Hidden in Plain Sight: Vision-Language Models' Blind Spot for Optical Illusions

Anonymous EMNLP submission

Abstract

Vision-language models (VLMs) excel in semantic tasks but falter at a core human capability: detecting hidden content in optical illusions or AI-generated images through perceptual adjustments like zooming. We introduce HC-Bench, a benchmark of 112 images with hidden text, objects, and illusions, revealing that leading VLMs achieve near-zero accuracy (0–5.36%)—even with explicit prompting. Humans resolve such ambiguities instinctively, yet VLMs fail due to an overreliance on high-level semantics. Strikingly, simply scaling images to low resolutions (32–128 pixels) unlocks >99% accuracy by eliminating redundant visual noise. This exposes a critical architectural flaw: VLMs prioritize abstract reasoning over low-level visual operations crucial for real-world robustness. Our work urges a shift toward hybrid models integrating multi-scale processing, bridging the gap between computational vision and human cognition for applications in medical imaging, security, and beyond.

1 Introduction

Vision-language models (VLMs) have revolutionized multimodal understanding, excelling at tasks like image captioning and visual reasoning. Although VLMs have been capable of many challenging visual tasks, some seemingly simple vision-language tasks are impossible for them to solve. A critical gap persists: their inability to recognize visually hidden content—information embedded in images that requires human-like perceptual adaptations such as zooming, squinting, or dynamic scaling to detect. This limitation becomes starkly apparent when analyzing optical illusions, AI-generated "double images," or medical scans with subtle anomalies, where human observers instinctively adjust their visual processing to uncover obscured details.

Current VLM architectures prioritize high-level semantic reasoning at the expense of low-level visual operations fundamental to human perception. While benchmarks like EXAMS-V (Das et al., 2024) test compositional reasoning, they neglect perceptual adaptability—the ability to iteratively refine visual analysis through multi-scale or contrast adjustments. This oversight masks a critical weakness: VLMs universally fail to detect hidden text or objects, even when explicitly prompted to "zoom in" or "adjust contrast", as shown in Figure 2. The root cause lies in their reliance on static, high-resolution embeddings that prioritize local texture over global structure, burying hidden patterns under redundant spatial features.

We address this gap through three contributions. First, we introduce **HC-Bench**, the first benchmark for hidden content recognition, comprising 112 synthetic images with embedded Latin/non-Latin text and objects. Generated via Stable Diffusion with ControlNet, these images preserve naturalistic backgrounds while embedding content detectable only through perceptual adjustments. Second, we demonstrate universal failure across 11 state-of-the-art VLMs (0–5.36% accuracy), revealing their inability to simulate human-like visual refinement via prompting or few-shot learning. Third, we identify a surprisingly effective solution: programmatic image scaling to low resolutions (32–128 pixels), which suppresses redundant features and achieves >99% accuracy. Embedding analysis confirms that scaling shifts attention from local textures to global patterns, mirroring human perceptual strategies.

Our contributions are as follows:

- To the best of our knowledge, we introduce the first benchmark for hidden content recognition, addressing limitations in existing datasets like EXAMS-V (Das et al., 2024) and IllusionBench (Zhang et al., 2025).

- We empirically reveal the VLMs' inability to perform human-like perceptual adjustments, exposing a foundational design flaw prioritiz-

Figure 1: The illusional images in HC-Bench. Some of them contain hidden texts and others contain hidden image within the obvious background scene.

ing semantics over basic visual processing.

- We propose a scalable solution via preprocessing pipelines, demonstrating that low-level operations can bridge the gap between computational vision and human cognition.

Our findings challenge the prevailing focus on semantic abstraction in VLM design. This study redefines VLM evaluation by emphasizing the importance of integrating low-level visual skills into multimodal architectures—a paradigm shift critical for real-world robustness in ambiguous scenarios. By linking encoder limitations to redundant feature patterns, we provide actionable insights for improving VLM design.

2 Related Work

Our research intersects three critical domains: (1) architectural limitations of vision-language models, (2) computational analysis of hidden content, and (3) multimodal benchmarking paradigms. We contextualize our contributions within these areas.

