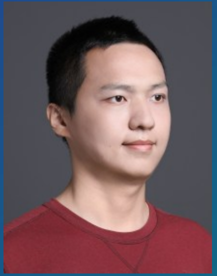
SemVink: Advancing VLMs' Semantic Understanding of Optical Illusions via Visual Global Thinking

Sifan Li[1]

Yujun Cai[2]



Yiwei Wang[1]

[1] University of California, Merced

[2] University of Queensland

EMNLP 2025

The Problem: VLMs Cannot See Hidden Content

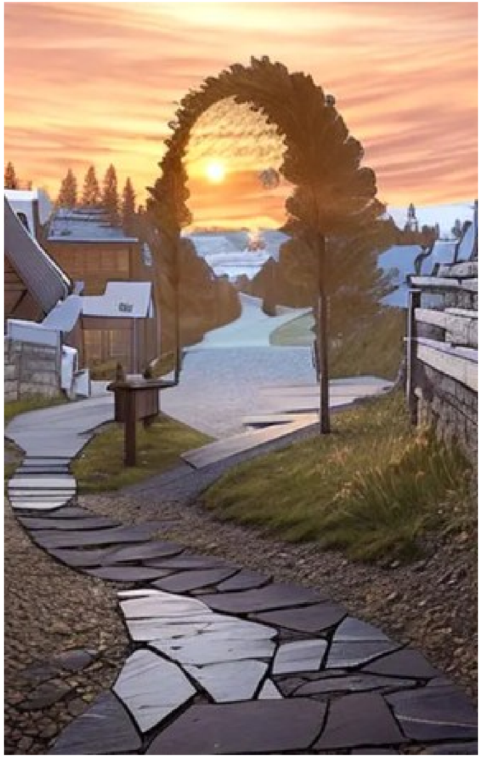
Vision-language models (VLMs) excel at semantic tasks like image captioning and visual reasoning, but they **fail at a core human capability** : detecting hidden content in optical illusions or AI-generated images.

Humans instinctively adjust their visual processing through **perceptual adaptations** like zooming, squinting, or dynamic scaling to uncover obscured details.

VLMs prioritize high-level semantics over low-level visual operations

Static, high-resolution embeddings bury hidden patterns under redundant spatial features

Critical gap between computational vision and human cognition



HC-Bench: A New Benchmark for Hidden Content Recognition

We introduce **HC-Bench**, a benchmark dataset of 112 synthetic images with embedded hidden texts and objects, generated using Stable Diffusion with ControlNet to preserve naturalistic backgrounds.

