



# PIXELS, PATTERNS, BUT NO POETRY: TO SEE THE WORLD LIKE HUMANS

## *A Preliminary*

Hongcheng Gao<sup>\*1</sup>, Zihao Huang<sup>\*1</sup>, Lin Xu<sup>\*1</sup>, Jingyi Tang<sup>\*1</sup>, Xinhao Li<sup>2</sup>  
 Yue Liu<sup>3</sup>, Haoyang Li<sup>4</sup>, Taihang Hu<sup>5</sup>, Minhua Lin<sup>6</sup>, Xinlong Yang<sup>7</sup>  
 Ge Wu<sup>5</sup>, Balong Bi<sup>1</sup>, Hongyu Chen<sup>8</sup>, Wentao Zhang<sup>†7</sup>

<sup>1</sup>University of Chinese Academy of Sciences    <sup>2</sup>Nanjing University    <sup>3</sup> National University of Singapore  
<sup>4</sup> BUPT    <sup>5</sup> Nankai University    <sup>6</sup>The Pennsylvania State University    <sup>7</sup>Peking University    <sup>8</sup>BJTU

Project Page: <https://TuringEyeTest.github.io>

## ABSTRACT

Achieving human-like perception and reasoning in Multimodal Large Language Models (MLLMs) remains a central challenge in artificial intelligence. While recent research has primarily focused on enhancing reasoning capabilities in MLLMs, a fundamental question persists: *Can Multimodal Large Language Models truly perceive the world as humans do?* This paper shifts focus from reasoning to perception. Rather than constructing benchmarks specifically for reasoning, we introduce the *Turing Eye Test (TET)*, a challenging perception-oriented benchmark comprising four diagnostic tasks that evaluate MLLMs' performance on synthetic images that humans process intuitively. Our findings reveal that state-of-the-art MLLMs exhibit catastrophic failures on our perceptual tasks trivial for humans. Both in-context learning and training on language backbone—effective for previous benchmarks—fail to improve performance on our tasks, while fine-tuning the vision tower enables rapid adaptation, suggesting that our benchmark poses challenges for vision tower generalization rather than for the knowledge and reasoning capabilities of the language backbone—a key gap between current MLLMs and human perception. *This is a preliminary version that only contains a subset of TET tasks. We will release the full set of TET with more diverse tasks and explore methods to improve visual generalization in the next version.*

## 1 INTRODUCTION

Large Language Models (LLMs) have demonstrated powerful capabilities across various tasks (Radford et al., 2018; 2019; Brown et al., 2020; Achiam et al., 2023). This breakthrough has catalyzed the development of multimodal architectures that extend beyond text to encompass visual understanding. Compared to LLMs, the main characteristic of Multimodal Large Language Models (MLLMs) is their ability to directly recognize and understand images (Liu et al., 2023; 2024a;b). The most popular method is to integrate a vision encoder with a language model (Li et al., 2023; Chen et al., 2024b;a); this process works by first projecting the image's features into the language model's embedding space, allowing the model to process both visual and text inputs seamlessly to generate a coherent response. Unified models (Team, 2024a; Zhou et al., 2024; Wang et al., 2024b; Luo et al., 2025), another paradigm, take this integration further by using a single architecture that processes both visual and textual tokens natively within the same parameter space, eliminating the need for separate vision and language encoders.

Recent MLLMs have achieved remarkable performance across numerous visual benchmarks like MMMU (Yue et al., 2024a), MathVista (Lu et al., 2024), and MathVision (Wang et al., 2024a). However, these benchmarks primarily evaluate the *knowledge and reasoning capabilities of the language backbone* rather than fundamental visual perception abilities. To explore the edge of MLLM

\*Equal contribution. † Corresponding author.



Figure 1: Evaluation cases for each category, including the four tasks: HiddenText, 3DCaptcha, ColorBlind, and ChineseLigatures. The text beneath each image in every subset represents the corresponding ground truth. The third line of **chinese characters**, read from left to right, symbolizes *marital bliss, a serendipitous union, a perfect match, dreams fulfilled, and flawless*.

in visual perception, we create a benchmark named ***Turing Eye Test (TET)*** with specialized tasks featuring unique visual challenges: concealed text within scenic imagery, which tests holistic pattern recognition in composite images; 3D Captchas, which evaluates 3D spatial character recognition capabilities; Chinese character compositions, which challenges the decomposition and recognition of complex character structures; and color blind test charts, which assess pattern perception within dot matrix arrangements. Although these images are immediately recognizable to human observers, they remain entirely undecipherable to state-of-the-art models (Table 1), and even increasing the rollout count fails to improve performance (Fig. 2).

To further analyze the reasons for model failure, we conducted a preliminary analysis and found three key insights. First, Grad-CAM reveals that models typically fail to correctly locate target regions in both the vision tower and language backbone. Second, in-context learning and fine-tuning the language backbone provide no improvement, while fine-tuning the vision tower enables rapid adaptation, *confirming our benchmark evaluates visual capabilities rather than knowledge and pure reasoning abilities*. Third, downsampling improves HiddenText performance more than blurring without resizing, which aligns with how vision encoders partition images into fixed-size patches during processing and highlights limitations in current visual encoding architectures. In the next version, we will release more diverse tasks with exploring methods to improve visual generalization, such as integrating reasoning into the perception stage, aligning both pixel and semantic levels, etc.

## 2 EMPIRICAL STUDY

### 2.1 DATASET

We create four specialized datasets of ***TET*** to evaluate the perception edge of MLLMs: (1) **HiddenText**, comprises scale-variant items where text is rendered as shapes within the figure, appearing as text when reduced and resolving into a complete image when magnified, which contains 150 images. (2) **3DCaptcha**: Involves recognition challenges constructed with curved characters in the three-dimensional space, which consists of 150 Captchas. (3) **ColorBlind**: similar to Ishihara tests (Ishihara, 1951), but augmented with confounding colored dots that are chromatically similar to the central character to increase difficulty. We generate 150 such test images. (4) **ChineseLigatures**: features complex glyphs synthesized through character decomposition, morphological transformation, and fusion of multiple Chinese characters, which include 40 different words or phrases. We show some cases in Fig. 1, the creation details can be found in Appendix A.1.