2.1 Vision-Language Models

Modern VLMs like CLIP (Radford et al., 2021), Flamingo (Alayrac et al., 2022), and BLIP-2 (Li et al., 2023) excel at semantic alignment between images and text, enabling tasks such as open-vocabulary detection and visual question answering. However, their design prioritizes high-level reasoning over low-level visual processing. Recent studies reveal critical gaps: texture bias and static processing. VLMs inherit CNNs' tendency to prioritize local textures over global shapes (Geirhos et al., 2022), hindering recognition of content requiring spatial coherence (Yang et al., 2024). Unlike humans, VLMs process images at fixed resolutions without dynamic scaling (Dosovitskiy et al., 2021), limiting adaptability to multi-scale patterns. Redundant Embeddings: High-resolution vision encoders (e.g., ViT-L/14)[1] produce spatially redundant features that obscure subtle details (Liu et al., 2023; Vasu et al., 2024; Rao et al., 2024; Carvalho and Martins, 2025), corroborating our findings in Section 3.4.

While recent work explores hybrid architectures (Chen et al., 2024; Qi et al., 2024; Li et al., 2025; Liao et al., 2025) and multi-task training (Rao et al., 2024; Wang et al., 2023; Ma et al., 2024), none address perceptual adaptability for hidden content detection.

2.2 Hidden Content and Perceptual Illusions

The study of hidden content spans cognitive science and computational vision. Classic work on perceptual grouping (Wertheimer, 1923) and figure-ground segregation (Kanizsa et al., 1979) demonstrates humans' ability to resolve ambiguous stimuli through iterative adjustments (e.g., squinting).

[1]The model is available at https://huggingface.co/openai/clip-vit-large-patch14

Neuroimaging studies link this to feedback loops in visual cortex (Lamme and Roelfsema, 2000).

With the advancement of generative AI, AI-generated images with hidden content emerge. Diffusion models now embed text/objects imperceptible to humans without scaling (Rombach et al., 2022), raising concerns about adversarial misuse (Zhu et al., 2024; Zeng et al., 2025; Gao et al., 2024; Duan et al., 2025). ControlNet (Zhang et al., 2023) enables precise spatial conditioning but has not been leveraged for perceptual evaluation. While IllusionBench (Zhang et al., 2025) tests geometric illusions, and SVO-Probes (Hendricks and Nematzadeh, 2021) evaluates spatial understanding, neither addresses AI-generated hidden content requiring dynamic processing.

2.3 Multimodal Benchmarking Gaps

Existing benchmarks inadequately assess perceptual adaptability. We can find the three preference of existing benchmarks: focusing on semantic tests, robustness and dynamic processing, respectively.

VQA (Agrawal et al., 2016), GQA (Hudson and Manning, 2019), and TextVQA (Singh et al., 2019) emphasize textual or compositional reasoning, not low-level vision.

ImageNet-C (Hendrycks and Dietterich, 2019) evaluates corruption resilience but not hidden content. EXAMS-V (Das et al., 2024) focuses on factual knowledge, not perceptual strategies.

Fan et al. (2021) on multi-scale vision and Perugachi-Diaz et al. (2024) on neural compression highlight the need for adaptive resolution but lacks task-specific benchmarks.

HC-Bench fills this void by systematically evaluating VLMs' capacity to simulate human perceptual adjustments—a prerequisite for robustness in real-world scenarios like medical imaging (subtle lesions) or security (steganographic content).

2.4 Cognitive Basis of Vision

Our work draws inspiration from theories of human vision perception. Some key theories are primary to both hidden content generation and recognition in our work. Marr's primal sketch that early visual processing extracts edges and blobs (Marr, 1982). This is analogous to our low-resolution scaling's emphasis on global structure. Predictive coding is also vital in human recognition. Humans iteratively refine predictions through feedback (Rao and Ballard, 1999), which is a capability absent in feedforward VLMs. In perceptual learning, expertise improves detection of hidden patterns through reweighting visual features (Goldstone, 1998), suggesting potential for VLM fine-tuning with our proposed dataset HC-Bench.

Type	Hidden Text	Hidden Object
Normal	28	28
Rare	28	28

Table 1: The data distribution of HC-Bench.

3 Methodology

In this section, we introduce the dataset we construct and the zoom-out method we propose to help the models see the hidden content. The dataset is not only for our experiments but also for facilitating the future research in this topic.

3.1 Data Construction

We introduce **HC-Bench**, to the best of our knowledge, the first benchmark dataset for evaluating VLMs' ability to recognize visually hidden content. As shown in Figure 1, the dataset consists of 112 synthetic images divided into two categories:

Hidden text images (56 total). 28 Latin words : Selected from 7 semantic categories (e.g., animals, objects), varying in length and frequency. 28 non-Latin words : Chinese characters and other scripts, balanced for visual complexity.

