FADE: Probing the Limits of VLMs on fine-grained OCR

Deep Shah
Google LLC
shahdeep@google.com

Nehal Kathrotia
Google LLC
nehalk@google.com

Sanket Badhe
Google LLC
sanketbadhe@google.com

Abstract

Multimodal Large Language Models (MLLMs) have achieved remarkable success in semantic visual reasoning, yet their capacity for fine-grained, low-level perception remains critically under-evaluated. This perceptual fragility limits their reliability in noisy, real-world environments where visual signals are degraded. Furthermore, existing benchmarks often entangle visual perception with language priors, masking these underlying deficits. To address this, we introduce the **FAint numeric Detection Evaluation (FADE)** dataset, a novel evaluation suite designed to probe the limits of zero-shot Optical Character Recognition (OCR) in frontier MLLMs. By embedding synthetic, strictly numerical sequences over cluttered natural backgrounds at varying levels of transparency (α), FADE explicitly disentangles pure visual perception from semantic predictability. We evaluate state-of-the-art models including Gemini 3.0, Claude 4.5 Sonnet, and Gemma 3 against a specialized UNet segmentation baseline. Our results reveal a striking limitation in frontier architectures: while they achieve near-perfect transcription at high visibility, their performance collapses under high transparency. Conversely, the UNet pipeline maintains robust spatial grounding, significantly outperforming generalist models at the lowest visibility thresholds. FADE provides a reproducible dataset to expose and diagnose the perceptual breakage points of modern multimodal systems.

Dataset: FADE on Hugging Face

1 Introduction

The trajectory of Vision-Language Models (VLMs) and Multimodal Large Language Models (MLLMs) has been marked by extraordinary advancements, fundamentally bridging the gap between visual perception and natural language understanding (Achiam et al., 2023; Team et al., 2023). Trained on vast web-scale image-text datasets, frontier architectures such as GPT-4, Gemini, and Claude have achieved unprecedented success in high-level semantic tasks, including visual question answering (VQA), complex scene interpretation, and nuanced image captioning (Liu et al., 2023; Li et al., 2023). Consequently, these models are increasingly deployed in real-world, open-ended environments where they are expected to act as general-purpose visual agents.

Current MLLMs often rely heavily on powerful language priors and global semantic context, rather than robust, low-level visual processing (Tong et al., 2024; Villa et al., 2025). While they excel at identifying prominent objects or reasoning about a scene's overall narrative, their performance deteriorates sharply when confronted with fine-grained perceptual tasks that lack semantic anchors. One of the most demanding tests of this capability is Optical Character Recognition (OCR) in the wild. In real-world scenarios, text is frequently obscured by poor lighting, complex background textures, or partial transparency—conditions where context alone cannot recover the missing characters (Zhu et al., 2024).

The inability of VLMs to reliably perform robust text-in-image extraction carries profound security and safety implications for digital platforms. As automated content moderation systems increasingly rely on multimodal models to enforce community guidelines, malicious actors have adapted by embedding harmful content such as hate speech, scam URLs, phone-numbers or illicit narratives directly into images and memes (Wang et al., 2025, 2024). Because standard text-only filters cannot parse image pixels, a VLM's OCR along with the MLLM reasoning capability serves as the primary line of defense. When VLMs fail to detect visually perturbed, highly transparent, or cluttered text, bad actors can successfully execute visual prompt injections and moderation evasion tactics. Consequently, poor OCR performance directly translates to a degraded safety posture, allowing harmful content to

propagate across digital ecosystems unchecked.

Despite these high-stakes implications, existing evaluation paradigms have largely failed to isolate and quantify this specific perceptual fragility. Traditional OCR benchmarks predominantly feature high-contrast, opaque text, or rely on semantic contexts (like street signs or document headers) that allow language models to seamlessly "guess" obscured words (Singh et al., 2021). Conversely, modern VLM benchmarks tend to entangle visual perception with complex logical reasoning, making it difficult to determine whether a model failed because it could not reason, or simply because it could not *see* the underlying visual signal (Villa et al., 2025; Tong et al., 2024). There is a need for a controlled, objective benchmark that tests the exact breakage thresholds of multimodal perception without the confounding variable of semantic predictability.

