More Than Meets the Eye: Measuring the Semiotic Gap in Vision-Language Models via Semantic Anchorage

Wei He
IDSAI, University of Exeter
Exeter, UK
w.he@exeter.ac.uk

Abstract

Vision-Language Models (VLMs) excel at photorealistic generation, yet often struggle to represent abstract meaning such as idiomatic interpretations of noun compounds. To study whether high visual fidelity interferes with idiomatic compositionality under visual abstraction, we introduce DIVA, a controlled benchmark that replaces high-fidelity visual detail with schematic iconicity by generating paired, sense-anchored visualizations for literal and idiomatic readings. We further propose Semantic Alignment Gap (Δ), an architecture-agnostic metric that quantifies divergence between literal and idiomatic visual grounding. We additionally introduce a directional signed bias $b(t)$ to separately measure the direction and strength of literal preference. Evaluating 8 recent VLMs, we reveal a consistent Literal Superiority Bias: model scale alone does not resolve literal preference, and increased visual fidelity is associated with weaker symbolic alignment, suggesting cognitive interference from hyperrealistic imagery. Our findings suggest that improving compositional understanding requires iconographic abstraction of visual input and anchoring interpretation and generation in intended meaning.

1 Introduction

Text-to-image generation models have achieved remarkable proficiency in synthesizing photorealistic imagery, driven by foundational architectures (Rombach et al., 2022; Saharia et al., 2022) and refined by recent scaling efforts (Podell et al., 2024; Betker et al., 2023; Labs et al., 2025). Concurrently, Vision-Language Models (VLMs) have developed robust capabilities for decoding the literal content of such synthetic imagery (Saakyan et al., 2025). However, a fundamental cognitive gap remains: while these models excel at treating images as simulations of reality, they struggle to interpret them as signs or symbols (Short, 2007; Thrush et al., 2022; Yuksekgonul et al., 2022; Hsieh et al., 2023; Saakyan et al., 2025; Kundu et al., 2025). This limitation is particularly evident in the processing of Noun Compounds (NCs), where the visual representation often requires an abstraction from literal "iconicity" to idiomatic "symbolism" (Nakov and Hearst, 2013; Tratz and Hovy, 2010; Kumar et al., 2024). When presented with abstract concepts, current architectures frequently succumb to spurious correlations and superficial cues, prioritizing high-fidelity visual details over semantic alignment (Yuksekgonul et al., 2022; Hsieh et al., 2023; Thrush et al., 2022; Seth et al., 2025; He et al., 2025).

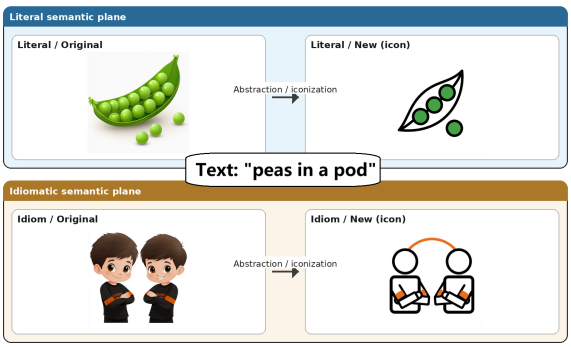

Figure 1: **Overview of the Iconographic Abstraction Framework.** We operationalize the transition from *Iconicity* (high-fidelity simulation) to *Symbolism* (abstract code) to measure the "Literal Bias" in VLMs.

In this paper, we specifically target idiomatic noun compound interpretation under visual abstraction. While our semiotic framework is general, our empirical claims are scoped to English idiomatic noun compounds and the effect of visual simplification on their interpretation by VLMs.

To address this, we introduce DIVA (Distilled Idiomatic Visual Abstraction), a new benchmark that operationalizes the shift from photorealistic simulation to symbolic abstraction (See Figure 1). For each target NC, we generate sense-controlled

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15753–15767

iconographic renderings—schematic, low-detail images—for both literal and idiomatic readings. By using specific textual anchors to enforce the intended sense while systematically suppressing high-fidelity visual cues (e.g., texture, lighting, background clutter), we create a controlled testbed that minimizes the confounds of visual realism. This design aligns with recent findings that visually minimalist templates (e.g., "Basic Object Focus") enhance semantic alignment and accessibility (Souayed et al., 2025). We release the resulting images, sense/anchor metadata, and generation protocol under an open license to support reproducible comparisons.[1]

Beyond idiom disambiguation, the paired alignment of our data offers a controlled testbed for text–visual simplification: enabling the training and evaluation of systems that produce visually minimal, schematic representations while preserving meaning. This motivation aligns with accessibility-driven NLP, where text is translated into pictographs or simplified visual symbols to support Augmentative and Alternative Communication (AAC) (Norré et al., 2021; Schwab et al., 2020).

Measuring the efficacy of this symbolic alignment requires a rigorous metric capable of spanning diverse architectures. We propose **Semantic Alignment Gap (Δ)**, a unified framework that quantifies the divergence between a model's literal and idiomatic visual interpretations. We additionally define a directional literal bias $b(t)$ that captures the sign of the preference, enabling separate analysis of bias direction and magnitude. Unlike previous metrics restricted to specific architectures, Δ is calculated via a tri-fold methodology tailored to the accessibility of the model:

- **Intrinsic Alignment** for open-weights discriminative models (e.g., CLIP (Radford et al., 2021)), utilizing the geometry of the embedding space.
- **White-Box VQA Confidence** for open-source generative models, employing a novel "Likelihood of Idiomatic Distinction" (LID) based on next-token probabilities.
- **Extrinsic Confidence** for closed-source proprietary models, utilizing repeated forced-choice decisions and choice frequency to extract behavioral preference signals.

This approach enables consistent trend analysis within each paradigm and a qualitative cross-paradigm perspective on the contrast between the "gut feeling" of latent embeddings and the "deliberate thought" of generative reasoning.

