICC (see 3.2)

Model	Params Size (B)	🔄	⚙	🧭	A	↕	📍	🔍	📷	🎨	Model Average
OpenAI: DALL·E 3 (OpenAI, 2024b)	–	55.56	80.00	40.00	42.86	50.00	56.25	57.14	66.67	85.71	**59.35**
Google: Gemini 2.5 Flash Image (Fortin et al., 2025)	–	88.89	93.33	93.33	57.14	66.67	81.25	74.29	100.00	92.86	**83.08**
Stability AI: Stable Diffusion 3.5 Large (Stability AI, 2024)	8.1	25.93	46.67	20.00	14.29	25.00	12.50	40.00	66.67	64.29	**35.04**
Category Average		**56.79**	**73.33**	**51.11**	**38.10**	**47.22**	**50.00**	**57.14**	**77.78**	**80.95**	**59.16**

Table 4: Pair Explicit Accuracy for Image Generation Models: Based on AMVICC (see 3.2)

and direction (od), positional and relational context (pr), and presence of specific features (pf).

Of the three models evaluated, Google Flash achieved the highest performance with only two failure modes across pair and individual explicit accuracy in qc and tx. Inversely, Stable Diffusion 3.5 performed worse than the other two models with all categories dropping below the standard for failure modes in pair explicit accuracy. Despite DALL·E 3 achieving moderate performance, its failure across 7 categories indicated IGM's shortcomings in generating images.

3.3 Cross-Examination of IGMs and MLLMs

Collectively, certain categories such as quantity and count constituted failures in both IGMs and MLLMs, with both modalities performing notably poorly on them. Other common failure groupings included viewpoints and perspectives and state and condition. However, while both MLLMs and IGMs generally tend to perform worse across all categories, MLLMs performed significantly better in textual contexts (tx) and IGMs performed significantly better for positional and relational context despite Stable Diffusion's low accuracy. Pair accuracy for both MLLMs and IGMs was also lower than all individual accuracies due to the requirement for both images in a pair or both answers in a pair to be correct in order to be considered a correct pair. Image generation models depicted a larger disparity in the capabilities of each model with Stable Diffusion unable to follow elementary instructions in differentiating between the implicit and explicit prompts and Gemini 2.5 Flash Image consistently accurately adding components based on the explicit instructions. These results underscore the need for more intensive testing into failure modes of MLLMs and IGMs in order to cross-reference influencing factors and improve visual intelligence and understanding across the field of machine learning.

3.4 Ablation Studies

To further explore the robustness and reliability of model behavior, we conducted a series of ablation studies designed to test sensitivity to prompt phrasing, model randomness, and architectural differences. These studies aimed to isolate which factors most influenced success or failure across tasks.

3.4.1 Linguistic Sensitivity

In order to understand whether prompt wording and adaptation to questions directly affected the outcome and accuracies demonstrated from image generation models, we changed the wording of 20 prompts and tested them on OpenAI's DALL·E 3 to determine whether the accuracies would fall in the same range as the original tests. We utilized ChatGPT-5 to improve prompt wording by adding context clues and disregarded the original prompt constraint of explicit prompts only having the new specific component in addition to the pair-generic implicit prompt. We used a randomly generated interval of the prompts in order to ensure generalization of the sample to the population. However, based on the overall accuracy of the prompts, it is clear that adding more targeted language does not help improve model accuracy except for a small decrease in pair accuracy for the category PF (Presence of Specific Features) in explicit prompts.

Pair Implicit Types	🎨	🔍	📍	🔄	⚙	**A**
Pair Implicit (C)	100.00	100.00	100.00	100.00	100.00	100.00
Pair Implicit (W)	100.00	100.00	100.00	100.00	100.00	100.00
Pair Explicit (C)	100.00	100.00	100.00	100.00	100.00	0.00
Pair Explicit (W)	100.00	75.00	100.00	100.00	100.00	0.00

Table 5: Linguistic Sensitivity Trials: Pair Implicit and Explicit Accuracies for Reworded Prompts. (C) denotes control/original wording; (W) denotes reworded prompts.