To address this critical gap, we introduce the **FAint numeric Detection Evaluation**, a novel evaluation suite explicitly designed to stress-test the limits of fine-grained visual grounding and zero-shot OCR in frontier VLMs. Our dataset embeds synthetic, strictly numerical sequences, ensuring no linguistic context can aid prediction over highly cluttered, diverse backgrounds sampled from the COCO dataset. Crucially, we isolate visual degradation as a single independent variable by systematically modulating the transparency (α) of the watermark across a gradient of visibility, ranging from highly visible ($\alpha = 0.8$) to near-imperceptible ($\alpha = 0.2$).

By benchmarking state-of-the-art models-including Gemini 3.0, Claude 4.5 Sonnet, and Gemma 3 against a specialized UNet segmentation baseline, we reveal a striking deficit in frontier architectures. While VLMs achieve near-perfect transcription at high visibility, their zero-shot capabilities collapse under high transparency and background clutter. In contrast, our dedicated segmentation pipeline maintains robust spatial grounding, significantly outperforming the generalist models at the lowest visibility thresholds. Through this work, we provide the community with a reproducible framework to diagnose and rectify the low-level perceptual shortcomings of modern multimodal systems.

2 Related Work

2.1 Multimodal Large Language Models

With the remarkable advancements of Large Language Models (LLMs), recent research has extended their capabilities to multimodal domains by integrating visual information, giving rise to Multimodal Large Language Models (MLLMs) (Achiam et al., 2023; Team et al., 2023; Li et al., 2023; Liu et al., 2023). These models typically align visual features from pre-trained image encoders with LLMs via modality adaptation layers (Dosovitskiy et al., 2020). Early works like BLIP-2 (Li et al., 2023) pioneered this architecture by pre-training on image-text datasets and fine-tuning on task-specific benchmarks, while subsequent models like LLaVA (Liu et al., 2023) advanced this approach by leveraging synthetic instruction-following data. Frontier models exhibit strong performance in complex scene interpretation (Zhou et al., 2023). Yet, because these architectures leverage joint vision-language spaces, evaluating whether their success arises from robust spatial perception or powerful language priors remains an ongoing methodological challenge. This reliance on semantic priors can overshadow a critical need to scrutinize their fundamental, low-level perception and spatial grounding skills.

2.2 Evaluating Fine-Grained Perception

While MLLMs excel at global image understanding, they often struggle with fine-grained tasks requiring precise recognition, localization, and data extraction (Li et al., 2024; Zhou et al., 2025). Existing benchmarks designed to probe these limitations face primary methodological challenges. In many traditional perception benchmarks, visual assessment is often entangled with reasoning; questions focusing on semantic concepts allow models to rely on language priors rather than pure visual input. Furthermore, to improve evaluation reliability, many benchmarks (such as MME (Fu et al., 2023) and SEED-Bench (Li et al., 2024)) adopted multiple-choice formats drawn from existing Visual Question Answering (VQA) datasets, which raises concerns about data contamination and true zero-shot evaluation.

Other benchmarks have focused on specific spatial reasoning deficits, such as object counting in dense scenes (Amini-Naieni and Zisserman, 2025) or chart comprehension (Masry et al., 2022; Methani et al., 2020). While these datasets have

driven progress in visual data extraction, they largely feature high-contrast, clearly delineated targets.

2.3 Visual Noise and Character Recognition

The capacity of MLLMs to comprehend abstract or low-signal visual data, such as reading text overlaid on complex backgrounds remains a critical frontier. While traditional Optical Character Recognition (OCR) (Wang et al., 2023) systems are highly specialized, frontier MLLMs are increasingly expected to perform zero-shot text extraction in the wild. Our Watermark Benchmark Dataset departs from existing evaluation methodologies by explicitly disentangling perception from semantics. By utilizing numerical watermarks with modulated transparency (α) across cluttered COCO backgrounds, we remove any possible semantic cues; the model must perceive the underlying patterns directly. This provides a reproducible benchmark to identify the exact breakage thresholds of multimodal perception, testing how deeply frontier models can ground themselves spatially to extract low-signal numeric data.

3 The Watermark Benchmark Dataset

A key contribution of this work is the curation and release of a specialized evaluation dataset designed to probe the limits of fine-grained visual reasoning and Optical Character Recognition (OCR) in frontier Vision-Language Models (VLMs). While standard benchmarks focus on high-contrast, legible text, this dataset introduces controlled transparency to identify the exact breakage thresholds of multimodal perception.