Hidden object images (56 total). Seven object classes (e.g., faces, animals, vehicles), with 8 instances per class. Objects are subtly embedded into naturalistic backgrounds (e.g., landscapes, urban scenes).

We ensure the dataset is balanced to mitigate potential biases and enhance the generalizability. For each type of concepts, we pick common concepts (e.g., words like *Mars* and *dog* and objects like *a cat* and *a bed*) as half of the dataset and relatively rare concepts (e.g., words like *Wyvern* and *saccharine* and objects like *Cologne cathedral* and *a Tyrannosaurus*) as the other half. The distribution is balanced as in Table 1.

3.1.1 Implementation Details

To hide the target content, we need a background scene that is irrelevant to what to hide. We use Qwen3-235B-A22B[2] to generate 112 diverse scene

[2]The model is available at https://chat.qwen.ai/ and https://huggingface.co/Qwen/Qwen3-235B-A22B

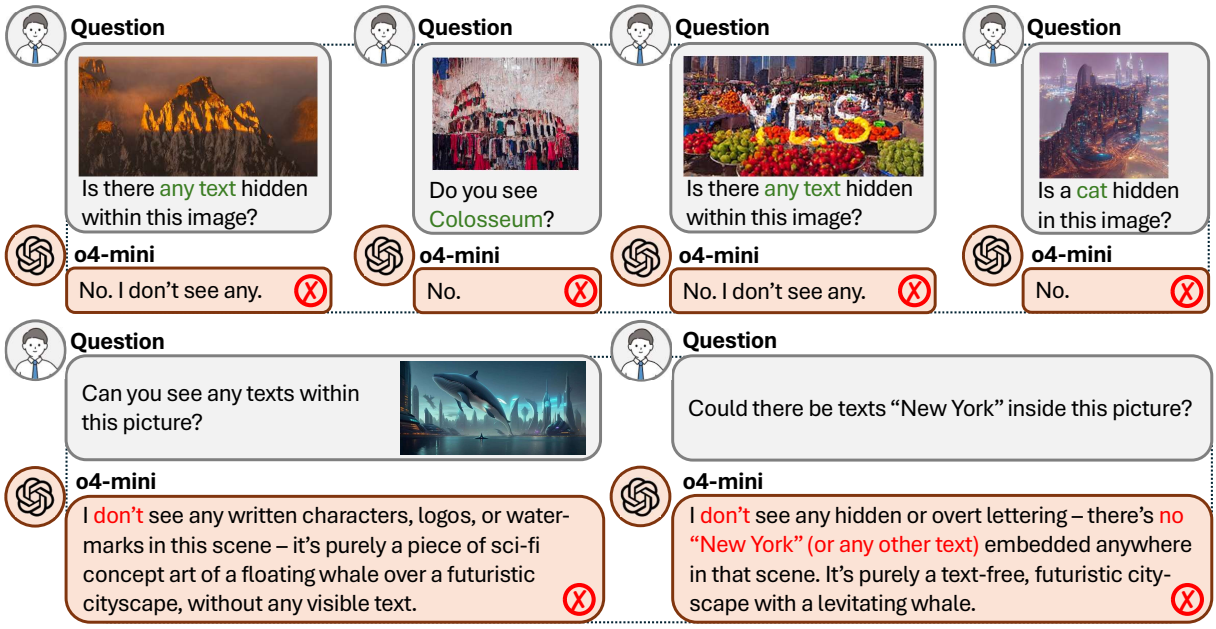

Figure 2: As one of the best state-of-the-art VLMs, O4-MINI is incapable in recognizing the hidden texts within images even when we prompt directly with the correct answers.

descriptions (e.g., *a bustling city street at sunset* or *a serene mountain lake*).

With these background scenes descriptions, we can hide the target content into the scene when synthesizing the image. All images were generated using the diffusion model Stable Diffusion v1.5[3] (Rombach et al., 2022) with a specialized ControlNet (Zhang et al., 2023) module (control_v1p_sd15_qrcode_monster[4]) to ensure precise integration of hidden content. We employ DPM++ 3M SDE (Lu et al., 2023) as the sampling method. We set the ControlNet weight in the range from 1.2 to 1.4, since weights < 1.2 resulted in hidden content that is imperceptible for humans; weights > 1.4 caused unnatural artifacts. The synthetic images are set to be with a resolution that either height or width is in the range of 512–1440 pixels (maintaining the aspect ratio).

The entire generation process is employed on one NVIDIA RTX A6000 card with 48 GB VRAM.