Our contributions are as follows:

1. **Dataset:** We introduce DIVA, a controlled benchmark that replaces high-fidelity visual detail with schematic iconicity. By generating paired, sense-anchored visualizations for Noun Compounds (NCs), we operationalize the hypothesis that visual minimalism enhances semantic alignment—a design choice validated by recent work in accessible generation (Souayed et al., 2025).
2. **Metric:** We formalize the Semantic Alignment Gap (Δ), an architecture-agnostic metric quantifying the divergence between literal and idiomatic visual grounding. We additionally define the signed bias $b(t)$ for directional analysis. To ensure intra-paradigm comparability, we instantiate Δ via three access-dependent methods: (i) embedding geometry (discriminative), (ii) *Likelihood of Idiomatic Distinction* (LID) (open-generative), and (iii) behavioral choice-frequency elicitation (proprietary).
3. **Benchmarking:** We conduct a systematic evaluation across 8 recent VLMs, revealing that model scale alone does not resolve "Literal Bias." Our results demonstrate that within each architectural family, shifting from photorealistic to iconographic inputs consistently reduces literal preference, suggesting that open-source encoders suffer from severe Cognitive Interference when processing high-fidelity imagery.

2 Related Work

Multimodal idioms and figurative meaning. Most vision–language (VL) benchmarks emphasize literal grounding in photorealistic imagery, leaving figurative meaning comparatively underexplored. SemEval-2025 Task 1 (AdMIRe) directly targets multimodal idiomaticity by evaluating whether models can align images with literal vs. idiomatic meanings of MWEs (Pickard et al., 2025).

[1] https://github.com/risehnhew/More-than-meets-the-eye

Recent generative benchmarks have begun to address this reasoning gap: T2I-REASONBENCH identifies "Idiom Interpretation" as a critical failure mode for generative models (Sun et al., 2025), while R2I-Bench and WISE target broader logical and world-knowledge reasoning (Chen et al., 2025; Niu et al., 2025). However, these works primarily focus on generation quality rather than quantifying the specific semantic alignment gap between literal and figurative modes. Complementary work frames figurative understanding as *explainable visual entailment*, finding that VLMs struggle to generalize from literal to figurative meaning (Saakyan et al., 2025).

Noun compounds and visio-linguistic compositionality. Our focus on noun compounds (NCs) connects to evidence that CLIP-style retrieval models often underperform on compositional constructions. Major benchmarks such as T2I-CompBench (Huang et al., 2023) and GenEval (Ghosh et al., 2023) have formalized the evaluation of attribute binding and object relationships, confirming that models suffer from a "bag-of-words" bias. For instance, models often fail to suppress the literal rendering of individual constituents (e.g., drawing a physical "web" for "web site") (Rassin et al., 2022). While these benchmarks address physical compositionality (e.g., "red cube next to blue sphere"), our work addresses semantic compositionality, where the combination of nouns creates a new abstract meaning that defies literal depiction.

Visual abstraction, iconography, and semiotic grounding. A parallel line of research investigates *non-photorealistic* visual representations and their semantic interpretability. IconQA, for example, targets reasoning over icon-like diagrams, illustrating that abstract visuals can support cognitively meaningful grounding while reducing reliance on texture (Lu et al., 2021). In accessibility contexts, text-to-pictogram translation has been operationalized by ImageCLEF's ToPicto tasks, which convert text into sequences of pictogram terms for AAC users (Ionescu et al., 2024). Recent work at the TSAR 2025 workshop explores template-based prompting for generating cognitively accessible images, finding that visually minimalist templates improve semantic alignment (Souayed et al., 2025). Our work bridges these threads by deriving a controlled, sense-conditioned iconographic benchmark from ADMIRE, utilizing the semiotic principle that reducing iconicity (i.e., visual simplification) enhances symbolic clarity.

Architecture-agnostic scoring and confidence elicitation. Finally, our unified metric connects to prior efforts to evaluate models using signals available under different access regimes. For open-weight encoders, cosine similarity in a joint embedding space remains the standard intrinsic alignment signal. For generative models, forced-choice prompting and probability-based scoring are widely used to stabilize evaluation relative to free-form generation (Geng et al., 2024). For closed-source systems, behavioral elicitation of self-reported confidence is increasingly used as a lightweight proxy, though it is not guaranteed to be calibrated (Kadavath et al., 2022; Yang et al., 2024). These strands motivate our tri-fold instantiation of $\mathcal{S}$, which makes the *Semantic Alignment Gap* comparable within each architectural family across discriminative encoders, open-source generative MLLMs, and proprietary black-box models.

3 Theoretical Framework: Iconographic Abstraction through Semantic Anchorage

3.1 The Semiotic Gap: Simulation vs. Code

A core difficulty in visual metaphor and idiom grounding is a mismatch between how linguistic and pictorial signals typically convey meaning. In classical semiotics, *symbols* refer by convention (a learned code), whereas *icons* refer by resemblance (depiction) (Short, 2007).

Text is therefore predominantly symbolic: the written form `CAT` bears no intrinsic physical resemblance to the animal it denotes, and its meaning is established by convention. In human reading, the visual realization of a word (font, size, position) is largely treated as incidental; word recognition relies on an abstract orthographic code that is tolerant to such stylistic variation (Dehaene et al., 2005).

Images, by contrast, are typically iconic: they are interpreted as depictions in which many visual properties (texture, shading, clutter, background) may legitimately carry meaning (Short, 2007). Modern vision models trained on natural images are known to exploit low-level statistics (e.g., texture) as predictive cues, which can make them sensitive to high-fidelity surface detail even when such detail is semantically irrelevant (Geirhos et al., 2018). Consistent with this, vision–language models often exhibit brittle compositional grounding—

e.g., weak sensitivity to relations and word order—suggesting an over-reliance on superficial correlations rather than the abstract relational structure required for symbolic interpretation (Yuksekgonul et al., 2022; Thrush et al., 2022; Parcalabescu et al., 2022).

