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

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 texts and objects, revealing that leading VLMs achieve **near-zero accuracy (0–5.36%)** even with explicit prompting. We propose **SemVink (Semantic Visual Thinking)** by simply scaling images to low resolutions, which unlocks **over 99% accuracy** by eliminating redundant visual noise.

The Problem

Current VLM architectures prioritize **high-level semantic reasoning** at the expense of **low-level visual operations** fundamental to human perception.

VLMs universally fail to detect hidden text or objects, even when explicitly prompted to "zoom in" or "adjust contrast"

Figure 1: Examples of illusional images with hidden texts or objects within background scenes

SemVink Solution

Simple Yet Effective

Scaling images to **low resolutions (32–128 pixels)** eliminates redundant visual noise and achieves **>99% accuracy**

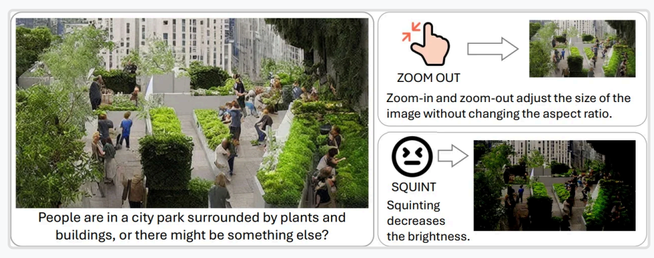

Figure 3: Two methods to help humans recognize hidden content: zoom out and squint

How it works: Downscaling forces models to focus on **global patterns** rather than local textures, mirroring human perceptual strategies.

HC-Bench Dataset

A benchmark dataset for evaluating VLMs' ability to recognize visually hidden content.

112	56	56
Total Images	Hidden Text	Hidden Objects

Generated using Stable Diffusion with ControlNet, balanced for common and rare concepts across Latin/non-Latin scripts and multiple object categories.

Experimental Results

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**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$
o4-mini	0	0	0	0	0	0	0	0	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$
Gemini 2.5 Pro	0	0	0	0	0	0	0	0	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$
Grok 3	0	5.36	0	8.93	0	5.36	0	5.36	**98.21**$_{\pm 6.21}$	**100.0**$_{\pm 0.0}$
Mistral	0	0	0	10.71	0	0	0	5.36	**96.43**$_{\pm 6.43}$	**100.0**$_{\pm 0.29}$
Claude 3.7 Sonnet	0	0	1.78	3.57	0	0	0	0	**98.21**$_{\pm 6.43}$	**100.0**$_{\pm 6.43}$
LLaVA-v1.5-7B	0	0	0	0	0	0	0	0	**91.07**$_{\pm 12.05}$	**98.21**$_{\pm 6.21}$
Doubao-1.5-pro	0	0	0	0	0	0	0	0	**96.43**$_{\pm 6.43}$	**98.21**$_{\pm 6.21}$
Kimi-VL-A3B-Thinking	0	0	0	0	0	0	0	0	**94.64**$_{\pm 6.64}$	**91.07**$_{\pm 12.05}$
Qwen2-VL-7B-Instruct	1.78	3.57	3.57	3.57	1.78	3.57	1.78	3.57	**100.0**$_{\pm 0.43}$	**96.43**$_{\pm 12.86}$
Qwen2-VL-72B-Instruct	1.78	1.78	5.36	3.57	1.78	3.57	1.78	3.57	**100.0**$_{\pm 6.64}$	**100.0**$_{\pm 6.43}$
DeepSeek-VL2	0	0	0	0	0	0	0	0	**92.86**$_{\pm 6.86}$	**94.64**$_{\pm 6.64}$

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**$_{\pm 98.11}$
o4-mini	0	0	0	0	**94.34**$_{\pm 94.34}$
Gemini 2.5 Pro	0	0	0	0	**90.57**$_{\pm 90.57}$
Grok 3	0	1.89	0	0	**92.45**$_{\pm 90.56}$
Mistral	0	3.77	1.89	0	**94.34**$_{\pm 90.57}$
Claude 3.7 Sonnet	0	1.89	0	0	**98.11**$_{\pm 96.22}$
LLaVA-v1.5-7B	0	0	0	0	**96.23**$_{\pm 96.23}$
Doubao-1.5-pro	0	0	0	0	**88.68**$_{\pm 88.68}$
Kimi-VL-A3B-Thinking	0	0	0	0	**86.79**$_{\pm 86.79}$
Qwen2-VL-7B-Instruct	0	0	0	0	**94.34**$_{\pm 94.34}$
Qwen2-VL-72B-Instruct	0	0	0	0	**96.23**$_{\pm 96.23}$
DeepSeek-VL2	0	0	0	0	**84.90**$_{\pm 84.90}$

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

Table 4: Recognition accuracy across different VLMs with SemVink zoom-out method

Key Finding: Larger models (O4-MINI, Gemini 2.5 Pro, Qwen2-VL-72B) achieve **perfect 100% accuracy** with SemVink, while baseline methods yield 0–5.36%.

Universal VLM Failure

We evaluated **12 state-of-the-art VLMs** including O3, O4-MINI, Gemini 2.5 Pro, Claude 3.7 Sonnet, and others.

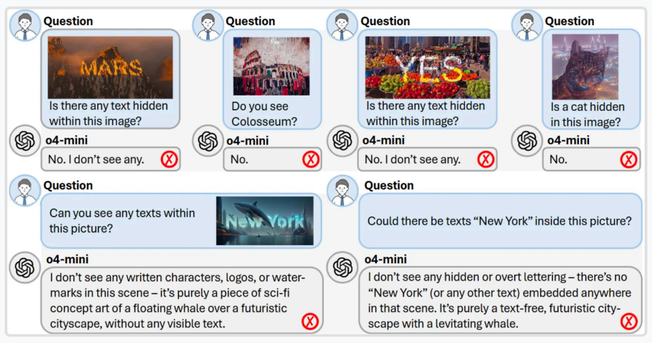

Figure 2: Evaluation protocol with direct questions for hidden text and objects

- **Zero-shot:** 0–5.36% accuracy
- **With hints:** No improvement
- **Prompt engineering & few-shot:** Failed

Why Does It Work?

Embedding Redundancy Analysis:

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

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.

Figure 4: Embedding visualization showing redundancy at different resolutions

Key Contributions

1	2	3
HC-Bench	**Revealed Flaw**	**SemVink**
Introduced benchmark for hidden content recognition	Exposed architectural flaw in VLM design	Proposed scalable solution achieving >99% accuracy

Contact Information

Sifan Li: sflijohn@foxmail.com
Yujun Cai: yujun.cai@uq.edu.au
Yiwei Wang: yiweiwang2@ucmerced.edu

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