3.4.2 IGM Stochasticity

To evaluate the significance of model stochasticity in IGMs, we tested the 10 prompt-pairs through 3 trials, generating 60 total implicit images and 60 total explicit images for 20 prompts. We utilized DALL-E 3 (the median performance model between Gemini 2.5 Flash Image and Stable Diffusion 3.5 Large) and ran an identical experiment pipeline to the main experiment. Through the findings, we concluded that while prompts could individually vary with accuracy with certain prompts only scoring accurately on two of the three tests, individual variance did not drastically affect the overall accuracy of the test set in the sample. This highlighted a negligible role of sampling variance in IGM failure modes and suggested that conceptual misunderstanding rather than model stochasticity model, accounted for the principal model accuracy.

Tests	Test 1	Test 2	Test 3
Individual Implicit	100.00	100.00	100.00
Individual Explicit	90.00	85.00	90.00
Pair Implicit	100.00	100.00	100.00
Pair Explicit	80.00	70.00	80.00

Table 6: IGM Stochasticity Trials: Individual and Pair Implicit and Explicit Accuracies for Three Separate Trials

4 Discussion

Our findings indicated that IGMs generally exhibited equal or higher levels of failure compared to MLLMs. However, category-specific analysis revealed that performance varied between the two, with each model type performing better in different category-specific tasks. Outliers on both ends of the spectrum included Meta: Llama 3.2 90B Vision Instruct and Google: Gemini 2.5 Flash Image, which achieved the best results, and xAI: Grok 4 and Stability AI: Stable Diffusion 3.5 Large, which showed the worst performance of their modalities. Each model exhibited fluctuations in performance compared to other models, alternating between producing stronger and weaker results. For instance, models of both modalities failed in Quantity and Count, but IGMs outperformed MLLMs in Positional and Relational Context while MLLMs outperformed IGMs in Text. However, if all these models are trained on the same data structure and similar data (ex. image-caption pairs in DALL-E 3), this could indicate that size is not relevant to the elementary visual understandings of either VLMs or IGMs (OpenAI, 2024b).

Furthermore, image generation models struggle significantly with text, camera angling to remove specific features, and quantity and count. Google Flash by far outperformed Stable Diffusion and DALL·E 3 in image realism and consistency, and a notable disparity among image generation models in quality emerged through our tests.

However, as observed in human evaluation, all image generation models were often unable to leave out specific features in each category and were unable to manipulate viewpoints to hide specific components as prompted, especially when "no" or "without" was included. This suggests that image generation models, despite the quality of the images, still struggle with elementary instruction following for certain phrasing. Sometimes, components instructed to be partially hidden were fully shown, and components instructed to be fully hidden were still slightly seen, indicating that some generated images just barely failed to meet the entirety of their prompts' requirements; this reduced the overall accuracies of the image generation models. Even though Gemini 2.5 Flash Image's capability far outperformed the other two IGMs, it still struggled with these same underlying issues that diminished its accuracy. For instance, when all 3 models were asked to generate a keyboard for one of the prompts, the generated keyboard quality was drastically better and more realistic for Gemini 2.5 Flash Image compared to the other two models. However, all three models failed to follow the implied instructions when prompted to create an image in contexts of higher difficulty, where it wasn't systematically stated how to achieve the image. This limited their ability to accomplish the prompt's direct requirement of having or not having a specific element in the generated image[1]. For example, one of the explicit prompts instructed the IGM to generate a computer keyboard with the Z key hidden. In the prompt, it was not expressly stated that the IGM would have to orient the image angle in a way where the Z key is hidden; the model would have to understand the implied instruction in order to satisfy the prompt.

Some models also indicated struggles with understanding contextual cues and alignment with natural human thought. For instance, if asked to produce a stripe down the middle of a car, Stable Diffusion would produce a stripe across the horizontal middle of the car while DALL-E 3 and Gemini 2.5 Flash produced a clear stripe across the middle of the top of the car as many humans would naturally think.

Interestingly, the architectures of the best and worst performing models of different modalities offered key insights and introduced new questions about the relevance of various architectures in model performance for elementary visual understanding and depiction. For example, Llama 3.2 90B Vision-Instruct, a two-stage vision encoder added on to a frozen LLM, which often doesn't outperform the more popular models such as GPT-4o on complex tasks, easily outperformed GPT-4o across all but two categories, Text and Color and Appearance. Gemini 2.5 Flash Image, a sparse mixture-of-experts (MoE) transformer, outperformed DALL-E 3 even though they were both

[1] An explicit prompt specifies the content that must be included in the image, whereas explicit instruction specifies how that content should be achieved or generated.

trained with a natively multi-modal architecture and similarly structured pairs of image and text data.