Table 1: **Performance Evaluation.** Pass@1 and pass@32 (%) of 15 MLLMs on four tasks of **TET**.

| Models/Tasks         | HiddenText |         | 3DCaptcha |         | ColorBlind |         | ChineseLigatures |         |
|----------------------|------------|---------|-----------|---------|------------|---------|------------------|---------|
|                      | Pass@1     | Pass@32 | Pass@1    | Pass@32 | Pass@1     | Pass@32 | Pass@1           | Pass@32 |
| OpenAI o1            | 0          | 0       | 0         | 0       | 0          | 1.33    | 0                | 0       |
| Claude-4-Sonnet      | 0          | 0       | 0         | 0       | 0          | 0       | 0                | 0       |
| Gemini 2.5 Pro       | 0          | 0       | 0         | 0       | 0          | 0       | 2.5              | 5       |
| Seed-1-6-250615      | 0          | 0       | 0         | 0       | 0          | 0       | 2.5              | 2.5     |
| Qwen2.5VL-72B        | 0          | 0       | 0         | 0       | 0          | 0       | 0                | 0       |
| Qwen2.5VL-7B         | 0          | 0.67    | 0         | 0       | 0          | 0       | 0                | 2.5     |
| QVQ-72B              | 0          | 0       | 0         | 0       | 0          | 0       | 0                | 0       |
| Qwen2.5-Omni-7B      | 0          | 0       | 0         | 0       | 0          | 0       | 0                | 2.5     |
| InternVL3-78B        | 0          | 0       | 0         | 0       | 0          | 0       | 0                | 0       |
| MiniCPM-o-2.6        | 0          | 0       | 0         | 0       | 0          | 0       | 0                | 2.5     |
| Show o2              | 0          | 0       | 0         | 0       | 0          | 0       | 0                | 0       |
| Bagel                | 0          | 0       | 0         | 0       | 0          | 0       | 0                | 0       |
| Janus-pro            | 0          | 0       | 0         | 0       | 0          | 0       | 0                | 0       |
| kimi-vl-a3b          | 0          | 0       | 0         | 0       | 0          | 0       | 0                | 5       |
| kimi-vl-a3b-thinking | 0          | 0       | 0         | 0       | 0          | 0       | 0                | 0       |

## 2.2 SETTINGS

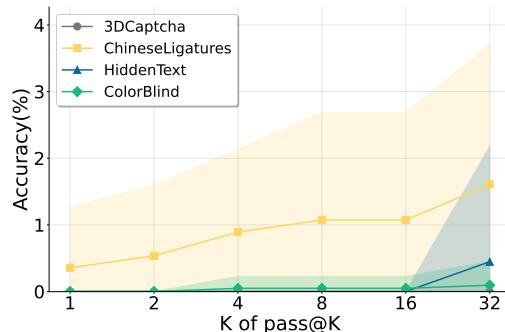
To comprehensively evaluate performance across diverse model architectures, we conduct extensive experiments on 15 models of different structures and types. These models can be divided into the following three categories: (1) unified multimodal models: Show o2 (Xie et al., 2025), Bagel (Deng et al., 2025), and Janus-pro (Chen et al., 2025) (2) API-based closed-source models: Claude 4-Sonnet (Anthropic, 2025), Gemini 2.5 Pro (Team, 2025b), OpenAI o1 (Jaech et al., 2024), and Seed-1-6-250615 (Team, 2025a) (3) open-source API models: Qwen2.5VL-72B (Bai et al., 2025), QVQ-72B (Team, 2024b), Qwen2.5-Omni-7B (Xu et al., 2025), InternVL3-78B (Zhu et al., 2025), MiniCPM-o-2.6 (Team, 2025c), kimi-vl-a3b (Team et al., 2025), kimi-vl-a3b-thinking (Team et al., 2025), and Qwen2.5VL-7B (Bai et al., 2025). We maintain original inference settings for unified models and configure all others with temperature 0.3 and 16384 max tokens.

We employ Pass@1 and Pass@K metrics (Chen et al., 2021) to evaluate model performance. Pass@1 measures the percentage of problems solved correctly in a single generation, reflecting the model’s accuracy on the specified tasks. Pass@K evaluates the percentage of problems for which at least one correct solution is obtained across K independent generations, capturing the exploration capability and solution diversity of the models. These metrics collectively provide a comprehensive evaluation framework that assesses both the task-specific accuracy and the problem-solving robustness through diverse solution exploration across our experimental tasks.

## 2.3 EVALUATION

Table 1 presents the performance of state-of-the-art multimodal language models on our proposed visual understanding tasks. Some response cases can be found in Appendix A.5. While contemporary MLLMs have demonstrated remarkable capabilities across a broad spectrum of multimodal reasoning-based tasks, our evaluation reveals significant limitations when confronted with the specific perceptual challenges presented in our dataset.

The results indicate that even the most advanced models struggle considerably with these tasks. Across all four evaluation domains—*HiddenText*, *3DCaptcha*, *ColorBlind*, and *ChineseLigatures*—the performance remains substantially

Figure 2: **Pass@k Results.** Mean and variance curves of pass@k on four tasks of **TET**.

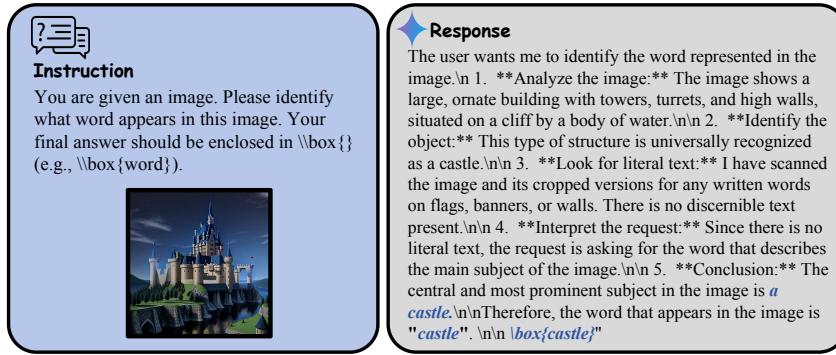


Figure 3: **Model response on question of HiddenText.** The goal is to identify the hidden word in an image. Gemini-2.5-Pro-0506 answers the hidden word as “castle”.

below acceptable thresholds. Most models achieve zero success rates on pass@1 evaluation, with only marginal improvements observed in certain cases.

Notably, some models demonstrate modest gains when evaluated with pass@32 compared to pass@1 metrics. For instance, several models, including Qwen2-VL-7B shows incremental improvements from 0% to 0.67% on specific tasks. However, these improvements, while statistically measurable, remain practically negligible and far from the performance levels required for reliable deployment. The consistently low performance across diverse model architectures and scales suggests that these perceptual challenges represent fundamental limitations in current MLLMs capabilities rather than model-specific deficiencies.

Fig. 2 illustrates the performance curves showing how average performance across different tasks varies with increasing values of K in pass@K, along with the corresponding standard deviations. As K increases, existing MLLMs exhibit minimal variation in average performance across different tasks. Even in cases where certain tasks demonstrate performance improvements, the peak performance gains remain below 4%. This finding suggests that those tasks cannot be effectively addressed through exploration within the reasoning space, which fundamentally differs from previous datasets. The relatively flat performance curves across all evaluated tasks further emphasize that the bottleneck lies not in the breadth of reasoning exploration, but rather in the visual perception capabilities required for accurate image understanding and interpretation. The response case presented in Fig. 3 further corroborates this conclusion. More cases can be found in Fig. 13 of Appendix A.5.

### 3 PRELIMINARY ANALYSIS

#### 3.1 INTERPRETATION OF MLLMs WITH GRAD-CAM

To explore why the model cannot perceive images accurately, we conduct an analysis of the model’s Grad-CAM following previous works (Selvaraju et al., 2017; Zhang et al., 2024b) for all datasets. We systematically examine two representative models from the Qwen2.5-VL series (7B and 72B parameters). Our analysis encompasses attention maps from both the visual backbone and language encoder components, uniformly sampled across multiple layers to capture attention evolution throughout the architectural depths of each component.