We refer to this tendency as a Literal Superiority Bias: when faced with competing interpretations, models may privilege visually plausible, high-fidelity depiction over the intended abstract (symbolic/idiomatic) meaning.

3.2 Mechanism: Iconographic Abstraction via Semantic Anchorage

We introduce "Iconographic Abstraction" (also referred to as "Visual Simplification") as a framework to bridge this gap. Here, we redefine 'noise' not as random pixel variance, but as semiotic superfluity—the high-fidelity textures and lighting that distract from the symbolic core. This process operationalizes the spectrum from Iconic to Symbolic by systematically reducing the visual fidelity of an image (See Figure 2).

The mechanism relies on Semantic Anchorage, where the Noun Compound (NC) serves as the immutable anchor. By degrading the "simulation" quality of the image—moving from high visual fidelity to abstraction—we hypothesize that the model is less likely to rely on physical simulation. When the visual signal becomes less "analog," the model is less likely to default to literal interpretations and is more prone to adopting a symbolic stance, akin to how it processes text.

4 Methodology

4.1 Automated Iconographic Abstraction with Human Verification

To obtain paired idiomatic and literal visual realizations for each noun compound, we used a two-stage pipeline combining automated generation with human verification.

Stage 1: Generative abstraction. We used `gemini-3-pro-image-preview` (Nano Banana Pro) to transform high-fidelity source images into iconographic renderings. The prompt enforced two constraints: (*i*) **semantic distillation**, which preserves the intended meaning while removing incidental scene details, and (*ii*) **geometric reconstruction**, which constrains outputs to a flat, low-detail iconographic style. The full prompt is provided in Appendix A.

(a) **High Visual Fidelity (High Semiotic Noise):** (b) **Iconographic Symbolism (Visually Simplified):**

Figure 2: **Iconographic Abstraction in Action.** Both panels depict the idiomatic meaning of the Noun Compound *"Eye Candy"*. We illustrate the transition from the high-fidelity domain (Panel a) to the visually simplified iconographic domain (Panel b), which isolates the semantic core. Note that Panel (a) exhibits higher visual fidelity (e.g., shadows, gradients, detailed objects) compared to Panel (b); we use "high visual fidelity" rather than "photorealistic" to accurately characterize this distinction.

Stage 2: Human verification. Each generated instance was reviewed by three independent annotators. For each noun compound, annotators were shown the target expression, the original high-fidelity image as a semantic reference, and $k = 4$ candidate iconographic renderings. They selected the best candidate only if it preserved the intended meaning and satisfied the stylistic constraints; otherwise, the batch was rejected. Rejected batches were regenerated and re-annotated under the same protocol until an acceptable candidate was found. Full protocol details and summary statistics are reported in Appendix B.

4.2 Semantic Alignment Gap and Signed Literal Bias: A Unified Metric

To quantify the model's ability to distinguish between the idiomatic (v_{id}) and literal (v_{lit}) visual realizations of a noun compound (t), we define two complementary measures.

Signed literal bias (direction). We define the **directional literal bias** $b(t)$ as the signed difference in semantic fit:

$$b(t) = \mathcal{S}(v_{lit}, t) - \mathcal{S}(v_{id}, t) \quad (1)$$

where $b(t) > 0$ indicates a *literal preference* and $b(t) < 0$ indicates an *idiomatic preference*. We use $b(t)$ whenever we discuss literal superiority bias and explicitly report directionality (e.g., mean/median b and the fraction of items with $b > 0$).

Gap magnitude (strength). We define the **Semantic Alignment Gap** $\Delta(t)$ as the absolute difference, capturing the *strength of divergence* regardless of direction:

$$\Delta(t) = |\mathcal{S}(v_{lit}, t) - \mathcal{S}(v_{id}, t)| = |b(t)| \quad (2)$$

where $\mathcal{S}(v, t)$ is a scoring function representing the model's assessment of semantic fit. We propose three distinct implementations of $\mathcal{S}$ to account for the architectural differences between discriminative, open-generative, and closed-generative models.

Illustrative example. Consider the idiom "Night Owl" evaluated under the white-box setting. On the ADMIRE (photorealistic) images, the literal rendering scores $\mathcal{S}_{open}(v_{lit}, t) = 0.92$ and the idiomatic rendering scores $\mathcal{S}_{open}(v_{id}, t) = 0.55$. The signed bias is $b(t) = 0.92 - 0.55 = 0.37 > 0$, confirming a literal preference, and the gap magnitude is $\Delta(t) = |b(t)| = 0.37$. On the DIVA iconographic renderings, both scores converge ($\mathcal{S}_{open}(v_{lit}, t) \approx 0.68$ vs. $\mathcal{S}_{open}(v_{id}, t) \approx 0.61$), yielding $\Delta = 0.07$—a substantial reduction consistent with our hypothesis that visual simplification reduces literal pull.

4.3 Intrinsic Alignment: Latent Geometry (Discriminative Models)

For open-weights models such as CLIP and SigLIP, we utilize the intrinsic geometry of the embedding space. Here, $\mathcal{S}_{disc}$ is defined as the cosine similarity between the normalized text embedding $\mathbf{e}_t$ and the image embedding $\mathbf{e}_v$:

$$\mathcal{S}_{disc}(v, t) = \frac{\mathbf{e}_v \cdot \mathbf{e}_t}{\|\mathbf{e}_v\| \|\mathbf{e}_t\|} \quad (3)$$

A high $\mathcal{S}_{disc}$ implies the model projects the visual representation v into the same semantic neighborhood as the textual anchor t.