As a result, this could create systems-level deployment challenges due to a lack of accuracy in elementary reasoning, which could lead to long-term oversights in basic tasks, essentially risking efficiency and scalability. It is necessary to perform more in-depth testing to uncover the basis for why image generation models and multimodal LLMs seem to fail and succeed in differing categories. We hope our work provides the foundational data to understand where current models succeed and where they fail.

5 Related Works

5.1 Failure Modes in Image Generation Models

Text-to-image generation models such as DALL·E 3 and Stable Diffusion have made rapid progress in image quality, but continue to face challenges in commonsense reasoning, fairness, and scene composition. Recent evaluations have shown systematic biases and reasoning failures in these models, raising questions about their true semantic understanding. Commonsense-T2I Challenge showed major failures in reasoning; DALL·E 3 scored approximately only 48% accuracy (Fu et al., n.d.). A biased survey identified a lack of evaluation frameworks and coverage of non-binary identities (Subramonian et al., n.d.). Similarly, a diffusion model survey highlighted specific weaknesses like generating multiple objects and rare concepts; proposed layout and attention improvements sought to improve the model (Zhang and Wang, n.d.). Although this work identifies critical weaknesses in generative performance, it remains unclear whether these are shared with interpretive failures in vision-language models or whether they have been directly compared to correlating tasks within varied-architecture MLLMs.

5.2 Visual Reasoning Challenges in Visual Language Models

Visual Language Models (VLMs) like GPT-4o and Gemini 2.5 Pro have become central to visual reasoning tasks, yet they often falter on simple image-based questions. Efforts to improve VLMs have been centered around better pretraining, alignment, and hallucination reduction using methods like VILA, CogVLM2, and SIMA. VILA showed improved in-context learning and world knowledge from interleaved pretraining (Lin et al., n.d.). SIMA reduced hallucinations and boosted VQA benchmark accuracy via visual critic metrics (Wang et al., n.d.). CogVLM2 achieved SoTA across multiple visual benchmarks with efficient architecture (Hong et al., n.d.). Despite these advances, prior work focuses solely on improving VLMs without evaluating whether these errors also emerge during generative tasks. Current existing studies don't test model performance on aligned image/question pairs.

6 Conclusion

In this work, we introduced a novel benchmark, AMVICC or Assessment of Modality-specific Visual Intelligence, Comprehension, and Creation, to evaluate the cross-modal failure modes of multimodal large language models and image generation models in order to gain insight into the commonalities and distinctions. We concluded that not only do IGMs and MLLMs share certain common failure modes and differ on others, they also diverge within specific modalities to create model-specific failure modes that could be attributed to a wide range of factors. Future work can expand the MMVP or AMVICC benchmarks to increase the range of visual understanding categories evaluated or improve visual understanding on specific models to improve accuracy for specific categories. Further extensions of this paper can replicate tests to prove accuracy on a larger scale with more resources.

7 Limitations

7.1 Methodology Limitations

The primary limitation present within this methodology is the conversion from the MMVP to specific prompts that cover the same visual element as the questions. As the prompts are written by 2 separate members of our research team, albeit following a strict linguistic structure, there is inherent prompt design bias. This hinders our ability to definitively state that the translation of categories and tasks tested can be completely translated to image generation models. However, the structure that we utilized in order to define the creation of the prompts, as outlined in Section 3, ensures that each prompt follows the same structure and inherits the same information from each question to ensure rigorous alignment.

Furthermore, each prompt and each question could fall under multiple categories. However, to

allow for predominantly accurate findings, we assigned each prompt and question to only one category. However, while their success and failures could also influence the accuracy of other categories that they could fall under, it is not incorporated into the final numbers. However, each prompt is double-checked by multiple human prompt writers to optimize categorization in order to mitigate this issue.

Another limitation includes the unbalanced model usage of IGMs compared to MLLMs. Due to the lack of availability of image generation models through API keys and time constraints, we were unable to test as many IGMs as MLLMs. This imbalance means that our accuracy averages for our IGM could potentially be less representative of the overall failure modes of all IGMs compared to the representation offered by the MLLM accuracy averages.