**Information flow in image encoder.** As illustrated in Fig. 4, 5, 6, and 7, given the lack of interaction between the visual encoder and text input, the image encoder’s primary role is to extract a generic representation of images to assist the LLM decoder. Analysis shows that for most tasks, while ViT does allocate attention across various regions of the image, this attention is often directed outside the target character regions or only captures partial segments of them. This suggests that the model’s image encoder struggles to effectively focus on the textural features corresponding to character regions and instead prioritizes object-level features within the image. Such disparities in visual attention prevent the model from truly comprehending the image content, thereby impairing its ability to address relevant tasks. Notably, in the case of the 3DCaptcha task, ViT’s recognition pattern appears to be guided by specific regions rather than adopting a global view. This behavior

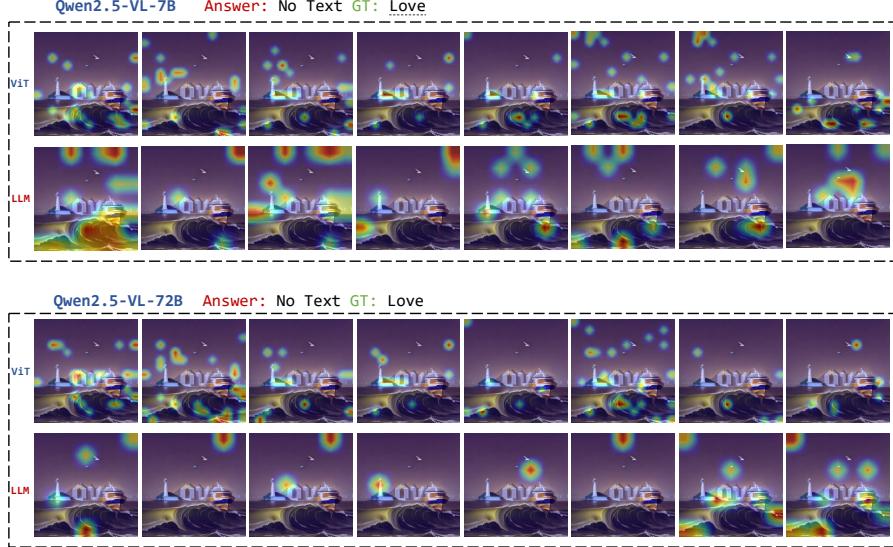


Figure 4: Grad-CAM of Qwen2.5-VL Series Models on HiddenText subset.

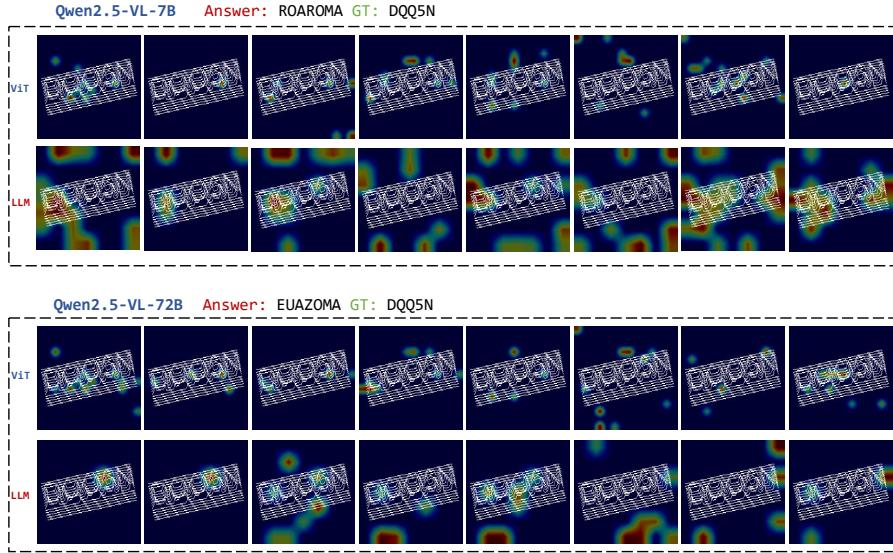


Figure 5: Grad-CAM of Qwen2.5-VL Series Models on 3DCaptcha.

may stem from inherent limitations in the generalization and robustness of ViT when processing out-of-domain inputs.

**Information flow in LLM decoder.** As illustrated in Fig. 4, 5, 6, and 7, regarding the information flow within the LLM decoder, we found that LLM decoders of different sizes (7B and 72B) exhibit no substantial differences in their attention patterns across our four tasks. It is observed that, except for the *ChineseLigature* task, the LLM decoder consistently fails to focus on the precise regions containing text or character information; instead, it scatters attention over irrelevant regions or completely ignores critical visual elements. Such inconsistency between the model’s attention patterns and the actual locations of important visual features indicates fundamental limitations in the model’s visual perception ability, making it difficult to focus on the key regions relevant to task resolution.

**Speculations on the causes of failure across different tasks.** As shown in Fig. 4, for *HiddenText*, we observed that MLLMs struggle to recognize the overall symbols formed by objects in the image, indicating a lack of a global perspective for comprehension. As shown in Fig. 6, for *Chi-*

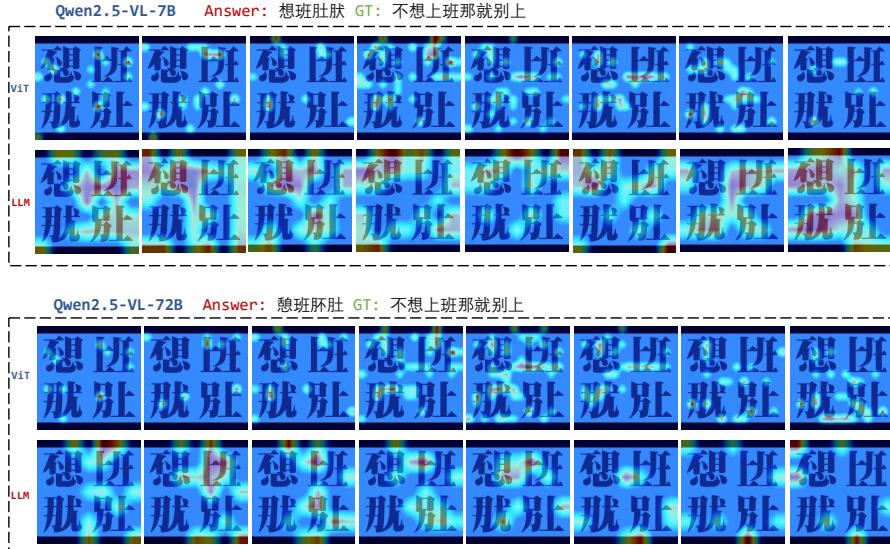


Figure 6: Grad-CAM of Qwen2.5-VL Series Models on ChineseLigature subset.