3.1 Composition and Diverse Backgrounds

To ensure the dataset reflects the complexity of natural scenes, background images were sampled from the COCO (Common Objects in Context) (Lin et al., 2014) dataset. These images encompass a diverse variety of textures, lighting conditions, and cluttered environments. Synthetic watermarks comprising random numerical digit sequences were overlaid onto these backgrounds. The digits are colored white, isolating transparency (opacity) as the single independent variable, neutralizing color-contrast bias.

3.2 Fine-Grained Transparency Modulation (α)

The primary feature of this dataset is its systematic modulation of the alpha blending parameter (α). This design allows researchers to evaluate how models transition from confident recognition to perceptual failure as a visual signal degrades.

We generate four distinct subsets across a gradient of visibility (See Fig. 1)

- **High Visibility ($\alpha = 0.8$):** Serves as the control baseline to establish upper-bound model performance.
- **Medium Visibility ($\alpha = 0.5$):** Mirrors standard, visible watermarking standards.
- **Subtle Visibility ($\alpha = 0.3$):** Designed to test the boundary of standard feature extraction.
- **Low Visibility ($\alpha = 0.2$):** Pushes the boundaries of native multimodal zero-shot capabilities.

By structuring the dataset around these transparent text, we provide a reproducible benchmark for evaluating how deeply frontier models can ground themselves spatially to extract low-signal text.

3.3 Dataset Composition and Structural Properties

To ensure the dataset (Fig. 1) is robust and statistically sound, we standardized the structural properties of the generated images and text overlays.

- **Scale and Splitting:** The dataset contains $2,600$ images designated for training the segmentation baselines. For evaluation, a separate test bank of $1,000$ images was evaluated independently across each of the four transparency levels ($\alpha \in \{0.2, 0.3, 0.5, 0.8\}$) where lower α corresponds to more transparent, totaling $4,000$ test inferences per model.
- **Resolution:** All images are standardized to a resolution of 480×640 pixels in RGB format.
- **Watermark Characteristics:** The watermark vocabulary is strictly numerical, consisting of digits 0–9. To maintain uniformity and control for length-based bias, each watermark contains exactly nine digits. As seen in Fig. 2, digits are uniformly distributed for all the different α.

Figure 1: Visual examples of watermarked images with varying opacity levels.

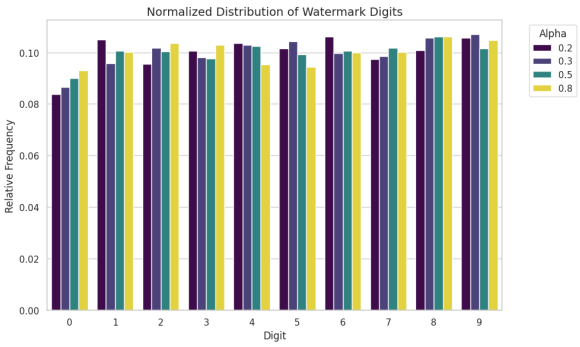

Figure 2: Distribution of digits in the dataset

- **Spatial Grounding:** To isolate the visual degradation caused by transparency from the difficulties of edge-of-frame detection (Chen et al., 2025), all digits are statically positioned at the center of the image.

The richness of the benchmark stems from the background imagery. Sourced from the COCO dataset, these backgrounds capture complex, everyday scenes where diverse objects are depicted in their natural context. The models are therefore subjected to realistic challenges such as background occlusion, varying object scales, cluttered visual gradients, and intricate spatial relationships.

4 Setup

4.1 Baseline: UNet Architecture and Training

To perform watermarking mask generation, we employed a UNet (Ronneberger et al., 2015) architecture characterized by its symmetrical encoder-decoder structure and skip connections (See 2). This design ensures that high-resolution spatial features from the contracting path are preserved and combined with the upsampled output to produce precise segmentation masks, predicting the watermarked text in our case.

- **Inference Pipeline:** During the inference stage, the trained UNet model processes the input image to generate a predicted binary mask. This mask serves as a spatial filter to isolate potential watermark regions. The segmented regions defined by the mask are then passed to an Optical Character Recognition (OCR) engine, which performs the final text prediction and extraction.

- **Data Representation:** The model was trained using pairs of source images and their corresponding binary watermarking masks. An example of this training input is illustrated in Figure 3.