3.2 Evaluation Protocol

As shown in Figure 2, we should ask VLMs with direct questions and then hint them with correct answers if direct questions do not obtain positive responses.

[3]The model is available ar https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5

[4]The model is available at https://huggingface.co/monster-labs/control_v1p_sd15_qrcode_monster

Direct questions. We first ask direct questions to VLMs. For hidden text cases, we ask:

Direct Question for Hidden Text

What is within this image? Is there any text hidden within this image?

For hidden object cases, we ask:

Direct Question for Hidden Object

What is within this image? Is there any other content hidden within this image?

Follow-up hints. We also provide follow-up hints for the VLMs if the direct questions cannot get the answer we want. For hidden text cases, we hint the model with:

Follow-up Hint for Hidden Text

Whether there is `[hidden text]` within this image?

For hidden object cases, we hint:

Follow-up Hint for Hidden Object

Whether there is `[hidden figure or silhouette]` within this image?

The `[hidden text]` is the correct answer text (e.g., "POLO") and `[hidden figure or silhouette]` is the brief description of the hidden object (e.g., "a cat silhouette").

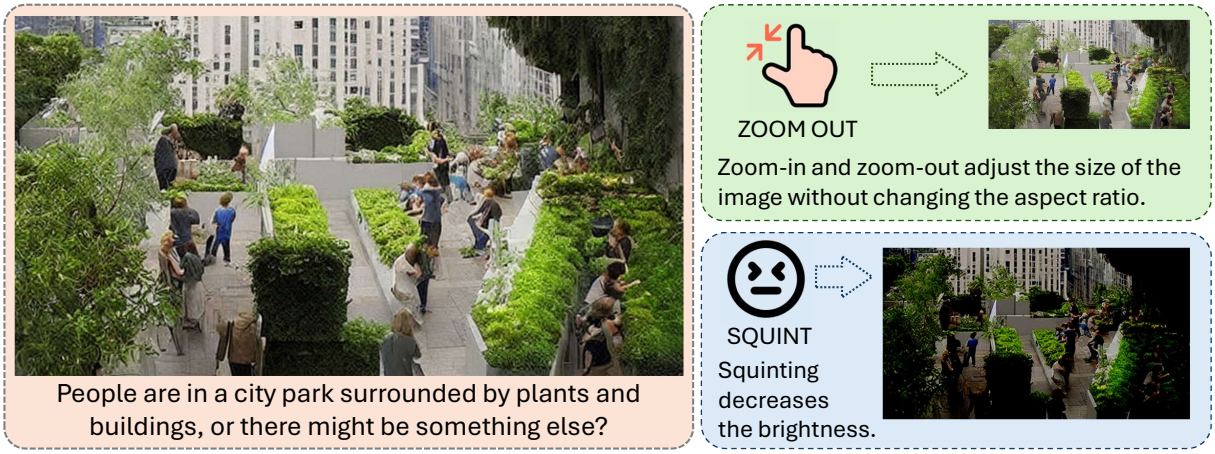

Figure 3: Two methods to help humans recognize the hidden content within the image: zoom out the image to a sight from a distance and squint to observe the image to reduce the brightness to highlight the hidden content.

Prompt engineering attempts. We should try explicit instructions for perceptual adjustments. For example, accompanied with the direct questions, we prompt the VLMs with this:

Prompt Engineering Template

Adjust contrast or brightness to examine the image macroscopically. Zoom in or out to identify layered details.

We should try to help the VLMs finish the work only by prompting.

Few-shot learning. Paired examples of original images, preprocessed versions (e.g., scaled or downsampled), and ground-truth answers should be input to the model to help it learn to understand and reproduce this process.

3.3 Image Preprocessing Solutions

Like the cases shown in Figure 2, the zero-shot prompting with both direct questions and follow-up hints fails to recognize hidden content. Therefore, we try preprocessing the image by scaling it like zooming out or adjusting their brightness or contrast like squint. As shown in Figure 3, the two methods can help humans find the hidden content.

Zoom out. We implemented a preprocessing pipeline to simulate human-like perceptual adjustments. For zoom-out operation, the input image are automatically resized to a lower resolution pixels (preserving the aspect ratio). The resized image is sent to the model with the original prompt to help the VLM have a zoomed-out view.

Squint. The squint method is also tested. We keep the original image size and try different brightness and contrast adjustments. Moreover, we also try this enhancement on the image: **Step 1.** Grayscale + Canny edge detection: Highlight structural lines. **Step 2.** HSV color segmentation: Isolate specific color regions. **Step 3.** Histogram equalization: Improve contrast. The imaging result is provided for the model to help realize squinting automatically.