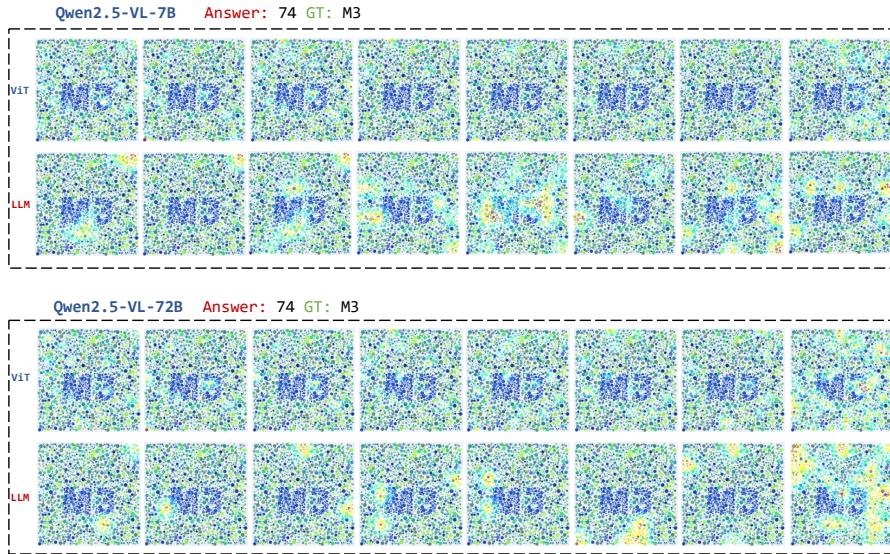


Figure 7: Grad-CAM of Qwen2.5-VL Series Models on ColorBlind.

*neseLigature*, it was noted that MLLMs can actually perceive the characters in the image but lack the imaginative capacity to extend them into commonly used sentences, instead engaging in rigid recognition of obscure characters. As shown in Fig. 5, for *3DCaptcha*, observations revealed that due to the particularity of its character composition pattern, the model is prone to interference and thus unable to distinguish the correct characters. As shown in Fig. 7, for *ColorBlind*, we found that MLLMs are easily disturbed by noise from surrounding circles of the same color, making it difficult to mentally construct the correct character from the overall shape formed by small circles of the same color, thereby resulting in recognition errors.

### 3.2 SUPERVISED FINETUNING

To explore whether the lack of domain knowledge causes the inability of the model to understand these tasks, we conduct supervised fine-tuning (SFT) on Qwen2.5-7B-VL using domain-specific data from the corresponding tasks, examining whether the model can acquire knowledge through

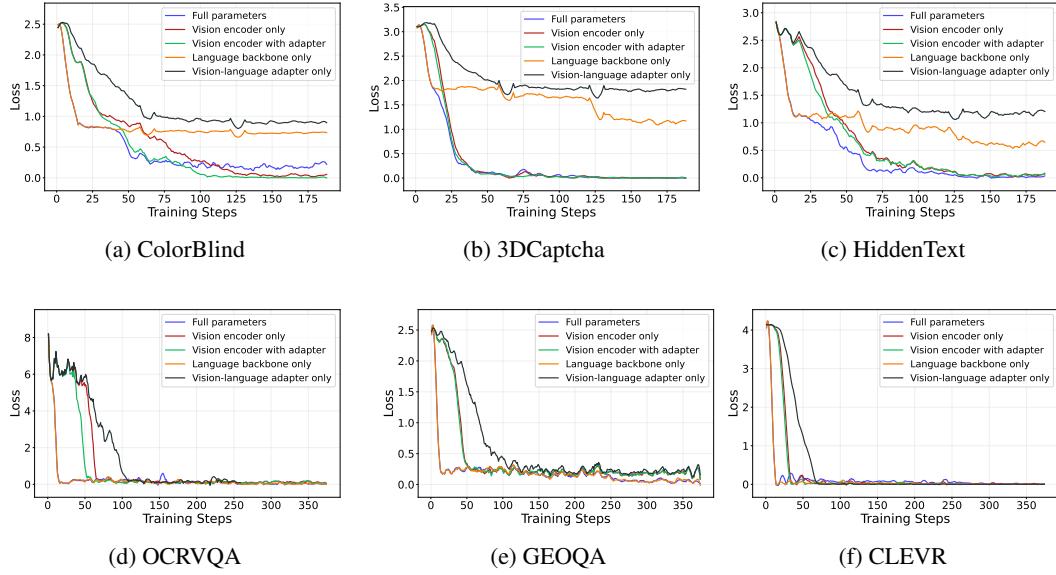


Figure 8: **Training Curves.** Training loss curve for different settings on finetuning parameters for both our tasks and traditional tasks.

SFT and handle the tasks effectively. We design five different training configurations targeting different model components, selectively updating various architectural components to isolate the contribution of each module to the final performance. We examine five training configurations based on parameter updates: (1) Full parameters; (2) Vision encoder only; (3) Vision encoder with vision-language adapter; (4) Language backbone only; (5) Vision-language adapter only. Since constructing training data for *ChineseLigature* proved challenging, we mainly discussed the other three datasets here. We admit that *ChineseLigature* likely demands both perceptual and reasoning abilities, and we'll explore this in next version.

Table 2: **SFT Analyses.** Accuracy (%) of Qwen2.5-VL 7B after finetuning different parameters.

| Parameters Updated/Tasks     | HiddenText |         | 3DCaptcha |         | ColorBlind |         |
|------------------------------|------------|---------|-----------|---------|------------|---------|
|                              | Pass@1     | Pass@32 | Pass@1    | Pass@32 | Pass@1     | Pass@32 |
| W/O Training                 | 0          | 0       | 0         | 0       | 0          | 1.3     |
| Full Parameters              | 90.00      | 94.67   | 95.33     | 98.00   | 77.33      | 79.33   |
| Vision Encoder Only          | 86.67      | 94.67   | 94.00     | 98.00   | 87.33      | 98.67   |
| Vision Encoder with Adapter  | 82.00      | 94.67   | 95.33     | 97.33   | 99.33      | 99.33   |
| Language Backbone Only       | 0          | 2.67    | 0         | 0       | 0.67       | 14.00   |
| Vision-language Adapter Only | 0          | 5.33    | 0         | 0       | 1.33       | 6.67    |

Table 2 shows that fine-tuning the vision encoder is essential for performance improvements on our tasks, while updating parameters without the vision encoder has few effect, which indicate our benchmark is challenge for the vision tower rather than reasoning or knowledge of the language backbone. Furthermore, fig. 8 demonstrates that configurations excluding visual finetuning “Language backbone only” and “Vision-language adapter only”) plateau early and achieve suboptimal convergence, while those including visual fine-tuning reach lower loss values more efficiently. In contrast, all configurations converge similarly on existing datasets like OCRVQA, GEOQA, and CLEVR. This difference also indicates that our tasks require enhanced visual perception capabilities that go beyond language knowledge and reasoning improvements. Baseline datasets likely fall within the visual domain coverage of current MLLM pre-training data, requiring only reasoning enhancements rather than fundamental visual adaptation.

We also observed that the trajectories of language backbone fine-tuning and full parameter fine-tuning overlap in the early stages, suggesting that MLLMs initially focus on adjusting language parameters before fine-tuning other components. We will analyze this further in the next version.