- **Model Configuration:** The encoder follows a progressive series of filter depths (16, 32, 64, and 128), leading to a central bottleneck bridge of 256 channels. Each block in the network consists of two 3×3 convolutional layers followed by Batch Normalization and ReLU activation.

- **Training Parameters:** The model was optimized using the Adam algorithm with a learning rate of 0.001 and a momentum of 0.9. Training was conducted with a batch size of 32 on images resized to 480×640 pixels.

4.2 Vision Language Models Evaluation Setup

To assess the native text extraction and visual reasoning capabilities of frontier Vision-Language Models (VLMs), we conducted a series of direct image-to-text experiments. Each model is provided with the source image and a zero-shot prompt (See Fig. 4) instructing it to identify and extract the subtly embedded numbers within the image.

We evaluated the benchmark against a diverse range of state-of-the-art multimodal architectures:

- **Gemini 3.0 Pro and Flash:** Google's latest frontier models, designed for high-fidelity visual reasoning and multimodal data extraction.

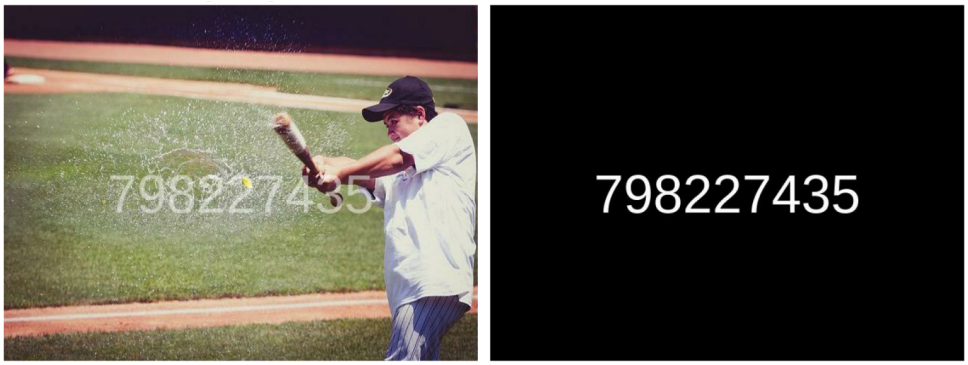

Figure 3: Image and corresponding mask used to train UNet network

Zero-Shot Prompt

You are an OCR.
Extract the 9-digit number subtly embedded in the foreground of the image.
Return the numbers between the tag `<number>` and `</number>`.
If you are not able to find any number return `<number>None</number>`.

Figure 4: The exact zero-shot instruction prompt provided to all Vision-Language Models during evaluation.

- **Gemma 3 27B:** A high-capacity open-weights model (Team et al., 2025) utilized to evaluate the performance of multimodal understanding in a resource-efficient framework.

- **Claude 4.5 Sonnet:** Anthropic's advanced multimodal model, included to provide a comparative baseline for cross-architecture robustness in transcribing text from complex visual contexts.

4.3 Evaluation Metrics

To rigorously assess the performance of both the segmentation and the subsequent text extraction by various Vision-Language Models (VLMs), we employed two primary metrics. These metrics evaluate the model's ability to correctly identify and transcribe the numeric sequences embedded within the images.

- **Exact Match (EM):** This is a strict metric that requires the predicted string to be identical to the ground-truth string. A score of 1 is assigned if the strings match perfectly (ignoring case), and 0 otherwise. This metric is particularly useful for assessing the reliability of the system in high-precision scenarios.

- **Character Error Rate (CER):** The CER provides a more granular view of the OCR performance by calculating the Levenshtein distance (the number of insertions, deletions, and substitutions required) between the predicted text and the ground truth, normalized by the length of the ground truth (Neudecker et al., 2021).

During evaluation, these metrics were tracked across varying levels of transparency (α) to determine the breakdown point for each model's visual reasoning capabilities.

5 Results

We evaluate our proposed UNet+OCR pipeline against four state-of-the-art Vision-Language Models (VLMs) across standard and cropped image settings. The primary metrics tracked are Exact Match Accuracy (Acc.) and Character Error Rate (CER) over a transparency gradient ($\alpha \in \{0.2, 0.3, 0.5, 0.8\}$). The results are summarized in Table 1.