112

Total Images

56

Hidden Text Images

56

Hidden Object Images

12

VLMs Tested

State-of-the-art VLMs achieve near-zero accuracy: 0–5.36%

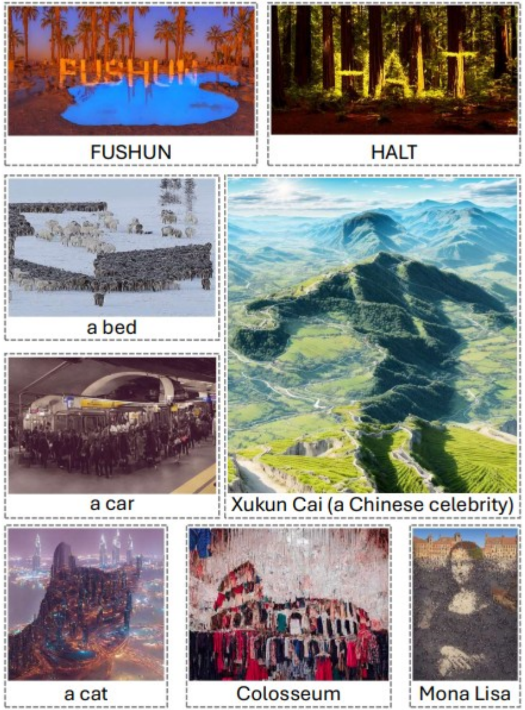
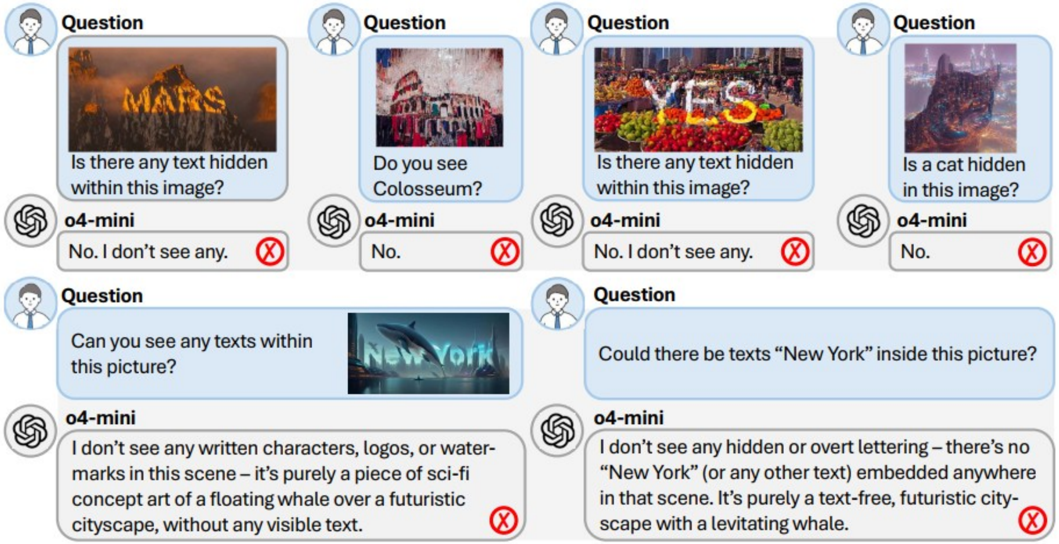


Figure 1: Illusional images can contain hidden texts or hidden images within the obvious background scenes.

Universal VLM Failure: Even SOTA Models Cannot See Hidden Content



Question

Is there any text hidden within this image?

o4-mini

No. I don't see any. ⓧ

Question

Do you see Colosseum?

o4-mini

No. ⓧ

Question

Is there any text hidden within this image?

o4-mini

No. I don't see any. ⓧ

Question

Is a cat hidden in this image?

o4-mini

No. ⓧ

Question

Can you see any texts within this picture?

o4-mini

I don't see any written characters, logos, or watermarks in this scene – it's purely a piece of sci-fi concept art of a floating whale over a futuristic cityscape, without any visible text. ⓧ

Question

Could there be texts "New York" inside this picture?

o4-mini

I don't see any hidden or overt lettering – there's no "New York" (or any other text) embedded anywhere in that scene. It's purely a text-free, futuristic cityscape with a levitating whale. ⓧ

Figure 2: As one of the best state-of-the-art VLMs, o4-mini is incapable in recognizing the hidden texts and objects within images even when we prompt directly with the correct answers. The hidden items in these images are *"MARS", Colosseum, "YES", a cat, and "NEW YORK"*, respectively.

12 State-of-the-Art VLMs Tested:

- O3
- Gemini 2.5 Pro
- Mistral
- LLaVA-v1.5-7B
- Kimi-VL-A3B
- Qwen2-VL-72B
- O4-MINI
- Grok 3
- Claude 3.7 Sonnet
- Doubao-1.5-Pro
- Qwen2-VL-7B
- DeepSeek-VL2

Methods Tested (All Failed):

- ✗ Zero-shot direct questions
- ✗ Zero-shot with follow-up hints
- ✗ Prompt engineering ("zoom in/out to examine layered details")
- ✗ Few-shot learning with examples

Root Cause:

VLMs rely on static, high-resolution embeddings that prioritize **local texture over global structure**, burying hidden patterns under redundant spatial features.