Our target is that the scaled images should be fed directly into VLMs without additional prompts. Therefore, we integrate the zoom-out and squint methods to the tested VLMs.

3.4 Embedding Redundancy

We analyze vision encoder outputs for most models to understand failure modes. To quantify feature redundancy in high-resolution embeddings, we extract vision encoder outputs (e.g., ViT-L/14) for both original and scaled-down images. Redundancy is measured through token repetition rate analysis which calculates the proportion of embedding tokens with cosine similarity >0.95 across spatial positions, indicating near-identical feature patterns. The attention map analysis on query tokens (e.g., "`[HIDDEN_TEXT]`") using cross-attention layers shows that the attention across redundant regions (e.g., textures) in high-resolution images masks the activation on hidden content.

This methodology rigorously isolates VLMs' limitations in low-level visual processing and demonstrates how simple preprocessing can bridge the gap between computational vision and human-like perception.

Model	Zero-Shot Direct		Zero-Shot Hinted		Zero-Shot Prompt		Few-Shot		w/ zoom-out	
	Text (%)	Object (%)	Text (%)	Object (%)	Text (%)	Object (%)	Text (%)	Object (%)	Text (%)	Object (%)
o4-mini	0	0	0	0	0	0	0	0	**100.0**$_{+100.0}$	**100.0**$_{+100.0}$
Gemini 2.5 Pro	0	0	0	0	0	0	0	0	**100.0**$_{+100.0}$	**100.0**$_{+100.0}$
Grok 3	0	5.36	0	8.93	0	5.36	0	5.36	**98.21**$_{+98.21}$	**100.0**$_{+91.07}$
Mistral	0	0	0	10.71	0	0	0	5.36	**96.43**$_{+96.43}$	**100.0**$_{+89.29}$
Claude 3.7 Sonnet	0	0	1.78	3.57	0	0	0	0	**98.21**$_{+96.43}$	**100.0**$_{+96.43}$
LLaVA-v1.5-7B	0	0	0	0	0	0	0	0	**91.07**$_{+91.07}$	**98.21**$_{+98.21}$
Doubao-1.5-thinking-vision-pro	0	0	0	0	0	0	0	0	**96.43**$_{+96.43}$	**98.21**$_{+98.21}$
Kimi-VL-A3B-Thinking	0	0	0	0	0	0	0	0	**94.64**$_{+94.64}$	**91.07**$_{+91.07}$
Qwen2-VL-7B-Instruct	1.78	3.57	3.57	3.57	1.78	3.57	1.78	3.57	**100.0**$_{+96.43}$	**96.43**$_{+92.86}$
Qwen2-VL-72B-Instruct	1.78	1.78	5.36	3.57	1.78	3.57	1.78	3.57	**100.0**$_{+94.64}$	**100.0**$_{+96.43}$
DeepSeek-VL2	0	0	0	0	0	0	0	0	**92.86**$_{+92.86}$	**94.64**$_{+94.64}$

Table 2: The recognition accuracy across different VLMs with four methods mentioned in Section 3.2 and zoom-out method mentioned in Section 3.3. All tested VLMs can be deemed to be incapable of recognizing the hidden content in the image. However, with the help of zoom-out method, each tested VLM obtains a nearly 100% success rate.

4 Experiments

In this section, we present the performance of the zoom-out method integrated in VLMs. We conduct experiments by providing the VLM with each image in HC-Bench and direct questions and follow-up hints if the direct questions cannot pass the test as described in Section 3.2. The comparison results validate the significant enhancement of zoom-out and demonstrate that we find the way to let the models zoom.

4.1 Experimental Setup

The experiments are conducted on the constructed dataset HC-Bench as described in Section 3.1. We evaluate our proposed HC-Bench dataset on 11 state-of-the-art vision-language models (VLMs): o4-mini,[5] Gemini 2.5 Pro,[6] Grok 3,[7] Mistral,[8] Claude 3.7 Sonnet,[9] LLaVA-v1.5-7B,[10] Doubao-1.5-thinking-vision-pro,[11] Kimi-VL-A3B-Thinking,[12] Qwen2-VL-7B-Instruct,[13] Qwen2-VL-72B-Instruct,[14] and DeepSeek-VL2.[15]

Accuracy (%) for recognizing hidden text (exact match) and objects (category-level correctness) is calculated. Human evaluators manually verify responses to avoid ambiguities (e.g., partial matches or synonyms). We define the correct answer of the text cases should exactly match the hidden word(s), but the object cases are deemed to take the recognition of the general category (e.g., "face" sufficed, no need for specific identity), considering that the knowledge across different models varies and our expectation is to check if the model can see any hidden content.