5.1 Overall Watermark Extraction Accuracy

As transparency increases (lower α), we observe a non-linear decay in accuracy across all tested models. At the control baseline of $\alpha = 0.8$, the Gemini family of models achieves near-perfect performance, with Gemini Flash 3.0 topping zero-shot accuracy at 0.955. However, performance deteriorates rapidly as α scales down to 0.2.

Our proposed UNet (Ours) segmentation pipeline significantly outperforms all zero-shot frontier models under standard image settings. At the most difficult visibility threshold ($\alpha = 0.2$ and $\alpha = 0.3$), UNet maintains an accuracy which is

Table 1: Comparison of Accuracy and CER across models at varying transparency levels (α).

		Sonnet 4.5		Flash 3.0		Pro 3.0		Gemma 3		UNet (Ours)	
Image	α	**Acc.**	**CER**	**Acc.**	**CER**	**Acc.**	**CER**	**Acc.**	**CER**	**Acc.**	**CER**
Standard	0.2	0.040	0.732	0.233	0.324	0.229	0.326	0.062	0.648	**0.478**	**0.210**
	0.3	0.104	0.546	0.446	0.165	0.441	0.173	0.141	0.480	**0.792**	**0.053**
	0.5	0.329	0.261	0.776	0.046	0.765	0.049	0.385	0.215	**0.834**	**0.043**
	0.8	0.672	0.075	**0.955**	**0.009**	0.952	0.009	0.768	0.050	0.948	0.014
Cropped	0.2	0.030	0.771	0.289	0.294	0.284	0.291	0.062	0.647	—	—
	0.3	0.088	0.581	0.480	0.149	0.493	0.144	0.140	0.479	—	—
	0.5	0.315	0.282	0.797	0.040	0.813	0.036	0.387	0.214	—	—
	0.8	0.628	0.095	0.951	0.010	0.947	0.010	0.766	0.050	—	—

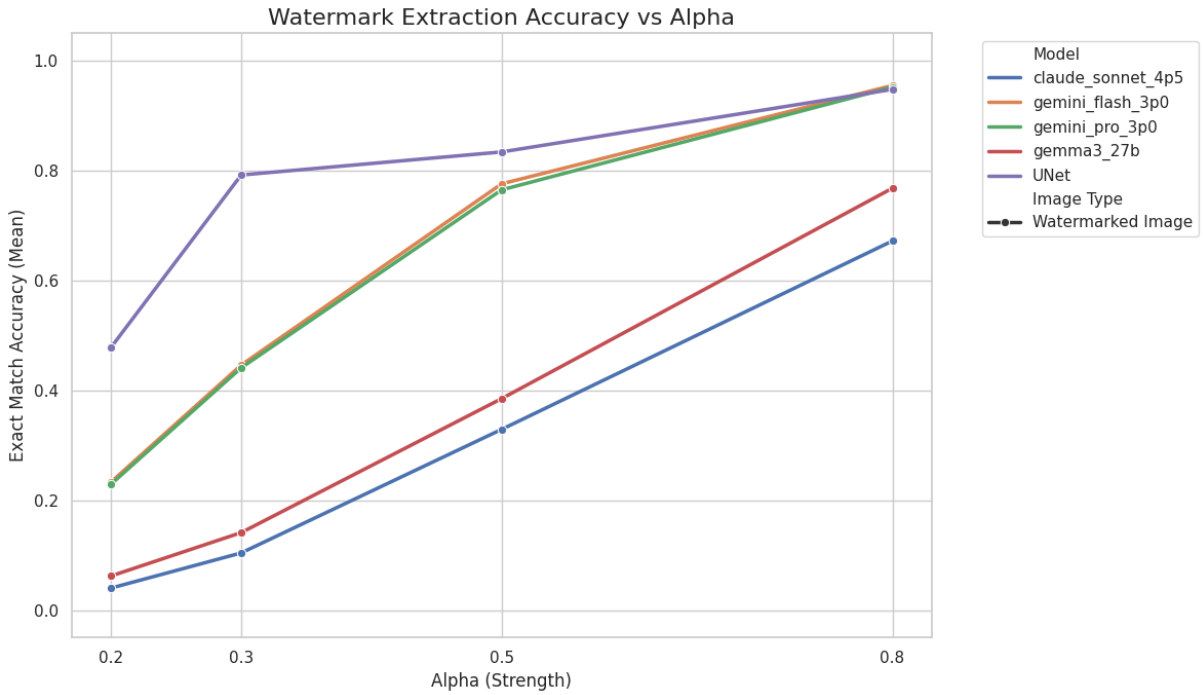

Figure 5: Accuracy detection at different alpha

double that of best performing VLM models. This validates our hypothesis that global attention mechanisms of current general-purpose VLMs fails to detect details when the visual signals are sparse.