The Solution: SemVink (Semantic Visual Thinking)

A Surprisingly Simple Solution

Scaling images to low resolutions (32–128 pixels)

How It Works:

- **Downscaling eliminates redundant visual noise**

  from high-resolution embeddings

- **Forces models to focus on global patterns**

  rather than local textures

- **Mirrors human perceptual strategies**

  like squinting to see hidden content

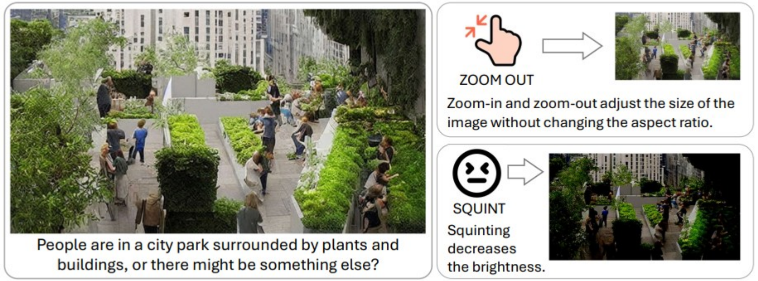

People are in a city park surrounded by plants and buildings, or there might be something else?
ZOOM OUT
Zoom-in and zoom-out adjust the size of the image without changing the aspect ratio.
SQUINT
Squinting decreases the brightness.

Figure 3: Two methods to help humans recognize the hidden content *a Labrador retriever* 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.

Accuracy jumps to

>99%

Experimental Results: Dramatic Improvement with SemVink

Key Findings

- **Universal failure** of baseline methods: All 12 VLMs achieve 0-5.36% accuracy with zero-shot, hints, prompt engineering, and few-shot learning.
- **Dramatic improvement** with zoom-out: Accuracy jumps to 91.07–100% across all models.
- Larger models (O4-MINI, Gemini 2.5 Pro, Qwen2-VL-72B) achieve **perfect 100%** accuracy.
- Even smaller models (Kimi-VL-A3B, LLaVA-v1.5-7B) exceed **90% accuracy**.

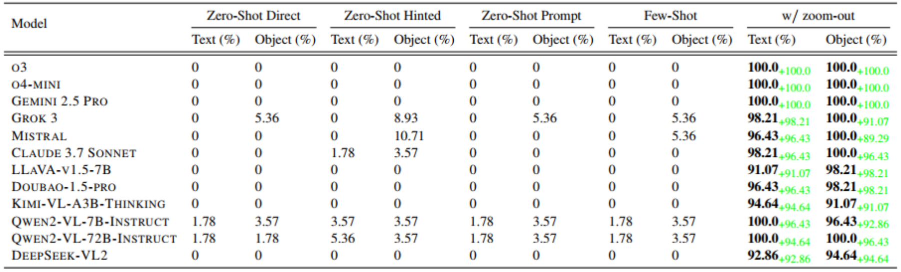
Baseline

0–5.36%

→

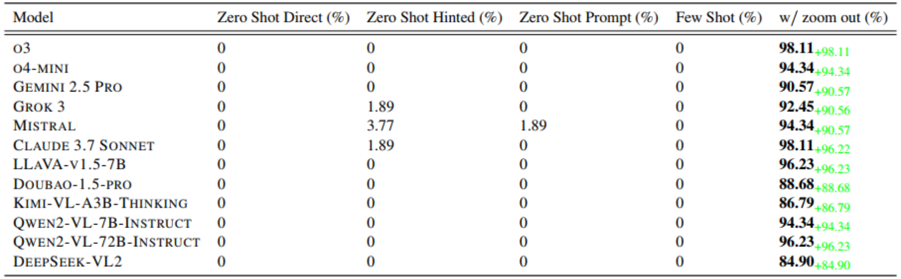
w/ SemVink

91–100%



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 (%)
o3	0	0	0	0	0	0	0	0	**100.0**$_{+100.0}$	**100.0**$_{+100.0}$
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-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 4: The recognition accuracy across different VLMs with four methods mentioned in Section 3.2 and SemVink zoom-out method mentioned in Section 3.3. All tested VLMs are incapable of recognizing the hidden content in the images. With the help of SemVink zoom-out, each tested VLM obtains a nearly 100% success rate.