Table 3: **On In-context Learning.** Pass@1 and pass@32 of three MLLMs with 3-example in-context learning on four tasks of **TET**.

| Models/Tasks    | HiddenText |         | 3DCaptcha |         | ColorBlind |         | ChineseLigatures |         |
|-----------------|------------|---------|-----------|---------|------------|---------|------------------|---------|
|                 | Pass@1     | Pass@32 | Pass@1    | Pass@32 | Pass@1     | Pass@32 | Pass@1           | Pass@32 |
| Qwen2.5VL-72B   | 0          | 0       | 0         | 0       | 0          | 0       | 0                | 0       |
| +ICL            | 0          | 2.0     | 0         | 0       | 0          | 1.33    | 0                | 0       |
| Gemini          | 0          | 0       | 0         | 0       | 0          | 0       | 2.5              | 5.0     |
| +ICL            | 0          | 0       | 0         | 0       | 0          | 4.0     | 7.5              | 20.0    |
| Seed-1-6-250615 | 0          | 0       | 0         | 0       | 0          | 0       | 2.5              | 2.5     |
| +ICL            | 0          | 0       | 0         | 0       | 0.67       | 0.67    | 0                | 5.0     |

### 3.3 IN-CONTEXT LEARNING

To explore whether different learning paradigms can influence knowledge acquisition and enhance model perceptual patterns, we investigate an in-context learning approach that enables the model to acquire knowledge through contextual examples. Similar to SFT, we utilize domain-specific data that aligns with the corresponding dataset to serve as examples, providing these examples within the context to facilitate the model’s learning of relevant perceptual patterns. Specifically, during inference, we augment the context of each data query by incorporating three image-answer pairs as demonstrations.

As shown in Table 3, incorporating the same-domain data as exemplars yields virtually no improvement in model performance on the corresponding tasks. This finding demonstrates that the capabilities required for these tasks cannot be enhanced through knowledge supplementation via in-context learning. The key to resolving the model’s perceptual limitations does not lie in providing additional knowledge through exemplar data, but rather points to more fundamental architectural or representational deficiencies which cannot be addressed through contextual demonstrations alone.

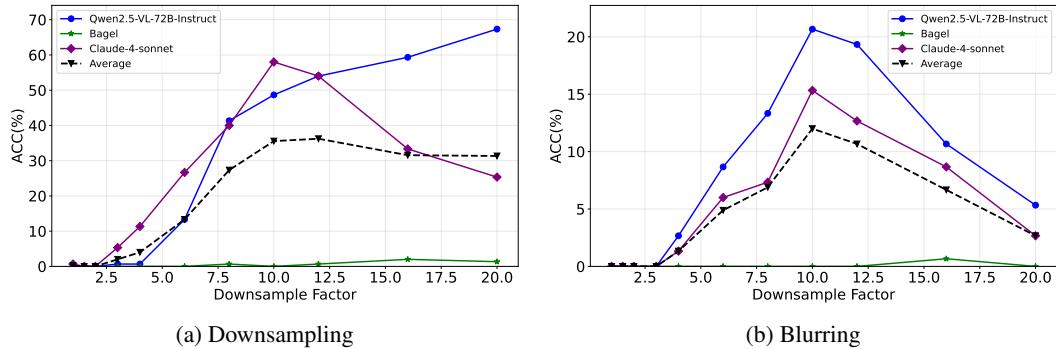


Figure 9: **Downsampling and Blurring.** Pass@1 accuracy of four MLLMs on *HiddenText* under downsampling and blurring conditions.

### 3.4 SAMPLING ON IMAGES

Since performance on *HiddenText* is somewhat dependent on image resolution, we test two different resolution settings: (1) direct downsampling by BOX filtering (Gonzalez & Woods, 2018) and (2) blurring by downsampling and upsampling back to original resolution. Fig. 9 illustrates the results between downsample factor and model performance under these two resolution settings.

As shown in Fig. 9a, as the downsample factor increases, model performance under low resolution gradually improves. Unlike the other models, Claude shows a performance degradation in Set-

tings with high downsampling factor. At extremely low resolutions, the original image degrades into simple character patterns, making the task approximate to OCR, where the model can readily identify authentic characters. However, as shown in Fig. 9b, under the blurring setting, model performance exhibits an inverted bell-shaped curve, with peak performance inferior to that of direct downsampling. This improvement aligns with how vision encoders partition images into fixed-size patches. Downsampling simplifies both patch content and reduces the total number of patches while highlighting character textures within each patch, whereas blurring introduces additional noise and makes character textures more ambiguous.

## 4 CONCLUSION

In this study, we introduced the Turing Eye Test (TET), a perception-oriented benchmark that reveals fundamental limitations in current Multimodal Large Language Models’ visual understanding capabilities. Through four diagnostic tasks involving concealed text, 3D Captchas, Chinese character compositions, and color blind test charts, we demonstrated that state-of-the-art MLLMs exhibit catastrophic failures on perceptual tasks that humans solve intuitively. Our analysis reveals that these failures stem from limitations in the vision tower’s generalization abilities rather than deficiencies in language reasoning or knowledge. While in-context learning and language backbone fine-tuning proved ineffective, targeted fine-tuning of the vision tower enabled rapid adaptation, highlighting a critical gap between current MLLM architectures and human-like visual perception. These findings underscore the need for improved visual generalization methods in MLLMs and establish TET as a valuable diagnostic tool for evaluating genuine perceptual capabilities beyond traditional reasoning-focused benchmarks.

Inspired by recent advances in large reasoning models (Guo et al., 2025; Jaech et al., 2024) that improve generalization for language understanding, injecting reasoning capabilities into the perception stage may also enhance vision encoder generalization. For example, we could train vision transformers using GRPO (Shao et al., 2024) while keeping the LLM backbone parameters frozen. Future work will include the full set of TET’s tasks and explore some methods to bridge the perception gap between MLLMs and human visual understanding.

## REFERENCES

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Anthropic. Introducing Claude 4. Anthropic News, May 2025. URL <https://www.anthropic.com/news/clause-4>. Accessed: [Insert Access Date].
- Maximilian Augustin, Yannic Neuhaus, and Matthias Hein. Dash: Detection and assessment of systematic hallucinations of vlms. *arXiv preprint arXiv:2503.23573*, 2025.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023. URL <https://arxiv.org/abs/2309.16609>.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Ruisheng Cao, Fangyu Lei, Haoyuan Wu, Jixuan Chen, Yeqiao Fu, Hongcheng Gao, Xinzhuang Xiong, Hanchong Zhang, Wenjing Hu, Yuchen Mao, et al. Spider2-v: How far are multimodal agents from automating data science and engineering workflows? *Advances in Neural Information Processing Systems*, 37:107703–107744, 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024a.

Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-star: Benchmarking video-lmms on video spatio-temporal reasoning. *arXiv preprint arXiv:2503.11495*, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning, 2025. URL <https://arxiv.org/abs/2501.00321>.

Hongcheng Gao, Jiashu Qu, Jingyi Tang, Baolong Bi, Yue Liu, Hongyu Chen, Li Liang, Li Su, and Qingming Huang. Exploring hallucination of large multimodal models in video understanding: Benchmark, analysis and mitigation, 2025. URL <https://arxiv.org/abs/2503.19622>.

Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Pearson, New York, 4th edition, 2018.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in lmms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.

Shinobu Ishihara. Tests for colour-blindness, 1951.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Wonjun Lee, Doehyeon Lee, Eugene Choi, Sangyoong Yu, Ashkan Yousefpour, Haon Park, Bumsub Ham, and Suhyun Kim. Elite: Enhanced language-image toxicity evaluation for safety. *arXiv preprint arXiv:2502.04757*, 2025a.