5.2 Character-Level Performance and Error Rate

While accuracy measures perfect string matches, the Character Error Rate (CER) (Fig. 6 gives a more granular view of localized failures. We observe that models often perceive the presence of a watermark but struggle to transcribe all nine digits without substitutions or deletions.

Gemma 3 27B suffers the highest degradation in character fidelity, yielding a CER of 0.648 at $\alpha = 0.2$ on standard imagery. In contrast, the UNet pipeline holds a low CER of 0.210 under identical conditions. The tracking of $(1 - \text{CER})$

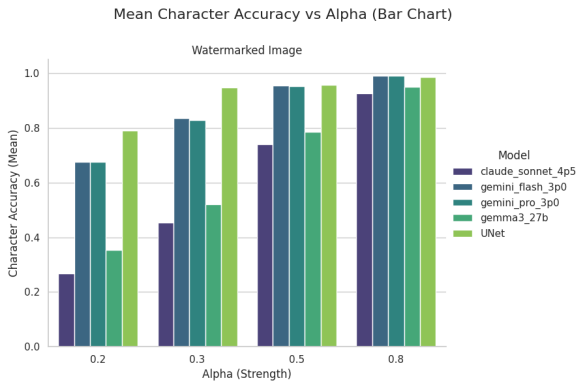

Figure 6: Mean Character Accuracy as a function of Watermark Strength (Alpha). Higher Alpha values indicate more visible watermarks, leading to significantly higher extraction accuracy across all tested vision-language models.

as a proxy for character-level accuracy reveals that while zero-shot models fragment on long numerical strings, segmentation-coupled OCR reliably preserves token-by-token alignment even against high-clutter COCO backgrounds.

5.3 Impact of Image Cropping on Detection Success

To improve the attention signals which are driving force of Vision transformer, we cropped the image which only captures the region containing the numeric text (see B). Comparing the Standard and Cropped partitions in Table 1 reveals a distinct architectural behavioral shift. Isolating the watermark via localized bounding box improves the zero-shot extraction.

For instance, Gemini Flash 3.0 jumps from 0.233 Accuracy at $\alpha = 0.2$ on standard imagery to 0.289 when restricted to cropped dimensions. The improvements are also observed at $\alpha = 0.3\&0.5$. No strong improvements were observed for images with $\alpha = 0.8$, which is primarily due to the fact that text is already very clear and visible.

5.4 Digit-Specific Confusion Analysis

To understand the precise failure modes of frontier VLMs, we analyze the confusion matrix of prediction mistakes for Gemini 3 Pro at $\alpha = 0.3$ (Fig. 7). To ensure a direct character-to-character comparison, this analysis is restricted to watermarked images where the predicted string length matches the ground truth exactly.

The visual data reveals several structural insights into model perception under high transparency:

- High-Frequency Morphological Confusion: The most significant error mode involves true digit 9 being misidentified as 0 (52 instances) and true digit 3 being misidentified as 8 (51 instances). These high-frequency errors suggest that subtle closures in curved digits are easily obscured by background textures, leading the model to hallucinate complete ellipsoids or connected loops.

- Vertical and Diagonal Feature Loss: There is a notable density of errors involving the misidentification of true digit 8 as 3 (45 instances) and true digit 6 as 0 (39 instances). Furthermore, true digit 3 is frequently confused with 2 (40 instances) or 5 (38 instances), highlighting a difficulty in resolving the specific orientation of horizontal and diagonal strokes when α is low.