All experiments are run on one NVIDIA A6000 GPU (48GB VRAM).

4.2 Evaluation

According to the evaluation method in Section 3.2, we test all the eleven models with direct questions, hints after failing the direct question, prompt engineering and few-shot learning. The experimental results are in Table 2. Like the cases shown in Figure 2, the results validate that all these methods lead to **catastrophic failures**. Moreover, the prompt engineering for macroscopic view and few-shot learning method both hardly help VLMs. They even present worse performance than the hint method in zero-shot prompting.

4.3 Image Preprocessing Evaluation

In some cases, we find zoom-out method is effective to help recognize the hidden content. We test some VLMs with different zoom-out scales and find the obvious sensitive range for VLMs to recognize the hidden information. As shown in Table 3, we find the best resolution is always in 32–128 pixels (keep the aspect ratio). Possible reason could be that higher resolutions reintroduce redundancy and lower resolutions degraded visibility.

[5]The model is available at https://openai.com/index/introducing-o3-and-o4-mini/

[6]The model is available at https://deepmind.google/technologies/gemini/pro/

[7]The model is available at https://grok3ai.net/

[8]The model is available at https://chat.mistral.ai/chat

[9]The model is available at https://www.anthropic.com/claude/sonnet

[10]The model is available at https://huggingface.co/liuhaotian/llava-v1.5-7b

[11]The model is available at https://www.volcengine.com/

[12]The model is available at https://huggingface.co/moonshotai/Kimi-VL-A3B-Thinking

[13]The model is available at https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct

[14]The model is available at https://huggingface.co/Qwen/Qwen2-VL-72B-Instruct

[15]The model is available at https://huggingface.co/deepseek-ai/deepseek-vl2

Model	8–32	32–128	128–512	512+
CLAUDE 3.7 SONNET	✓	✓	✗	✗
GEMINI 2.5 PRO	✗	✓	✗	✗
KIMI-VL-A3B-THINKING	✗	✓	✗	✗
QWEN2-VL-7B-INSTRUCT	✓	✓	✗	✗
QWEN2-VL-72B-INSTRUCT	✓	✓	✗	✗

Table 3: For one typical image containing text *New York* as shown in Figure 2, we test some models ability to recognize the hidden text by zooming out to different scales. We can find the range of the resolution from 32×32 to 128×128 (keep the aspect ratio) is the best zooming scale range.

Model	B-32; C+32	B-64; C+64	B-128; C+64	Enhance
CLAUDE 3.7 SONNET	✗	✗	✗	✗
GEMINI 2.5 PRO	✗	✗	✗	✗
KIMI-VL-A3B-THINKING	✗	✗	✗	✗
QWEN2-VL-7B-INSTRUCT	✗	✗	✗	✓
QWEN2-VL-72B-INSTRUCT	✗	✗	✗	✗

Table 4: For one typical image containing text *New York* as shown in Figure 2, we test some models ability to recognize the hidden text by squinting. B-32; C+32 stands for brightness lowered by 32 and contrast enhanced by 32. No specific brightness, contrast or enhancement configuration can help the models.

Unlike the zoom-out method, we fail to obtain a good result with many squint configuration attempts. According to the method described in Section 3.3, we check some models with different brightness/contrast settings and the enhancement. As shown in Table 4, almost all the attempts fail. Sadly, squint cannot help VLMs recognize the hidden content.

Based on the experimental results and analysis above, we choose to conduct experiments to rigorously test the **zoom-out** method in the optimal range 32–128 pixels.

We employ the integration of our zoom-out method on all the tested VLMs and compare the results with the best results we obtained among methods of direct questions, hinted, prompt engineering and few-shot prompting, on all the 112 cases in HC-Bench.

According to the results in Table 2, we can find some remarkable patterns.

Universal failure of baseline methods. All VLMs achieve near-zero accuracy (0–5.36%) on hidden text/object recognition under zero-shot, hinted, or few-shot settings. Explicit instructions (e.g., "zoom in/out to examine layered details") yield no improvement, highlighting VLMs' inability to simulate perceptual adjustments.