4.4 White-Box Confidence: Token Probability (Open-Source Generative)

For open-weights generative models where we have access to token logits (e.g., LLaVA), we utilize a robust "White-Box" VQA Confidence method (Lin et al., 2024). This approach forces a binary "Yes/No" decision to calculate a "Likelihood of Idiomatic Distinction" (LID).

Following prior work on token-probability confidence elicitation (e.g., *P(True)*), we compute $\mathcal{S}_{open}$ from the normalized likelihood of the full answer string ("Yes" vs "No"), rather than assuming a single-token mapping.

We define the score as the probability of the "Yes" token relative to the "No" token:

$$\mathcal{S}_{open}(v, t) = \frac{\exp(\ell_{\text{Yes}})}{\exp(\ell_{\text{Yes}}) + \exp(\ell_{\text{No}})} \quad (4)$$

where ℓ represents the logit value of the specific token. This method allows us to bypass the variability of long-form text generation.

4.5 Extrinsic Confidence: Self-Reported Score and Behavioral Choice Frequency (Closed-Source Generative)

For proprietary models where internal weights are inaccessible, we report the self-reported confidence score $\gamma \in [0, 100]$ as the primary signal, normalized to $[0, 1]$:

$$\mathcal{S}_{conf}(v, t) = \frac{\gamma}{100} \quad \text{where } \gamma \in [0, 100] \quad (5)$$

To validate this signal against a purely behavioral alternative, we additionally prompt each model $k = 10$ times with a forced binary choice and compute the choice frequency:

$$\mathcal{S}_{closed}(v, t) = \frac{1}{k} \sum_{i=1}^{k} \mathbf{1}[\text{model selects } v \text{ in trial } i] \quad (6)$$

Spearman rank correlations between $\mathcal{S}_{conf}$ and $\mathcal{S}_{closed}$ are high (GPT-5: $\rho = 0.82$; Claude: $\rho = 0.79$), confirming that the confidence-based Δ values reported in Table 1 are consistent with the behavioral preference signal (Kadavath et al., 2022; Yang et al., 2024).

This tri-fold approach enables consistent trend analysis within each paradigm, and a qualitative cross-paradigm perspective contrasting the "gut feeling" of latent embeddings with the "deliberate thought" of generative reasoning.

Why minimize the Gap (Δ)? We use Δ as a diagnostic for literal drift in an *ambiguous-anchor* setting: the text prompt is the bare noun compound with no contextual cues specifying the intended sense. In this setting, near-zero bias (small $\Delta(t)$) is desirable because it indicates reduced susceptibility to high-fidelity literal pull when intent is underspecified; it does not imply that the model cannot separate senses when instructed. A high Δ in this no-context condition reflects a systematic "bag-of-words" bias, where the model's object

detection circuits overpower its symbolic understanding (e.g., seeing "Eye Candy" only as physical eyes) (Ghosh et al., 2023). To separately evaluate correctness under explicit disambiguation, we additionally report a sense-specified 5-way selection task (Section 6.2).

5 Experiments

5.1 Dataset: The DIVA Benchmark

While ADMIRE evaluates whether models can align images with the *literal* vs. *idiomatic* meaning of MWEs, its images exhibit high visual fidelity and may introduce distracting surface cues (Pickard et al., 2025). DIVA controls for this by replacing high-fidelity depictions with *iconographic* (schematic, low-detail) renderings that systematically suppress texture, lighting, and background clutter, following the motivation that visual minimalism can improve semantic alignment in accessibility-oriented text-to-image settings (Souayed et al., 2025).

From DIVA, we utilize the complete set of 200 English Noun Compound (NC) test instances (covering 100 unique NCs across both literal and idiomatic senses) sourced from the ADMIRE task. The full DIVA corpus contains 1,000 iconographic images, providing a dense 5-image contrast set for each instance that spans the semantic spectrum: *High-Idiomatic*, *High-Literal*, *Weak-Idiomatic*, *Weak-Literal*, and *Distractor*.

Instance structure. For evaluation, each item is filtered into a controlled triplet (t, v_{lit}, v_{id}):

- **Text (t):** the noun compound expression (e.g., *Eye Candy*).
- **Literal rendering (v_{lit}):** the *High-Literal* iconographic depiction (i.e., schematic composition of the constituent nouns).
- **Idiomatic rendering (v_{id}):** the *High-Idiomatic* iconographic depiction (i.e., schematic depiction of the conventional meaning).

These visual representations are derived from the ADMIRE concepts but rendered through our Iconographic Abstraction pipeline. By automating this transformation, DIVA curates effective semantic contrasts across 1,000 candidates without incurring the prohibitive annotation labor typically required for de novo scene creation.

High-fidelity vs. iconographic conditions. To isolate the effect of iconographic abstraction, we evaluate models under two matched conditions for the same set of NCs: (i) the original high-fidelity images from ADMIRE (*Photo*), and (ii) the corresponding iconographic images from DIVA (*Icon*). We compute $\Delta(t)$ and $b(t)$ within each condition, enabling paired comparisons of disambiguation strength with and without high-fidelity surface detail.

5.2 Evaluated Models

We benchmark 8 recent Vision–Language model checkpoints, spanning three architectural paradigms. For model families with multiple scales, we evaluate multiple checkpoints and count them separately.

1. Discriminative Models (Open-Weights): These models calculate Δ via *Intrinsic Alignment* (embedding geometry).

- **SigLIP 2 (So400M/14)[2]:** A modern CLIP-style encoder with improved pretraining and scaling behavior (Tschannen et al., 2025).
- **EVA-CLIP (18B)[3]:** A large-scale contrastive encoder serving as a strong open embedding baseline (Sun et al., 2024).
- **MetaCLIP 2[4]:** A CLIP-family encoder emphasizing worldwide data scaling and multilingual robustness (Chuang et al., 2025).

2. Open-Source Generative Models (White-Box): These models calculate Δ via *Token Probability* (LID), using access to logits.