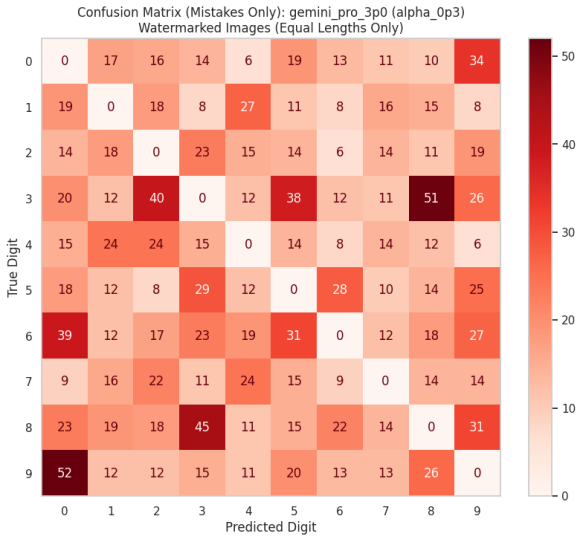

Figure 7: Confusion matrix (Mistakes Only) for Gemini 3 Pro at $\alpha = 0.3$, filtered for equal-length predictions. Rows represent the true digits, while columns represent the predicted digits.

- Structural Simplification and Reconstructive Bias: The model exhibits a tendency to hallucinate non-existent strokes when resolving faint visual signals, frequently misidentifying 1 as 4 (27 instances) and 0 as 9 (34 instances). Rather than failing gracefully by outputting a blank or a partial stroke, the model forces a completion of the character. This suggests a perceptual collapse driven by dataset priors, where the model relies on top-down linguistic and structural biases to fill in the blanks of ambiguous visual inputs, resulting in false positives for structurally similar digits.

6 Conclusion

In this work, we introduced the **FAint numeric Detection Evaluation (FADE)** benchmark to rigorously evaluate the fine-grained visual perception of frontier Multimodal Large Language Models (MLLMs). By systematically modulating the transparency (α) of numerical sequences against cluttered natural backgrounds, FADE provides a unique framework to disentangle pure visual grounding from semantic and linguistic predictability.

Our comprehensive evaluation of state-of-the-art models—including Gemini 3.0, Claude 4.5 Sonnet, and Gemma 3—reveals a significant percep-

tual gap in general-purpose architectures. While these models demonstrate high proficiency in high-visibility contexts, their performance decays non-linearly as transparency increases, often collapsing at $\alpha \leq 0.3$.

A critical finding of our error analysis is the presence of *reconstructive bias*: when faced with low-signal visual inputs, MLLMs tend to hallucinate non-existent strokes to complete digits likely based on dataset priors rather than failing gracefully. Furthermore, our experiments with image cropping demonstrate that reducing the spatial search space provides only modest improvements, suggesting that the bottleneck lies in the visual encoder's inability to register sparse signals rather than a failure of global attention.

As MLLMs are increasingly integrated into safety-critical domains such as content moderation and autonomous agents, addressing these low-level perceptual blind spots is important. We hope that FADE serves as a standard diagnostic tool for the community to facilitate the development of more resilient, spatially aware visual encoders capable of genuine fine-grained visual reasoning.

7 Future Directions

7.1 Analysis of Visual Token Perturbation

A critical next step is investigating the representational shift within the latent space. Future work should quantify how drastically the output embeddings of the Vision Transformer (ViT) (Dosovitskiy et al., 2020) differ between pristine backgrounds and their watermarked counterparts. Understanding the magnitude and nature of this token-level perturbation, particularly at low transparency (α) thresholds, will help isolate whether the primary failure mode is the visual encoder failing to register the weak signal or the language model failing to interpret it.

7.2 Task-Aware Visual Encoding

Current VLM architectures typically utilize a ViT to encode images into a static sequence of tokens, which are subsequently processed by the language model. While this query-agnostic approach is efficient, the generated visual tokens remain identical regardless of the user's prompt. This limits the encoder's ability to dynamically focus on task-relevant, low-contrast features. Recent research into task-aware or query-conditioned visual encoders (Ganz et al., 2024) presents a promising alternative. Applying such dynamic architectures to this benchmark would allow the visual encoder to actively search for and amplify fine-grained, α-transparent signals based on the specific extraction prompt, potentially bridging the perceptual gap observed in our zero-shot evaluations.

Limitations

While the FADE dataset provides a rigorous framework for evaluating visual perception, it currently has scope constraints. First, the watermark vocabulary is restricted to numerical digits. Incorporating full alphanumeric characters and symbols would provide a broader range of morphological complexities. Second, all text is statically positioned at the center of the image to isolate the variable of transparency. Shifting the spatial distribution of text such as placing watermarks near margins would determine how VLMs handle positional embeddings and edge-of-frame detection challenges.