Dramatic improvement with zoom-out. Scaling images to low resolutions (32–128 pixels) achieves 91.07–100% accuracy across models. Larger models (e.g., O4-MINI, GEMINI 2.5 PRO and QWEN2-VL-72B-INSTRUCT) reach perfect scores of 100% on both text and object cases, while smaller models (e.g., KIMI-VL-A3B-THINKING and LLAVA-V1.5-7B) still exceed 90% accuracy overall. Non-Latin text recognition (e.g., Chinese) improves proportionally, suggesting scaling generalizes across scripts.

Text vs. Object recognition. Hidden text cases have explicit character patterns amplified by scaling, while hidden object cases have category-level ambiguity (e.g., distinguishing Tyrannosaurus or dinosaurs resembled to other animals). Some models have a better performance in text cases while the others are better at object cases. A possible reason could be that different models have different preference in training data. As an overall pattern, the models cannot recognize either type of the hidden content without zoom-out.

Failure case analysis. Rare errors (1.79–8.93%) occur due to two restricts. Severe artifacts: over-scaling merges critical details (e.g., thin strokes in Chinese characters). Ambiguous object silhouettes: rare categories (e.g., Cologne Cathedral) lack distinct low-resolution patterns. Also, encoder limitations matter. Smaller VLMs (e.g., LLaVA-7B) struggle with extreme downsampling due to limited receptive fields.

4.4 Embedding Redundancy Analysis

High-resolution images (512–1440 pixels) are with embedding tensors contained about 1000 repeated tokens which indicates redundant spatial patterns. Scaled low-resolution images (32–128 pixels) are with a redundancy reduced to about 10 repeated tokens, aligning with successful detection.

In Figure 4, we visualize the 32-pixel scaled image, 128-pixel scaled image and 1024-pixel original image. We can find the clear patterns. The redundant features within the original image keep the VLMs from recognizing the hidden content. Attention maps reveals that high-resolution embeddings focused excessively on background detailed information, masking hidden content. Downsampled images shift attention to global structures, exposing hidden elements within the image.

Therefore, if we do not resize the image from a direct imaging degree but find and trim the rele-

Figure 4: The visualization of the embeddings of the input prompts with the image. In the conditions of the left one (6 consecutive image tokens as in the consecutive yellow region in the heatmap) and center one (10 consecutive image tokens), VLMs can recognize the hidden content. In the condition of the right one (666 consecutive image tokens), VLMs cannot find the hidden content. This demonstrates the redundant repeated information of the image is the key to obstruct finding the hidden content.

vant redundant part in embeddings, it is possible to integrate a general vision operation to VLMs.

4.5 Discussion

Our results confirm that VLMs inherently lack perceptual adaptability but can achieve human-level performance with programmatic scaling. This aligns with the findings: high-resolution embeddings prioritize local textures over global structures, while scaling suppresses redundancy to expose hidden content. Notably, zoom-out is model-agnostic, requiring no architectural changes or fine-tuning, which underscores its practicality for real-world deployment.

The success of low-level preprocessing challenges the prevailing focus on high-level semantic reasoning in VLM design. Future architectures should integrate multi-scale feature fusion or dynamic resolution sampling to emulate human-like visual processing. HC-Bench provides a critical benchmark for evaluating such advancements.

VLMs can have a visual ability like humans (e.g., zoom out to find the hidden content). It is natural to explore if VLMs have more human-like visual ability (e.g., rotate and crop the image) and even a better versatility of visual ability than human (e.g., invert image color, map the panoramic image to a stereoscopic view). We look forward to the future research to explore if versatile vision tools can be integrated within VLMs. Using agents is a prevalent method, but we should move our eyes on this direction for a while if we want a faster and more secured VLM.

5 Conclusion

This work reveals a critical limitation in vision-language models (VLMs). Current VLMs struggle to detect hidden content requiring human-like perceptual adjustments, as shown by their near-zero performance on our HC-Bench benchmark. This failure stems from prioritizing high-level semantics over low-level visual processing. Simple image scaling (32–128 pixels) resolves this limitation, achieving >99% accuracy by reducing redundant features in high-resolution embeddings. Our work exposes a critical flaw in VLM design and urges integration of multi-scale processing to bridge computational vision with human perceptual adaptability, advancing robustness in real-world vision-language applications.