7.2 Evaluation Limitations

Due to our MLLMs having been proven to have visual reasoning deficiencies, we chose to use human evaluators to determine the accuracy of the outputs produced by image generation. Despite the rubric outlined in Section 3's specificity to reduce subjectivity of human evaluators, there is still a chance of human subjectivity bias in the results. However, the specificity of the rubric limited the ability of the empirical data to represent the confounding factors of the data, such as the situational factors generated around the specific criteria (ex. Z key in a keyboard compared to an inaccurate depiction of a keyboard is still incorrect). These, due to computational power and human resources, limit the extent to which the failure modes can be understood from the data.

Furthermore, an OpenAI AI grader was utilized for MLLMs. This could skew the results due to a lack of a human counterpart in evaluations, but because there are answer choices present, an AI grader is only consolidating any potential responses from an MLLM into either 'a' or 'b' as an answer.

Another limitation encompasses the lack of a human performance control group for image generation performance due to the technological nature of the task that we are testing on IGMs. This requires us to understand the competencies and capabilities of models through relational comparison between models.

Another limitation arises from the closed-source nature of many of the models, because we are unable to look at the internal elements of the model and must only rely on surface-level documentation provided by commercial companies (ex. DALL-E 3).

References

Casper, S., Schulze, L., Patel, O., & Hadfield-Menell, D. (2024, March 8). Defending against unforeseen failure modes with latent adversarial training (preprint). arXiv. https://arxiv.org/abs/2403.05030

Fu, X., He, M., Lu, Y., Wang, W. Y., & Roth, D. (n.d.). Commonsense-T2I Challenge: Can text-to-image generation models understand commonsense? Retrieved from https://www.semanticscholar.org/paper/Commonsense-T2I-Challenge%3A-Can-Text-to-Image-Models-Fu-He/64a9a997d796678edc9d5693424d9feb2e9d3777

Krishnapriyan, A., Gholami, A., & others. (n.d.). Characterizing possible failure modes in neural models. Retrieved from https://www.semanticscholar.org/paper/Characterizing-possible-failure-modes-in-neural-Krishnapriyan-Gholami/3c4372b125d0744bb68bfca9f5d6b0abb85dd182

Lin, & Yin, . . . (n.d.). VILA: On pre-training for visual language models. Retrieved from https://www.semanticscholar.org/paper/VILA%3A-On-Pre-training-for-Visual-Language-Models-Lin-Yin/2141ed804636a1cf339d606cd03fd3b3e9582133

Pal, A., Karkhanis, D., Dooley, S., Roberts, M., Naidu, S., & White, C. (2024, February 20). Smaug: Fixing failure modes of preference optimisation with DPO-Positive (preprint). arXiv. https://arxiv.org/abs/2402.13228

Subramonian, V., & Wan, . . . (n.d.). Survey of bias in text-to-image generation. Retrieved from https://www.semanticscholar.org/paper/Survey-of-Bias-In-Text-to-Image-Generation%3A-and-Wan-Subramonian/8c323eca1406bd4020c98d6b5f00ff8f2b7f3340

Wang, & Chen, . . . (n.d.). Enhancing visual-language modality alignment in . . . . Retrieved from https://www.semanticscholar.org/paper/Enhancing-Visual-Language-Modality-Alignment-in-via-Wang-Chen/4499afc74bda1c7d521a516df040facfe39943ed

Zhang, Y., & Wang, . . . (n.d.). A survey of diffusion-based image generation issues. Retrieved from https://www.semanticscholar.org/paper/A-Survey-of-Diffusion-Based-Image-Generation-Issues-Zhang-Wang/49faa5c9bf6459a256f68872fb3b51df6b0a2dd8

Zhang,& Ochiai, . . . (n.d.). A design of interface for visually impaired people to . . . . Retrieved from

https://www.semanticscholar.org/paper/A-Design-of-Interface-for-Visual-Impaired-People-to-Zhang-Ochiai/d01bf3d6f885cacfad27724e4d09decf60ff6578

Hong, & Wang, . . . (n.d.). CogVLM2: Visual language models for image and video. Retrieved from https://www.semanticscholar.org/paper/CogVLM2%3A-Visual-Language-Models-for-Image-and-Video-Hong-Wang/3c83033c15e889302d0d21597e518a2f5c723291

Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. (2024). Meta.com. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/

The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. (2025). Meta.com. https://ai.meta.com/blog/llama-4-multimodal-intelligence/