Generative AI Usage

All study design, literature review, synthesis, and writing were conducted by the authors. Generative AI tools (Gemini) were used only for grammar checking and proofreading during the final polishing of the manuscript. No generative AI system was used to generate content, interpret prior work, or draw conclusions. The authors reviewed and approved all final text and remain fully responsible for the content of the paper.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Niki Amini-Naieni and Andrew Zisserman. 2025. Countgd++: Generalized prompting for open-world counting. *arXiv preprint arXiv:2512.23351*.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. 2025. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. *arXiv preprint arXiv:2503.01773*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words:

Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. 2024. Question aware vision transformer for multimodal reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13861–13871.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. Seedbench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

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

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pages 1527–1536.

Clemens Neudecker, Konstantin Baierer, Mike Gerber, Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. 2021. A survey of ocr evaluation tools and metrics. In *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing*, pages 13–18.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. 2021. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8802–8812.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9568–9578.

Andrés Villa, Juan Léon, Alvaro Soto, and Bernard Ghanem. 2025. Behind the magic, merlim: Multi-modal evaluation benchmark for large image-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 492–502.

Bing Wang, Shengsheng Wang, Changchun Li, Renchu Guan, and Ximing Li. 2024. Harmfully manipulated images matter in multimodal misinformation detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2262–2271.

Tong Wang, Ting Liu, Xiaochao Qu, Chengjing Wu, Luoqi Liu, and Xiaolin Hu. 2025. Glyphmastero: A glyph encoder for high-fidelity scene text editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28523–28532.

Xiao-Feng Wang, Zhi-Huang He, Kai Wang, Yi-Fan Wang, Le Zou, and Zhi-Ze Wu. 2023. A survey of text detection and recognition algorithms based on deep learning technology. *Neurocomputing*, 556:126702.

Chenyue Zhou, Mingxuan Wang, Yanbiao Ma, Chenxu Wu, Wanyi Chen, Zhe Qian, Xinyu Liu, Yiwei Zhang, Junhao Wang, Hengbo Xu, and 1 others. 2025. From perception to cognition: A survey of vision-language interactive reasoning in multimodal large language models. *arXiv preprint arXiv:2509.25373*.

Kankan Zhou, Eason Lai, Wei Bin Au Yeong, Kyriakos Mouratidis, and Jing Jiang. 2023. Rome: Evaluating pre-trained vision-language models on reasoning beyond visual common sense. In *Findings of the*

Association for Computational Linguistics: EMNLP 2023, pages 10185–10197.

Fengbin Zhu, Ziyang Liu, Xiang Yao Ng, Haohui Wu, Wenjie Wang, Fuli Feng, Chao Wang, Huanbo Luan, and Tat Seng Chua. 2024. Mmdocbench: Benchmarking large vision-language models for fine-grained visual document understanding. *arXiv preprint arXiv:2410.21311*.

A UNet configurations

A.1 Detailed Network Architecture

The specific layer-by-layer configuration of the UNet model is detailed in Table 2. The final output layer utilizes a 1×1 convolution with a Sigmoid activation function to generate a probability map for the binary mask.

Parameter	**Value**
Encoder Filter Sequence	16, 32, 64, 128
Bottleneck Bridge Channels	256
Decoder Filter Sequence	128, 64, 32, 16
Convolutional Kernel Size	3×3
Hidden Layer Activation	ReLU
Output Layer Activation	Sigmoid
Dropout Probability	0.1
Optimizer	Adam
Learning Rate	0.001

Table 2: Summary of UNet hyperparameters and architectural constants.

A.2 Data Augmentation Strategy

To increase dataset diversity and model generalization, the following stochastic augmentations were applied during the training phase:

- **Spatial Transformations:** Random horizontal and vertical flipping (p=0.5) and random rotations within a range of ± 15 degrees.
- **Scaling:** Random zooming with a scale factor between 0.8 and 1.2.
- **Intensity Adjustments:** Random contrast scaling applied between 0.5 and 1.5 to account for varying lighting conditions in the source images.

B Cropping algorithm

For each instance, we utilized the ground-truth binary mask to delineate the superimposed numerical text. We calculated a bounding box from the mask's top-left and bottom-right coordinates, applied a uniform 15-pixel padding to these boundaries, and used the resulting expanded region to crop the image.