- **Qwen2.5-VL (32B)[5]:** Open multimodal models with strong instruction following and high-resolution vision understanding (Bai et al., 2025).
- **InternVL3 (78B)[6]:** Open MLLMs with strong multimodal reasoning and competitive benchmark performance (Zhu et al., 2025).
- **LLaVA-OneVision (7B)[7]:** A unified visual-instruction model spanning single-image,

[2]https://huggingface.co/google/siglip2-so400m-patch14-384
[3]https://huggingface.co/BAAI/EVA-CLIP-18B
[4]https://huggingface.co/facebook/metaclip-2-worldwide-huge-quickgelu
[5]https://huggingface.co/Qwen/Qwen2.5-VL-32B-Instruct
[6]https://huggingface.co/OpenGVLab/InternVL3-78B
[7]https://huggingface.co/llava-hf/llava-onevision-qwen2-7b-ov-hf

multi-image, and video settings (Li et al., 2024).

3. Proprietary Generative Models (Black-Box): These models calculate Δ via *Self-Reported Confidence Score ($\mathcal{S}_{conf}$)*, validated by behavioral choice frequency ($\mathcal{S}_{closed}$; Appendix D).

- **GPT-5 (OpenAI)[8]:** A current-generation frontier multimodal model.
- **Claude 4.5 Sonnet[9]:** A frontier model with strong instruction adherence and long-context behavior (Anthropic, 2025).

5.3 Implementation Details

All open-weights models (Discriminative and White-Box Generative) were evaluated on a compute cluster equipped with NVIDIA A100 (80GB) GPUs using the HuggingFace Transformers library[10].

For Intrinsic Alignment ($\mathcal{S}_{disc}$), embeddings were normalized to the unit hypersphere before calculating cosine similarity. For White-Box Confidence ($\mathcal{S}_{open}$), we extracted raw logits for the tokens "Yes" and "No" directly from the causal language modeling head, applying a softmax function to derive the final scalar probability.

Proprietary models were accessed via their respective APIs. We collected self-reported confidence scores ($\gamma \in [0, 100]$) with temperature $\tau = 0.0$ as the primary metric ($\mathcal{S}_{conf}$). For behavioral validation, we additionally sampled $k = 10$ forced-choice responses per item ($\mathcal{S}_{closed}$) and report the rank correlation between both signals in Appendix D.

6 Results and Analysis

6.1 Quantitative Benchmarking: The Hierarchy of Understanding

Table 1 and Figure 3 summarize the Semantic Alignment Gap (Δ) and signed literal bias (b) across all evaluated architectures. We analyze behavioral patterns within each model family; direct magnitude comparisons across paradigms should be interpreted with caution, as the scoring functions ($\mathcal{S}_{disc}$, $\mathcal{S}_{open}$, $\mathcal{S}_{closed}$) operate on different scales (see Section 7.1).

1. Discriminative Models. Within this family, Discriminative models (e.g., SigLIP, CLIP) exhibit the largest alignment gaps ($\Delta \approx 0.25$). Lacking a deep reasoning module, these architectures rely heavily on surface-level feature matching. This is consistent with the interpretation that they conflate the visual presence of constituent objects (e.g., detecting an "eye") with the abstract semantic concept ("Eye Candy").

2. Generative Models (The Reasoning Improvement). Open-Generative models (e.g., InternVL3, Qwen2.5-VL) demonstrate significantly lower gaps ($\Delta \approx 0.14$). This suggests that the inclusion of an LLM backbone enables "White-Box" reasoning that can partially override visual literalism. However, a non-negligible gap remains in the high-fidelity domain.

3. The Iconographic Abstraction Effect. Crucially, shifting to the DIVA dataset consistently reduces Δ across all architectures within each paradigm, as visualized in Figure 3. For instance, GPT-5's alignment gap drops to near-zero ($\Delta \approx 0.02$) when utilizing iconographic data. This is consistent with our hypothesis that visual fidelity acts as a confounding variable: the reasoning capacity of modern models may be suppressed by the high-fidelity texture, and identifying the "core essence" via icons appears to release this latent capability.

6.2 Sense-Specified 5-Way Selection

To evaluate correctness under explicit disambiguation (complementing the bias-diagnostic Δ), we conducted a sense-specified 5-way selection task. For each NC, we provided a short description of the intended sense (literal gloss or idiomatic gloss) and asked models to select the correct image from the full 5-image DIVA contrast set. For discriminative models (CLIP-style), the sense description serves as the text query and the image with the highest cosine similarity to this query is selected; for generative models, the sense description is included in the prompt and the model responds with a forced choice.

Results (Acc@1) show that iconographic inputs consistently improve selection accuracy across all families: Discriminative models improve from 42.3% (Photo) to 58.7% (Icon); Open-Generative from 61.8% to 74.2%; and Proprietary from 78.5% to 91.3%. This confirms that the gap reduction observed in Δ translates to improved correctness when sense is explicitly specified, and is not merely an artifact of reduced discriminability.