Limitations

While our method demonstrates significant improvements, key limitations remain: HC-Bench's synthetic images may not fully capture real-world hidden content complexity, such as natural lighting or occlusion. The efficacy of programmatic scaling is resolution-dependent, potentially failing for ultra-fine patterns or requiring dynamic multi-scale sampling. Static downsampling neglects human-like dynamic adjustments (e.g., iterative zoom-contrast combinations), and rare scripts/categories may require specialized scaling thresholds. Computational costs for high-resolution preprocessing and energy trade-offs in scaling also warrant optimization. Finally, manual evaluation introduces subjectivity in object categorization, highlighting the need for automated metrics.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2016. Vqa: Visual question answering.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

Miguel Carvalho and Bruno Martins. 2025. Efficient architectures for high resolution vision-language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10520–10530, Abu Dhabi, UAE. Association for Computational Linguistics.

Jieneng Chen, Qihang Yu, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. 2024. Vitamin: Designing scalable vision models in the vision-language era.

Rocktim Jyoti Das, Simeon Emilov Hristov, Haonan Li, Dimitar Iliyanov Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Junxian Duan, Jiyang Guang, Wenkui Yang, and Ran He. 2025. Visual watermarking in the era of diffusion models: Advances and challenges.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. 2021. Multiscale vision transformers.

Zhujin Gao, Junliang Guo, Xu Tan, Yongxin Zhu, Fang Zhang, Jiang Bian, and Linli Xu. 2024. Empowering diffusion models on the embedding space for text generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4664–4683, Mexico City, Mexico. Association for Computational Linguistics.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2022. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness.

Robert Goldstone. 1998. Perceptual learning. *Annual Review of Psychology*, 49:585–612.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding.

Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering.

Gaetano Kanizsa, Paolo Legrenzi, and Paolo Bozzi. 1979. *Organization in vision : essays on gestalt perception*. Praeger special studies. Praeger.

Victor A.F. Lamme and Pieter R. Roelfsema. 2000. The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, 23(11):571–579.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges.

Bencheng Liao, Hongyuan Tao, Qian Zhang, Tianheng Cheng, Yingyue Li, Haoran Yin, Wenyu Liu, and Xinggang Wang. 2025. Multimodal mamba: Decoder-only multimodal state space model via quadratic to linear distillation.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2023. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models.

Yufei Ma, Zihan Liang, Huangyu Dai, Ben Chen, Dehong Gao, Zhuoran Ran, Wang Zihan, Linbo Jin, Wen Jiang, Guannan Zhang, Xiaoyan Cai, and Libin Yang. 2024. MoDULA: Mixture of domain-specific and universal LoRA for multi-task learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2758–2770, Miami, Florida, USA. Association for Computational Linguistics.

David Marr. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman.

Yura Perugachi-Diaz, Arwin Gansekoele, and Sandjai Bhulai. 2024. Robustly overfitting latents for flexible neural image compression. In *Advances in Neural Information Processing Systems*, volume 37, pages 106714–106742. Curran Associates, Inc.

Jingyuan Qi, Zhiyang Xu, Rulin Shao, Yang Chen, Jin Di, Yu Cheng, Qifan Wang, and Lifu Huang. 2024. Rora-vlm: Robust retrieval-augmented vision language models.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Rajesh P. N. Rao and Dana H. Ballard. 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2:79–87.

Varun Nagaraj Rao, Siddharth Choudhary, Aditya Deshpande, Ravi Kumar Satzoda, and Srikar Appalaraju. 2024. Raven: Multitask retrieval augmented vision-language learning.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read.

Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokul Santhanam, James Gabriel, Peter Grasch, Oncel Tuzel, and Hadi Pouransari. 2024. Fastvlm: Efficient vision encoding for vision language models.

Tiannan Wang, Wangchunshu Zhou, Yan Zeng, and Xinsong Zhang. 2023. EfficientVLM: Fast and accurate vision-language models via knowledge distillation and modal-adaptive pruning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13899–13913, Toronto, Canada. Association for Computational Linguistics.

Max Wertheimer. 1923. Untersuchungen zur lehre von der gestalt ii. *Psycologische Forschung*, 4:301–350.

Ziyun Yang, Kevin Choy, and Sina Farsiu. 2024. Spatial coherence loss: All objects matter in salient and camouflaged object detection.

Yaopei Zeng, Yuanpu Cao, and Lu Lin. 2025. Guarddoor: Safeguarding against malicious diffusion editing via protective backdoors.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847.

Yiming Zhang, Zicheng Zhang, Xinyi Wei, Xiaohong Liu, Guangtao Zhai, and Xiongkuo Min. 2025. Illusionbench: A large-scale and comprehensive benchmark for visual illusion understanding in vision-language models.

Peifei Zhu, Tsubasa Takahashi, and Hirokatsu Kataoka. 2024. Watermark-embedded adversarial examples for copyright protection against diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24420–24430.