Youngwan Lee, Kangsan Kim, Kwanyong Park, Ilcahe Jung, Soojin Jang, Seanie Lee, Yong-Ju Lee, and Sung Ju Hwang. Holisafe: Holistic safety benchmarking and modeling with safety meta token for vision-language model. *arXiv preprint arXiv:2506.04704*, 2025b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024b.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024. URL <https://arxiv.org/abs/2310.02255>.

Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jiwen Liu, Jifeng Dai, Yu Qiao, and Xizhou Zhu. Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24960–24971, 2025.

Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, and Zhe Gan. MIA-bench: Towards better instruction following evaluation of multimodal llms, 2025. URL <https://arxiv.org/abs/2407.01509>.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL <https://arxiv.org/abs/2402.03300>.

ByteDance Seed Team. Introduction to techniques used in seed1.6, June 2025a. URL [https://seed.bytedance.com/en/seed1\\_6](https://seed.bytedance.com/en/seed1_6). Technical Report.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024a.

Gemini 2.5 Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025b. URL <https://arxiv.org/abs/2507.06261>.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haoning Wu, Haotian Yao, Haoyu Lu, Heng Wang, Hongcheng Gao, Huabin Zheng, Jiaming Li, Jianlin Su, Jianzhou Wang, Jiaqi Deng, Jiezhong Qiu, Jin Xie, Jinhong Wang, Jingyuan Liu, Junjie Yan, Kun Ouyang, Liang Chen, Lin Sui, Longhui Yu, Mengfan Dong, Mengnan Dong, Nuo Xu, Pengyu Cheng, Qizheng Gu, Runjie Zhou, Shaowei Liu, Sihan Cao, Tao Yu, Tianhui Song, Tongtong Bai, Wei Song, Weiran He, Weixiao Huang, Weixin Xu, Xiaokun Yuan, Xingcheng Yao, Xingzhe Wu, Xinxing Zu, Xinyu Zhou, Xinyuan Wang, Y. Charles, Yan Zhong, Yang Li, Yangyang Hu, Yanru Chen, Yejie Wang, Yibo Liu, Yibo Miao, Yidao Qin, Yimin Chen, Yiping Bao, Yiqin Wang, Yongsheng Kang, Yuanxin Liu, Yulun Du, Yuxin Wu, Yuzhi Wang, Yuzi Yan, Zaida Zhou, Zhaowei Li, Zhejun Jiang, Zheng Zhang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Zijia Zhao, and Ziwei Chen. Kimi-VL technical report, 2025. URL <https://arxiv.org/abs/2504.07491>.

OpenBMB MiniCPM-o Team. Minicpm-o 2.6: A gpt-4o level mllm for vision, speech, and multi-modal live streaming on your phone, 2025c.

Qwen Team. Qvq: To see the world with wisdom, December 2024b. URL <https://qwenlm.github.io/blog/qvq-72b-preview/>.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multi-modal mathematical reasoning with math-vision dataset, 2024a. URL <https://arxiv.org/abs/2402.14804>.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b.

Zhaowei Wang, Wenhao Yu, Xiyu Ren, Jipeng Zhang, Yu Zhao, Rohit Saxena, Liang Cheng, Ginny Wong, Simon See, Pasquale Minervini, Yangqiu Song, and Mark Steedman. Mmlongbench: Benchmarking long-context vision-language models effectively and thoroughly, 2025. URL <https://arxiv.org/abs/2505.10610>.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024. URL <https://arxiv.org/abs/2407.15754>.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal lilm logical reasoning benchmark in visual contexts, 2024. URL <https://arxiv.org/abs/2407.04973>.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osword: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Shuai Bai, et al. Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy. *arXiv preprint arXiv:2412.02210*, 2024.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2024. URL <https://arxiv.org/abs/2308.02490>.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024a.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024b.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?, 2024a. URL <https://arxiv.org/abs/2403.14624>.

Xiaofeng Zhang, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. From redundancy to relevance: Enhancing explainability in multimodal large language models. *arXiv e-prints*, pp. arXiv–2406, 2024b.

Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, et al. Multitrust: A comprehensive benchmark towards trustworthy multimodal large language models. *Advances in Neural Information Processing Systems*, 37:49279–49383, 2024c.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836*, 2024.

## A APPENDIX

### A.1 CREATION DETAILS OF DATASETS

**HiddenText.** We construct the *HiddenText* dataset using a Stable Diffusion pipeline with ControlNet guidance to embed textual information within aesthetically pleasing artistic images. The generation process first renders input text as high-contrast binary images, then employs a ControlNet model to guide the diffusion process while preserving text readability through artistic transformation. By varying style prompts and generation parameters, we produce diverse artistic renderings where text content remains human-readable despite being visually integrated into complex visual compositions.

**3DCaptcha.** We construct the *3DCaptcha* dataset using an automated Python pipeline that transforms 2D text renderings into pseudo-3D wireframe visualizations. The generation process renders random 5-character alphanumeric sequences as 2D images, applies camera transformations with randomized viewpoints, maps pixel intensities to depth coordinates, and projects the results onto a standardized canvas using wireframe line drawing algorithms.

**ColorBlind.** We construct the *ColorBlind* dataset through an automated generation system that creates Ishihara-style color vision test images containing 2-character targets. The generation pipeline renders 2-character combinations (excluding visually confusing pairs such as 0-O, 1-I-L, 5-S, 8-B, Q-O) as text masks, then populates the image with randomly distributed colored dots using four distinct color schemes (red-green, green-red, orange-green, brown-green) to simulate different types of color vision deficiency tests. Character regions use uniform color palettes while background areas employ contrasting colors with 20% interference colors (similar but distinct hues) to increase recognition difficulty.

**ChineseLigatures.** We construct the *ChineseLigatures* dataset by leveraging GPT-4o’s visual generation capabilities to create Chinese compound character images. The construction process involved collecting representative examples of Chinese ligature characters as visual references, then designing carefully crafted prompts to guide GPT-4o in generating structurally accurate and visually consistent compound character variations.

## A.2 EVALUATION WITH HIGH TEMPERATURE

To further investigate whether the model’s failure stems from a fundamental limitation in visual perception or merely from a tendency to follow a single, incorrect reasoning path under default (low-temperature) settings, we conducted an additional evaluation using a high-temperature sampling strategy. In language model generation, increasing the temperature parameter enhances the randomness and diversity of the output. For this experiment, we specifically set the sampling temperature to 0.8 to encourage the model to explore a wider range of possibilities.

As shown in Table 4, the performance of all models across the four TET tasks remains extremely poor, with the vast majority of success rates (Pass@1 and Pass@32) still at or near zero. Although negligible improvements were observed in the Pass@32 metric for isolated tasks like ChineseLigatures, this is far from sufficient to alter the conclusion of the models’ overall failure.

This finding strongly supports that the bottleneck in current Multimodal Large Language Models is a fundamental deficiency in visual perception, not a lack of reasoning capability in the language backbone. Merely increasing the exploratory scope and diversity of the reasoning process (i.e., high-temperature sampling) cannot compensate for the failure at the initial stage of visual information extraction.