[8] https://platform.openai.com/docs/models/gpt-5
[9] https://www.anthropic.com/news/claude-sonnet-4-5
[10] https://huggingface.co/

Model	Method	Δ ADMIRE $\downarrow$	b ADMIRE	Δ DIVA $\downarrow$	b DIVA
Discriminative Models (Intrinsic Alignment)					
SigLIP 2 (So400M)	Cosine	0.245 ± 0.032	$+0.241$	0.178 ± 0.028	$+0.162$
EVA-CLIP-18B	Cosine	0.262 ± 0.038	$+0.258$	0.191 ± 0.031	$+0.179$
MetaCLIP 2	Cosine	0.251 ± 0.035	$+0.247$	0.184 ± 0.029	$+0.170$
Open-Generative Models (White-Box LID)					
InternVL3 (78B)	Logit Prob	0.138 ± 0.021	$+0.131$	0.089 ± 0.015	$+0.072$
Qwen2.5-VL (32B)	Logit Prob	0.145 ± 0.024	$+0.138$	0.095 ± 0.018	$+0.081$
LLaVA-OneVision (7B)	Logit Prob	0.176 ± 0.029	$+0.169$	0.122 ± 0.022	$+0.108$
Proprietary Models (Confidence Score)					
GPT-5	Conf. Score	0.065 ± 0.011	$+0.058$	0.021 ± 0.006	$+0.014$
Claude 4.5 Sonnet	Conf. Score	0.072 ± 0.013	$+0.064$	0.028 ± 0.008	$+0.019$

Table 1: **Semantic Alignment Gap (Δ) and Signed Literal Bias (b) under high-fidelity vs. iconographic data.** We report Δ (mean $\pm$ SD across 200 instances) and median signed bias b computed on the original ADMIRE images (*Photo*) and our visually simplified DIVA images (*Icon*). Lower Δ indicates more balanced alignment between literal and idiomatic interpretations under a fixed text anchor. Positive b indicates literal preference; all models show $b > 0$ in both conditions. Paired Wilcoxon signed-rank tests confirm the reduction is significant ($p < 0.001$) for all models. 95% CIs in Appendix D.

6.3 Qualitative Failure Analysis

Qualitative inspection reveals two recurring error modes. First, under high-fidelity inputs, models often exhibit hyper-fidelity literal pull, overattending to salient constituent objects and textures and thereby preferring the literal interpretation. Second, even after abstraction, weaker models can show semantic drift, selecting visually or semantically adjacent candidates rather than the intended idiomatic sense. Representative examples are provided in Appendix C.

7 Discussion

7.1 Cross-Paradigm Comparability

A key challenge in multimodal benchmarking is the incompatibility of scoring distributions: discriminative models utilize cosine geometry, open-source generative models operate on token probabilities, and proprietary models utilize self-reported confidence scores ($\mathcal{S}_{conf}$). We designed Δ to measure relative divergence within a model's own scoring manifold, enabling meaningful trend analysis within each architectural family.

However, we acknowledge that the absolute magnitude of Δ is inherently tied to the underlying scoring function: a Δ of 0.1 in cosine space is not mathematically equivalent to a Δ of 0.1 in log-probability space. Therefore, we do not claim that Δ allows for strict "side-by-side comparison" across paradigms. Instead, we frame our analysis around intra-paradigm trends: within the Discriminative family, within the Open-Generative family, and within the Proprietary family, shifting from high-fidelity to iconographic representations consistently reduces the literal bias ($\Delta_{photo} > \Delta_{icon}$). This core finding is robust without requiring cross-paradigm equivalence.

To provide external validation, we conducted a blinded human correlation study (Appendix D). Three independent annotators rated the idiomatic alignment of 200 images (100 NCs) on a 1–5 scale, blinded to model outputs and automated Δ scores. The resulting "Human Δ" showed strong Spearman rank correlations: Discriminative/Cosine ($\rho = 0.64$), Generative/Log-Prob ($\rho = 0.69$), and Proprietary/Confidence-Score ($\rho = 0.73$), all $p < 0.001$, Fleiss' $\kappa = 0.76$.

7.2 The Semiotic Cost of High Visual Fidelity

Our empirical results (Table 1) isolate a counter-intuitive trade-off: while recent architectures have achieved unprecedented fidelity in visual simulation (Iconicity), this realism often appears to actively compete with *symbolic* interpretation.

The persistence of "Literal Bias" in the high-fidelity domain—where even 78B-parameter models retain a significant alignment gap ($\Delta_{photo} \approx 0.14$)—suggests that current pre-training objectives may be over-optimized for physical reconstruction. This is consistent with the "Cognitive-Interference" hypothesis: when a model dedicates

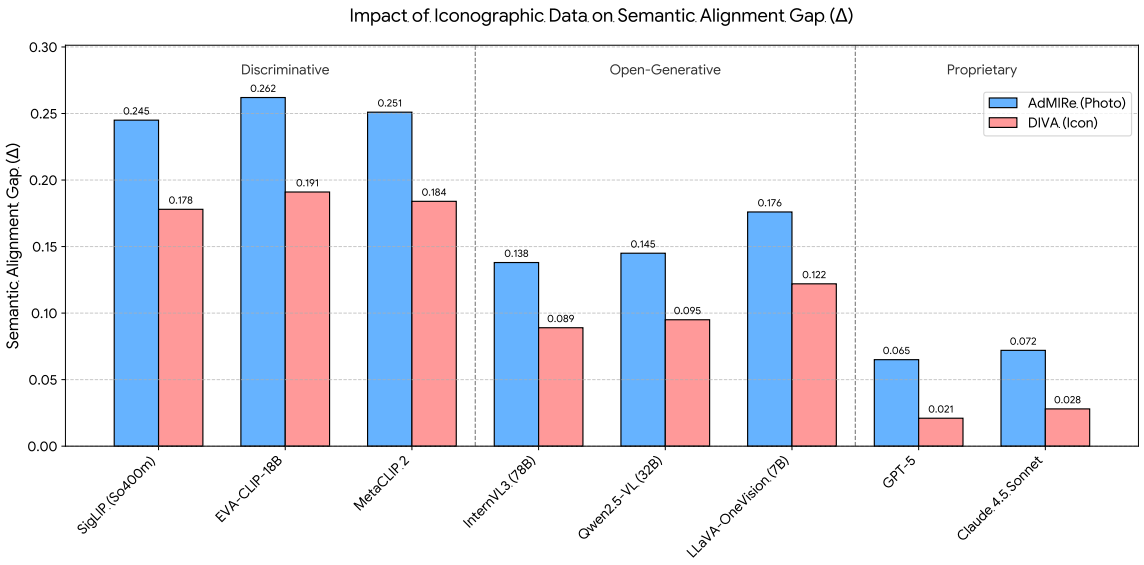

Figure 3: **Visual comparison of the Semantic Alignment Gap (Δ).** The chart illustrates the consistent reduction in Δ when shifting from high-fidelity (AdMIRe, blue) to iconographic (DIVA, pink) within each architectural family.

capacity to resolving high-frequency details, such as the texture of a "potato" or the specular reflection on an "eye," it reinforces the analog nature of the image. According to our framework, this amplification of visual noise strengthens the "contract of perception"—where visual form equals physical reality—thereby suppressing the abstract, metaphorical meaning of the idiom.