Model	Zero Shot Direct (%)	Zero Shot Hinted (%)	Zero Shot Prompt (%)	Few Shot (%)	w/ zoom out (%)
o3	0	0	0	0	**98.11**$_{+98.11}$
o4-mini	0	0	0	0	**94.34**$_{+94.34}$
Gemini 2.5 Pro	0	0	0	0	**90.57**$_{+90.57}$
Grok 3	0	1.89	0	0	**92.45**$_{+90.56}$
Mistral	0	3.77	1.89	0	**94.34**$_{+90.57}$
Claude 3.7 Sonnet	0	1.89	0	0	**98.11**$_{+96.22}$
LLaVA-v1.5-7B	0	0	0	0	**96.23**$_{+96.23}$
Doubao-1.5-pro	0	0	0	0	**88.68**$_{+88.68}$
Kimi-VL-A3B-Thinking	0	0	0	0	**86.79**$_{+86.79}$
Qwen2-VL-7B-Instruct	0	0	0	0	**94.34**$_{+94.34}$
Qwen2-VL-72B-Instruct	0	0	0	0	**96.23**$_{+96.23}$
DeepSeek-VL2	0	0	0	0	**84.90**$_{+84.90}$

Table 5: Validation of task difficulty on 53 internet-sourced hidden-content images, collected independently to reduce dataset-specific noise and biases.

Why Does It Work? Embedding Redundancy Analysis

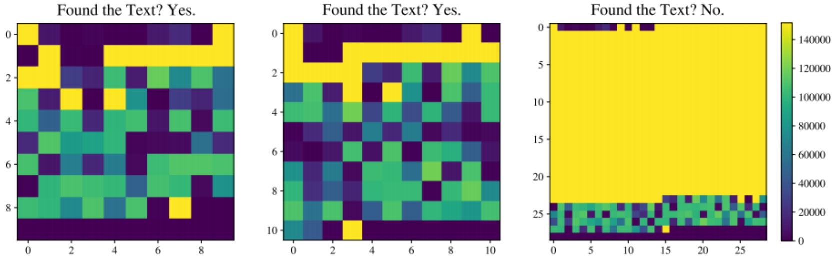


Figure 4: The visualization of the embeddings of the input prompts with the image. In the conditions of the left one (6 consecutive image tokens as in the consecutive yellow region in the heatmap) and center one (10 consecutive image tokens), VLMs can recognize the hidden content. In the condition of 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.

Resolution	Repeated Tokens	Detection
High (512-1440px)	~1000	✗ Failed
Low (32-128px)	~10	☑ Success

Embedding analysis reveals the root cause of VLM failure and explains why downscaling works effectively.

High-resolution embeddings contain **redundant spatial patterns** that obscure subtle details

Attention maps show that VLMs focus excessively on background textures, masking hidden content.

Downscaling shifts attention from local textures to global patterns

Implications and Future Directions

⚠ Critical Architectural Flaw

Current VLMs **lack integrated low-level visual operations** that are fundamental to human perception. They prioritize abstract reasoning over basic visual processing, making them brittle in real-world scenarios requiring perceptual adaptability.

Real-World Applications

- Medical Imaging
- Security Systems
- Adversarial Detection
- Steganography
- Quality Control

Paradigm Shift Needed

Multi-Scale Processing

Integrate dynamic resolution routing and learned scaling schedules into VLM architectures

Hybrid Models

Combine high-level semantic reasoning with low-level visual operations as first-class components

Adaptive Vision Tools

Elevate preprocessing operations to integrable visual tools within VLM pipelines

Human-Like Perception

Bridge the gap between computational vision and human cognition through perceptual adaptability

Conclusion

1

HC-Bench Benchmark

Introduced a benchmark of 112 synthetic images with hidden texts and objects, addressing limitations in existing datasets like EXAMS-V and IllusionBench.

2

Revealed VLM Limitations

Demonstrated universal failure of 12 state-of-the-art VLMs (0–5.36% accuracy) in hidden content recognition, exposing a foundational design flaw prioritizing semantics over basic visual processing.

3

SemVink Solution

Proposed a scalable solution via image scaling to low resolutions (32–128 pixels), achieving over 99% accuracy and demonstrating that low-level operations can bridge the gap between computational vision and human cognition.

Key Takeaway

Integrating low-level visual skills into multimodal architectures is crucial for building robust VLMs that can handle real-world ambiguous scenarios in medical imaging, security, and beyond.

Thank You!

Project Website & Code

https://johnnyzeppelin.github.io/vlm-semvink

Sifan Li
sflijohn@foxmail.com
UC Merced

Yujun Cai
yujun.cai@uq.edu.au
University of Queensland

Yiwei Wang
yiweiwang2@ucmerced.edu
UC Merced

Questions?