Table 4: Model Performance on Visual Understanding Tasks.

| Models               | HiddenText |         | 3DCaptcha |         | ColorBlind |         | ChineseLigatures |         |
|----------------------|------------|---------|-----------|---------|------------|---------|------------------|---------|
|                      | Pass@1     | Pass@32 | Pass@1    | Pass@32 | Pass@1     | Pass@32 | Pass@1           | Pass@32 |
| OpenAI o1            | 0          | 0       | 0         | 0       | 0          | 1.33    | 0                | 0       |
| Claude-4-Sonnet      | 0          | 0       | 0         | 0       | 0          | 0       | 0                | 0       |
| Gemini 2.5 Pro       | 0          | 0       | 0         | 0       | 0.67       | 0.67    | 0                | 0       |
| Seed-1-6-250615      | 0          | 0       | 0         | 0       | 0          | 0       | 0                | 2.5     |
| Qwen2.5VL-72B        | 0          | 0       | 0         | 0       | 0          | 0       | 0                | 0       |
| Qwen2.5VL-7B         | 0          | 0       | 0         | 0       | 0          | 0       | 0                | 5       |
| QVQ-72B              | 0          | 0       | 0         | 0       | 0          | 0.67    | 0                | 0       |
| Qwen2.5-Omni-7B      | 0          | 0.67    | 0         | 0       | 0          | 0       | 0                | 2.5     |
| InternVL3-78B        | 0          | 0       | 0         | 0       | 0          | 0.67    | 0                | 0       |
| MiniCPM-o-2.6        | 0          | 0       | 0         | 0       | 0          | 1.33    | 0                | 2.5     |
| kimi-vl-a3b          | 0          | 0       | 0         | 0       | 0          | 0.67    | 0                | 0       |
| kimi-vl-a3b-thinking | 0          | 0.67    | 0         | 0       | 0          | 0       | 0                | 5       |

## A.3 RELATED WORK

### A.3.1 MULTIMODAL LARGE LANGUAGE MODELS

The architecture of Multimodal Large Language Models (MLLMs) are commonly grouped into two design families.(1) The dominant paradigm is the modular Vision Language Model (VLM), which connects a pre-trained vision encoder with a Large Language Model (LLM) via a lightweight projection module. This flexible and widely adopted approach is represented by many of the open-source models evaluated in this study, including the Qwen2.5-VL series (Bai et al., 2025; 2023), QVQ-72B (Team, 2024b), InternVL3-78B (Zhu et al., 2025), MiniCPM-o-2.6 (Team, 2025c), and the Kimi-VL (Team et al., 2025) models, and is foundational to other influential models like the LLaVA series (Liu et al., 2023; 2024a;b) and BLIP-2 (Li et al., 2023). (2) Another line of work has developed unified models, which process visual and linguistic tokens within a more tightly coupled or even single framework to achieve a more seamless cross-modal interaction. This category includes models that we evaluate such as Show-o2 (Xie et al., 2025), Bagel (Deng et al., 2025), Janus-pro (Chen et al., 2025), as well as other notable frameworks such as Mono-InternVL (Luo et al., 2025) and Transfusion (Zhou et al., 2024). By covering both families, our work checks

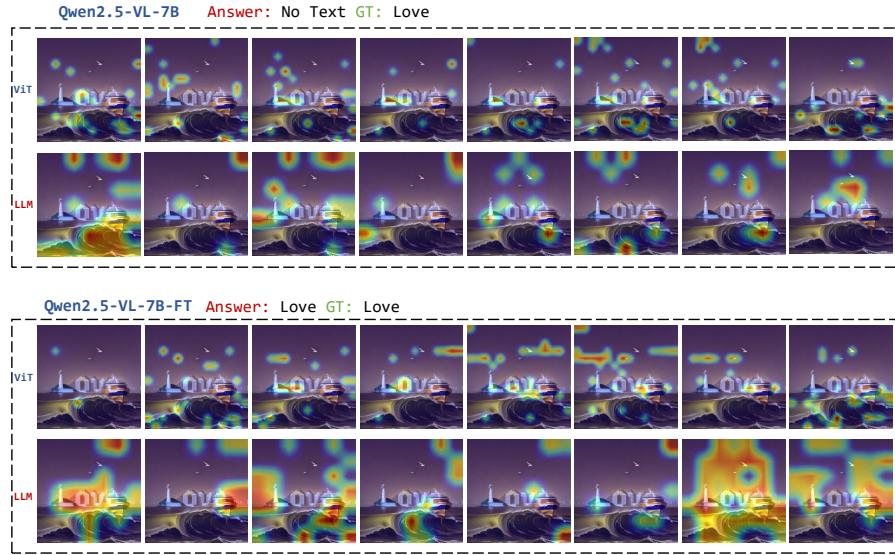


Figure 10: Grad-CAM of Qwen2.5-VL-7B before and after visual fine-tuning on HiddenText.

whether these different design choices fundamentally alter their visual perception capabilities in *TET*.

### A.3.2 EVALUATION BENCHMARKS FOR MLLMs

The evaluation landscape for MLLMs has rapidly expanded from foundational skill assessments to comprehensive evaluations of complex reasoning and robustness. These developments include the following areas: (1) **General visual question-answer benchmarks** that test diverse capabilities through open-ended formats, such as LLaVABench (Liu et al., 2023), MM-Vet (Yu et al., 2024), V-StaR (Cheng et al., 2025), Video-MMMU (Hu et al., 2025), LongVideoBench (Wu et al., 2024), MMLongBench (Wang et al., 2025) and MIA-Bench (Qian et al., 2025); (2) **Reasoning benchmarks**, including multi-disciplinary knowledge tests like MMMU (Yue et al., 2024a) and MMMU-Pro (Yue et al., 2024b), a suite of mathematical reasoning evaluations like MathVista (Lu et al., 2024), MathVerse (Zhang et al., 2024a), ScienceQA (Lu et al., 2022), DynaMath (Zou et al., 2024) and MathVision (Wang et al., 2024a), and logical problem-solving with LogicVista (Xiao et al., 2024); (3) **Domain-Specific benchmarks** that target specialized applications and task-specific scenarios, such as OCR-focused tasks like OCRCBenchV2 (Fu et al., 2025) and agentic tasks like OS-World (Xie et al., 2024) and Spider2-V (Cao et al., 2024) and CC-OCR (Yang et al., 2024); and (4) Trustworthiness benchmarks, that examine model reliability and safety aspects, such as evaluations for evaluating hallucination tendencies like HAVEN (Gao et al., 2025), MultiTrust (Zhang et al., 2024c) and DASH (Augustin et al., 2025), and comprehensive safety assessments for MLLMs like HoliSafe (Lee et al., 2025b) and ELITE (Lee et al., 2025a).

### A.4 INTERPRETING VISUAL FINE-TUNING WITH GRAD-CAM

Fig. 10, 11, 12 presents Grad-CAM visualizations of Qwen2.5-VL-7B before and after vision module fine-tuning across different datasets. Following vision module fine-tuning, the model demonstrates enhanced perceptual capabilities, as attention coverage over effective character regions across inter-module interactions increases. This phenomenon validates that targeted optimization of the vision module effectively improves the generalization of the model’s perceptual patterns.