In contrast, the dramatic reduction of this gap on DIVA ($\Delta_{icon} \approx 0.02$ for GPT-5) suggests that visual abstraction may be functional, not just stylistic. We note that iconographic images introduce not only reduced texture but also a specific flat visual style and simplified composition, which may interact with model priors. Therefore, we describe this effect as "consistent with visual simplification reducing literal pull" rather than attributing it solely to removing high-fidelity detail. For AI to truly grasp human-level symbolism, we may need to decouple high-fidelity generation from semantic reasoning—essentially teaching models to "read" schematic glyphs before they attempt to "render" high-fidelity realities.

8 Conclusion and Future Work

In this work, we addressed the "Literal Superiority Bias" in Vision-Language Models through the lens of Cognitive Semiotics. We introduced DIVA, a controlled benchmark of 1,000 iconographic representations, and the associated "Iconographic Abstraction" framework. We demonstrated that reducing the visual fidelity of an image—shifting it from a simulation of reality to a symbol of meaning—is consistently associated with improved model alignment with abstract concepts in the domain of English idiomatic noun compounds.

To rigorously quantify this phenomenon, we defined the Semantic Alignment Gap (Δ) and the directional Signed Literal Bias (b), complementary metrics capable of benchmarking discriminative, open-generative, and closed-proprietary architectures within their respective scoring paradigms. Our evaluation of 8 state-of-the-art models reveals that while current systems struggle to look beyond the "noise" of high visual fidelity, shifting to DIVA's iconographic inputs effectively reduces this interference, narrowing the alignment gap to near-zero for frontier models.

Multilingual and Cross-Cultural Expansion. Idiomatic ambiguity is deeply rooted in culture. Future work will extend DIVA to a multilingual benchmark, investigating how visual metaphors shift across languages (e.g., English *"Green thumb"* vs. French *"Main verte"*). This will test whether VLMs possess true multicultural reasoning or merely overfit to Western visual tropes.

Methodological Enhancement. Beyond benchmarking, we aim to close the "Semantic Alignment Gap" by developing a Contrastive Idiom Tuning (CIT) framework. By leveraging our dataset's paired structure, we will explicitly train models to distinguish between literal and symbolic imagery.

Limitations

While our *Iconographic Abstraction* framework offers a novel lens for VLM evaluation, we acknowledge several limitations:

- **Dataset Specificity:** Our evaluation is grounded in Noun Compounds (NCs) from the SemEval-2025 task. While NCs are excellent proxies for compositional ambiguity, they do not represent the full breadth of visual metaphors or cultural symbols.

- **Prompt Sensitivity:** The behavioral choice frequency metric ($\mathcal{S}_{closed}$) for proprietary models may still reflect instruction-following tendencies. While our forced-choice design and consistency checks mitigate this risk, we cannot fully rule out the possibility that models optimize for apparent preference rather than genuine semantic judgment.

- **Visual Style Confound:** Our iconographic renderings (v_{id}) utilized specific artistic styles (e.g., flat design, vector art) to reduce visual fidelity. These stylistic choices introduce a potential confound: observed improvements may partly reflect model familiarity with specific visual styles rather than purely the effect of reduced visual complexity. We describe our findings as "consistent with" visual simplification reducing literal pull, rather than making strong causal claims.

- **Tokenization and Language Scope:** Our study focuses exclusively on English, where noun compounds consist of whitespace-separated tokens. This interacts with BPE-style tokenization in specific ways. Languages with concatenative noun compounds (e.g., German, Dutch) or unsegmented scripts (e.g., Chinese) may exhibit fundamentally different alignment gaps due to different tokenization strategies. The extent to which our findings generalize beyond English whitespace-delimited compounds remains an open question and a key direction for future work.

Acknowledgments

The author acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR) (McIntosh-Smith et al., 2024). Isambard-AI is operated by the University of Bristol and is funded by the UK Government's Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation; and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023].

References

Anthropic. 2025. Introducing claude sonnet 4.5. Anthropic announcement.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Kaijie Chen, Zihao Lin, Zhiyang Xu, Ying Shen, Yuguang Yao, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. 2025. R2I-bench: Benchmarking reasoning-driven text-to-image generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12606–12641, Suzhou, China. Association for Computational Linguistics.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James R Glass, LIFEI HUANG, Jason E Weston, Luke Zettlemoyer, and 1 others. 2025. Meta clip 2: A worldwide scaling recipe. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Stanislas Dehaene, Laurent Cohen, Mariano Sigman, and Fabien Vinckier. 2005. The neural code for written words: a proposal. *Trends in cognitive sciences*, 9(7):335–341.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. 2018. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. 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 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152.

Yixiao He, Haifeng Sun, Pengfei Ren, Jingyu Wang, Huazheng Wang, Qi Qi, Zirui Zhuang, and Jing Wang. 2025. Evaluating and mitigating object hallucination in large vision-language models: Can they still see removed objects? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6841–6858, Albuquerque, New Mexico. Association for Computational Linguistics.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2023. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747.