Grok 4 | xAI. (2025). X.ai. https://x.ai/news/grok-4

Team, G., & Deepmind, G. (2025). Gemma 3 Technical Report. Retrieved September 27, 2025, from https://arxiv.org/pdf/2503.19786

Fortin, A., Guillaume Vernade, Kampf, K., & Ammaar Reshi. (2025, August 26). Introducing Gemini 2.5 Flash Image, our state-of-the-art image model. Googleblog.com. https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/

OpenAI. (2024). GPT-4o System Card. https://arxiv.org/pdf/2410.21276

Team, Q. (2025). Qwen2.5-VL Technical Report. https://arxiv.org/pdf/2502.13923

Pixtral Large | Mistral AI. (2024). Mistral.ai. https://mistral.ai/news/pixtral-large

Claude Opus 4.1. (2025). Anthropic.com. https://www.anthropic.com/news/claude-opus-4-1

System Card: Claude Opus 4 & Claude Sonnet 4. (2025). https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf

Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., & Xie, S. (2024). Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 9568-9578.

DALL·E 3 system card. (2024, February 14). Openai.com. https://openai.com/index/dall-e-3-system-card/

AI. (2024, October 22). Stability AI. Stability AI. https://stability.ai/news/introducing-stable-diffusion-3-5

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2024). A survey on multimodal large language models. In IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE [Journal-article].

Anis, A. M., Ali, H., Sarfraz, M. S., Cohere for AI Community, Arbisoft, & Karlsruhe Institute of Technology. (2025). On the Limitations of Vision-Language Models in Understanding Image Transforms. arXiv.

Chen, P., 1, Ye, J., 1, Wang, G., 1, Li, Y., 1, Shanghai AI Laboratory, University of Washington, Monash University, East China Normal University, University of Cambridge, Shanghai Jiao Tong University, The Chinese University of Hong Kong, Shenzhen, Shenzhen Research Institute of Big Data, Shenzhen Institute of Advanced Technology (SIAT), Chinese Academy of Sciences, Seibel, E. J., 2, He, J., 1, Qiao, Y., 1, Shanghai AI Laboratory, University of Washington, Monash University, . . . Qiao, Y., 1. (2024). GMAI-MMBench: A comprehensive multimodal evaluation benchmark towards general medical AI [Preprint]. arXiv. https://arxiv.org/abs/2408.03361v7 (Original work published 2408)

Sepehri, M. S., Fabian, Z., Soltanolkotabi, M., & Soltanolkotabi, M. (2024, September 23). MediConfusion: Can you trust your AI radiologist? Probing the reliability of multimodal medical foundation models [Preprint]. arXiv. https://arxiv.org/abs/2409.15477

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Meta AI, Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLAMA: Open and Efficient Foundation Language Models. arXiv. https://arxiv.org/pdf/2302.13971.pdf

OpenAI. (2024). GPT-4 Technical Report. In arXiv: Vol. 2303.08774v6 [Technical report]. https://arxiv.org/pdf/2303.08774.pdf (Original work published 2023)

Marioriyad, A., Department of Computer Engineering, Sharif University of Technology, Rezaei, P., Department of Computer Engineering, Sharif University of Technology, Soleymani Baghshah, M., Department of Computer Engineering, Sharif University of Technology, Rohban, M. H., & Department of Computer Engineering, Sharif University of Technology. (2025). Diffusion Beats Autoregressive: An Evaluation of Compositional Generation in Text-to-Image Models. arXiv.

Gokhale, T., Palangi, H., Nushi, B., Vineet, V., Microsoft Research, Horvitz, E., Kamar, E., Baral, C., & Yang, Y. (2023). Benchmarking Spatial Reasoning Abilities of Text-to-Image Generative Models. arXiv. https://arxiv.org/pdf/2212.10015

Xu, W., 1, Wang, J., 2, Wang, W., 3, Chen, Z., 3, Zhou, W., Yang, A., Lu, L., Li, H., Wang, X., Zhu, X., Wang, W., Dai, J., 5, Zhu, J., University of Science and Technology of China, Xi'an Jiaotong University,

Shanghai Artificial Intelligence Laboratory, SenseTime Research, & Tsinghua University. (n.d.). VisuLogic: A Benchmark for Evaluating Visual Reasoning in Multi-modal Large Language Models. axRiv. https://visulogic-benchmark.github.io/VisuLogic