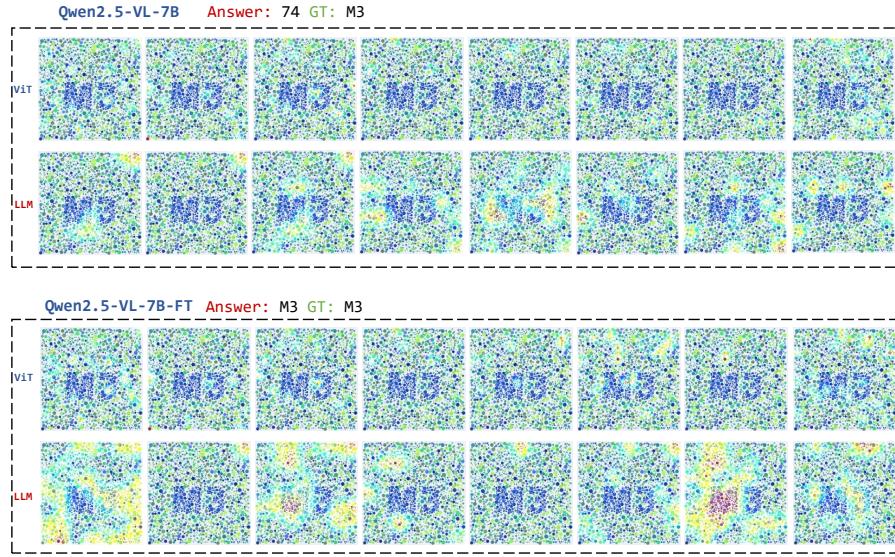


Figure 11: Grad-CAM of Qwen2.5-VL-7B before and after visual fine-tuning on ColorBlind.

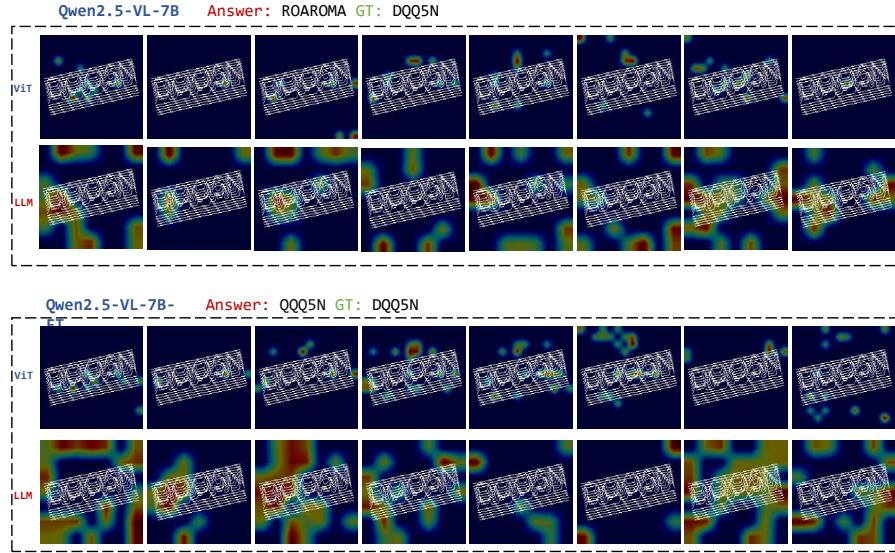


Figure 12: Grad-CAM of Qwen2.5-VL-7B before and after visual fine-tuning on 3DCaptcha.

### A.5 CASES OF FOUR TASKS

To intuitively demonstrate the challenges that the state-of-the-art Multimodal Large Language Models face on our tasks, this section provides specific failure cases. Fig. 13 illustrates the responses from the Gemini2.5Pro-0506 model on the four tasks of the TET. These cases clearly indicate that the model’s failure is not due to a lack of reasoning ability of the language backbone, but rather stems from fundamental deviations in the perception stage of visual input.

You are given an image. Please identify *what word appears* in this image. Your final answer should be enclosed in `\boxed{}`.



Based on a close examination of the image and its cropped sections, the most plausible *word* that appears in the image is formed by a distinct shape on the building. On the white, upper section of the building, below the three vertical black slats, there is a *circular object* that resembles a security camera or an outdoor light. *This circular shape can be interpreted as the letter 'O'*. Since 'O' is also a word in the English language, this is the most likely answer. Other text-like features, such as the sign on the right-hand side of the building, appear to be garbled and do not form a recognizable word, which is a common artifact in AI-generated images like this one. The circular object provides the clearest and most interpretable representation of a letter that is also a word.

HiddenText

Gemini-2.5pro-0506

You are given an image. Please identify the *characters* in this image. Your final answer should be enclosed in `\boxed{}`.



*This image does not contain any characters.* It appears to be a graphical representation, such as a waveform or a 3D surface plot. Therefore, there are no characters to identify.

3DCaptcha

Gemini-2.5pro-0506

You are given an image. Please identify the *Chinese phrase, sentence, idiom, proverb, or solar term displayed* in this image. Your final answer should be enclosed in `\boxed{}`.

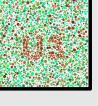


This is a highly stylized and auspicious composite Chinese character. *It is an artistic representation of the character 福 (luán)*. The standard character 福 refers to a mythical bird, similar to a phoenix, that symbolizes good fortune, peace, and marital harmony. This specific artistic version is created by combining several characters with positive meanings, all packed into a single glyph. \*\*王 (wáng)\*\*: Two "King" characters, symbolizing supreme status or great prosperity. \*\*好 (hǎo)\*\*: Meaning "good" or "order". \*\*禾 (hé)\*\*: Meaning "virtuous" or "beautiful". \*\*禾 (hé)\*\*: Meaning "grain," symbolizing a bountiful harvest. \*\*鸟 (niǎo)\*\*: The "bird" radical, indicating the nature of the character. All these components are combined to create a super-auspicious version of the already fortunate character 福.

ChineseLigatures

Gemini-2.5pro-0506

You are given an image. Please identify the *characters* in this image. Your final answer should be enclosed in `\boxed{}`.



The image contains characters formed by a pattern of dots. By focusing on the cluster of brown dots against the background of green and cyan dots, one can discern two letters. The letter *on the left is a 'U'*, and *the letter on the right is a 'C'*.

ColorBlind

Gemini-2.5pro-0506

Figure 13: Example responses from the Gemini model on the four tasks of TET. For each task, the model fails to provide correct analysis and conclusion based on its flawed initial perception.

**HiddenText.** In this figure, the model fails to perceive the macroscopic word *Star* formed by the arrangement of objects in the scene, instead focusing on describing the details of local objects and misidentifies a circular object on the building as the letter *O*.

**3DCaptcha.** When presented with characters that are distorted and stacked in three-dimensional space, the model cannot recognize them as a combination of letters and numbers. Instead, it interprets the image as a *graphical representation, such as a waveform or a 3D surface plot*, containing no specific characters.

**ChineseLigatures.** The model shows its limitations when dealing with creative glyphs synthesized from multiple components of Chinese characters. It attempts to deconstruct the whole into separate standard characters, but fails to perceive the holistic concept of the idiom *marital bliss*.

**ColorBlind.** The model struggles to distinguish the character pattern formed by dots of a specific color from the noisy background dots. In the case shown in the figure, it incorrectly identifies the pattern as *U* and *C*.