Huang, K., Duan, C., Sun, K., Xie, E., Li, Z., Liu, X., The University of Hong Kong, Tsinghua University, Huawei Noah's Ark Lab, & IEEE Publication Technology Department. (2021). T2I-CompBench++: an enhanced and comprehensive benchmark for compositional text-to-image generation. In JOURNAL OF LATEX CLASS FILES (Vol. 14, Issue 8, pp. 1–2) [Journal-article]. https://karine-h.github.io/T2I-CompBench-new/

Team, G., & Google. (2025). Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf

A Appendix

A.1 Prompt Sets for VLM and Image Generation

Below is the link to the prompts used to evaluate Visual Language Models (VLMs) and to guide the image generation process. These were adapted directly from the MMVP benchmark to ensure consistency across tasks: AMVICC-Benchmark Prompts

A.1.1 Categories (Defined):

1. Orientation and Direction: The model's ability to accurately detect the position, alignment, facing direction, or angles of objects in the image

2. Presence of Specific Features: The ability of a model to identify if specific visual characteristics, objects, or fine-grained attributes are explicitly present in an image.

3. State and Condition: This refers to the model's ability to be able to recognize the current status, phase, or physical condition of an object, entity, or scene that is being depicted in an image.

4. Quantity and Count: The model's ability to identify the number of objects, people, or elements in an image, including the tasks that involve counting, estimating quantities, or comparing amounts.

5. Positional and Relational Context: It refers to a model's ability to be able to understand the spatial relationships and relative positions between objects or entities within an image.

6. Color and Appearance: This refers to the ability of the model to perceive, recognize, and reason about colors, visual patterns, and image-level characteristics like tone, brightness, and artistic style.

7. Structural and Physical Characteristics: The model's ability to perceive and reason about the shape, material, construction, and physical properties of objects or elements within an image.

8. Text: The ability of a model to detect, recognize, and interpret written language (printed, handwritten, or stylized text) that appears within an image, and to reason about its content, meaning, and context.

9. Viewpoint and Perspective: It refers to the ability of a model to be able to recognize and reason about the camera or observer's perspective and angle relative to the objects or scene in an image, affecting how elements are visually presented.

Each task (image interpretation or image generation) was analyzed independently and comparatively across these dimensions to identify common and divergent failure modes.

A.2 Code Base

All code used in this study for model evaluation, result collection, and visualization is available at: `https://anonymous.4open.science/r/AMVICC-IGM-FC2A/README.md`

A.3 Experiments (Further Outlined)

This section outlines how we applied our methods to test the failure mode alignment between vision-language models (VLMs) and image generation models (IGMs), specifying the experimental conditions, controls, and design decisions underpinning our analysis.

Overview and Hypotheses: We test the core hypothesis: Do the failure modes of VLMs in visual reasoning correlate with the failure modes of IGMs when tasked with generating images that express those same visual concepts?

This hypothesis rests on two premises: If VLMs fail to understand a visual concept (e.g., object orientation), IGMs may also fail to generate that

concept reliably. Alternatively, divergence in failure patterns would suggest modality-specific weaknesses, pointing to differences in model architecture or training objectives.

Experimental Variations and Comparative Design: To probe our hypothesis and ensure robustness, we introduced several comparative and diagnostic experiments: Cross-Modality Comparison: VLM Task: Answer MMVP questions based on real and generated images. IGM Task: Generate images based on prompts derived from MMVP questions. Explicit vs. Implicit Prompting: We varied prompt specificity to test if IGMs struggled more with indirect language. This also enabled assessment of whether image failures propagated into VLM misinterpretation when fed generated content. *Ablation: Prompt Rewording:* For failure-prone prompts, we created reworded versions to test whether small linguistic changes improved generation accuracy or altered failure types. *Ablation: Repetition Analysis (Randomness Test):* For 30 selected prompts, DALL·E 3 was queried 5 times each. We analyzed generation consistency and its impact on downstream VLM accuracy. *Ablation: Architecture/Scale Variation:* We included related models with different parameter sizes (e.g., Llama 4 Maverick vs. Scout) to evaluate the impact of architecture vs. scale. All these comparisons allowed us to isolate not only when models failed, but why—whether due to conceptual, linguistic, architectural, or visual representation limitations. *Data Summary:* Below is a description of the experiment flow: VLM Baseline: Each of the 300 MMVP (image, question) pairs evaluated across 11 VLMs. IGM Prompting: 600 text prompts (explicit + implicit) derived from MMVP questions input to 3 IGMs. *Generated Image Evaluation:* Human annotators judged whether the images accurately captured the core visual concepts. Generated images were fed back into VLMs to answer the original questions. Cross-analysis: Accuracy and failure mode types were tracked and compared across tasks, models, categories, and prompt types.

A.4 VLM Image Analysis Responses

When prompted to analyze the images generated in Section 3, the VLMs produced outputs containing their answers and explanations. These responses were used to calculate individual and pair accuracy for VLM evaluation. One of the responses includes:


```
{
  "question_id": 26,
  "category": "ca",
  "question": "What color is the chicken's body?
               (a) Black (b) Red",
  "correct_answer": "(b)",
  "model_response": "The chicken's body is a
                     reddish-brown color, which
                     is typical for certain
                     breeds like Rhode Island
                     Reds. So, the correct
                     answer is: (b) Red",
  "is_correct": true,
  "gpt_grade": "yes"
}
```


A.5 Rubric for Image Generation & Vision Language Model Evaluation

Below is the rubric used to assess whether a generated image successfully followed a prompt. Access on GitHub: https://anonymous.4open.science/r/AMVICC-IGM-FC2A/README.md

	0	*1*
Implicit	Does not generate the scenario with each specific aspect mentioned in the prompt	Generates the scenario with each specific aspect mentioned in the prompt
Explicit	Does not generate the specific feature that the prompt asks for based on the visual understanding question	Generates the specific feature that the prompt asks for based on the visual understanding question

Table 7: Rubric for Image Generation Evaluation

	0	*1*
Question	Answers question incorrectly based on GPT_Grader	Answers question correctly based on GPT_Grader similarity index

Table 8: Rubric for Visual Language Model Evaluation

A.6 Generated Image Results & Extra Results

Figure 4 below outlines 5 images generated from 5 of the prompts[2]. These outputs were used as part of the image analysis phase to assess whether image generation models could accurately depict the components (see figure on the next page). Here is a link to the GitHub Repository labeling the implicit accuracies in the results section of the image generation models referenced in section 3 as produced by the image generation evaluation code: https://anonymous.4open.science/r/AMVICC-IGM-FC2A/README.md

More results available on Zenodo (redacted to protect anonymity and will be added back after review)

[2] All of the prompts are in scripts/AMVICC.csv, and the file is in the AMVICC-IGM GitHub Repository.

Can you see the key "Z" in the image?

(a) Yes
(b) No

Prompt: Generate a computer keyboard photographed at an angle where the Z key is hidden

Is the shark's belly visible in this image?

(a) Yes
(b) No

Prompt: Generate a shark photographed from below showing its belly

Does the elephant have long or short tusks?

(a) Long
(b) Short

Prompt: Generate an elephant with prominently long curved tusks

Is there a lemon inside the drink in the cup, or are all the lemons outside the drink?

(a) There is one inside
(b) All are outside

Prompt: Generate water in a glass with mint leaves and lemon slices, with all the lemons outside of the glass

Are there any words displayed on the vehicle's lightbar?

(a) Yes
(b) No

Prompt: Generate a police car with "POLICE" text clearly visible on the lightbar

	Question 1		Question 2		Question 3		Question 4		Question 5	
Model	**Ans.**	**Result**	**Ans.**	**Result**	**Ans.**	**Result**	**Ans.**	**Result**	**Ans.**	**Result**
OpenAI: DALL·E 3	a	×	b	×	a	✓	b	✓	b	×
Google: Gemini 2.5 Flash Image	a	×	a	✓	a	✓	b	✓	a	✓
Stability AI: Stable Diffusion 3.5 Large	b	✓	a	✓	b	×	a	×	b	×

Figure 4: Examples of specific IGMs' abilities to generate an image based on explicit prompts. We handpick 5 out of the 300 questions in the MMVP dataset to delineate disparities between the models. It is apparent that Google: Gemini 2.5 Flash Image was the most accurate, followed by OpenAI: DALL·E 3, and Stability AI: Stable Diffusion 3.5 Large, in that order. An important thing to note is that the IGMs don't directly state Yes or No or any of the answer choices, for that matter. However, based on the models' image generation, we can associate certain answer choices with the models. A ✓ indicates that the model generated an image in accordance with the given prompt, whereas an × indicates the opposite.