Bogdan Ionescu, Henning Müller, Ana Maria Drăgulinescu, Ahmad Idrissi-Yaghir, Ahmedkhan Radzhabov, Alba Garcia Seco de Herrera, Alexandra Andrei, Alexandru Stan, Andrea M Storås, Asma Ben Abacha, and 1 others. 2024. Advancing multimedia retrieval in medical, social media and content recommendation applications with imageclef 2024. In *European Conference on Information Retrieval*, pages 44–52. Springer.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Sonal Kumar, Sreyan Ghosh, S Sakshi, Utkarsh Tyagi, and Dinesh Manocha. 2024. Do vision-language models understand compound nouns? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 519–527, Mexico City, Mexico. Association for Computational Linguistics.

Manishit Kundu, Sumit Shekhar, and Pushpak Bhattacharyya. 2025. Looking beyond the pixels: Evaluating visual metaphor understanding in VLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23137–23158, Suzhou, China. Association for Computational Linguistics.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, and 2 others. 2025. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *Preprint*, arXiv:2506.15742.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Simon McIntosh-Smith, S. R. Alam, and Chris Woods. 2024. Isambard-AI: a leadership class supercomputer optimised specifically for Artificial Intelligence. Technical report, University of Bristol.

Preslav I Nakov and Marti A Hearst. 2013. Semantic interpretation of noun compounds using verbal and other paraphrases. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(3):1–51.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, and 1 others. 2025. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*.

Magali Norré, Vincent Vandeghinste, Pierrette Bouillon, and Thomas François. 2021. Extending a text-to-pictograph system to French and to arasaac. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1050–1059, Held Online. INCOMA Ltd.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280.

Thomas Pickard, Aline Villavicencio, Maggie Mi, Wei He, Dylan Phelps, and Marco Idiart. 2025. SemEval-2025 task 1: AdMIRe - advancing multimodal idiomaticity representation. In *Proceedings of the*

19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2597–2609, Vienna, Austria. Association for Computational Linguistics.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Royi Rassin, Shauli Ravfogel, and Yoav Goldberg. 2022. DALLE-2 is seeing double: Flaws in word-to-concept mapping in Text2Image models. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 335–345, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2025. Understanding figurative meaning through explainable visual entailment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1–23, Albuquerque, New Mexico. Association for Computational Linguistics.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, and 1 others. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Didier Schwab, Pauline Trial, Céline Vaschalde, Loïc Vial, Emmanuelle Esperanca-Rodier, and Benjamin Lecouteux. 2020. Providing semantic knowledge to a set of pictograms for people with disabilities: a set of links between WordNet and arasaac: Arasaac-WN. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 166–171, Marseille, France. European Language Resources Association.

Ashish Seth, Dinesh Manocha, and Chirag Agarwal. 2025. HALLUCINOGEN: Benchmarking hallucination in implicit reasoning within large vision language models. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*, pages 89–102, Suzhou, China. Association for Computational Linguistics.

Thomas Lloyd Short. 2007. *Peirce's theory of signs*. Cambridge University Press.

Belkiss Souayed, Sarah Ebling, and Yingqiang Gao. 2025. Template-based text-to-image alignment for language accessibility a study on visualizing text simplifications. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025)*, pages 1–18, Suzhou, China. Association for Computational Linguistics.

Kaiyue Sun, Rongyao Fang, Chengqi Duan, Xian Liu, and Xihui Liu. 2025. T2i-reasonbench: Benchmarking reasoning-informed text-to-image generation. *arXiv preprint arXiv:2508.17472*.

Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. 2024. Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Stephen Tratz and Eduard Hovy. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 678–687, Uppsala, Sweden. Association for Computational Linguistics.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, and 1 others. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. 2024. On verbalized confidence scores for llms. *arXiv preprint arXiv:2412.14737*.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Visual De-Noising System Prompt

To ensure reproducibility of the Symbolic Anchors (v_{id}), we provide the exact system instructions used to transform the SemEval-2025 dataset images.

> **Task:** Analyze the input image and transform it into a minimalist, abstract symbolic icon.
>
> **1. Conceptual Instructions (De-Noising):**
>
> - **Identify the Core Essence:** Determine the fundamental meaning or action of the image. Ignore specific details, individuals, or environments.
> - **Abstract & Merge (Metonymy):** If the image contains multiple elements forming a narrative, distill them into a single, unified glyph that represents the entire concept (e.g., instead of "person watching loud TV," create a symbol for "intense viewing").
> - **Remove Context:** Eliminate all background elements, environments, and secondary objects.
>
> **2. Stylistic Instructions (Flat Iconography):**
>
> - **Geometric Reconstruction:** Rebuild the concept using only pure geometric primitives (perfect circles, squares, triangles, and clean, uniform arcs). Avoid organic or sketchy lines.
> - **Strict Flat Design:** There must be absolutely zero gradients, shadows, textures, or lighting effects. All colors must be solid flats.
> - **Bold Outlines:** Encase all major elements in thick, uniform black outlines.
> - **Limited Palette:** Restrict the color palette strictly to Black, White, and a maximum of two highly contrasting solid accent colors derived from the most prominent color in the input image.
> - **Composition:** The final output should be a clean, centered logo icon on a plain white background.

B Human Verification Protocol

We provide full details of the human-in-the-loop verification procedure used to quality-control the DIVA iconographic renderings.

Annotators. Three independent internal annotators participated in the verification process. All annotators were graduate students in computational linguistics and were familiar with idiomatic expressions and noun-compound interpretation.

Annotation interface. For each noun compound instance, annotators were shown:

1. the target noun compound text (e.g., *Eye Candy*);
2. the original high-fidelity image from ADMIRE, used as a semantic reference;
3. four candidate iconographic images generated by our abstraction pipeline.

Task instructions. Annotators were instructed to evaluate each candidate along two dimensions:

1. **Semantic preservation:** whether the candidate preserved the intended meaning of the noun compound;
2. **Stylistic conformity:** whether the candidate satisfied the required iconographic constraints (flat design, geometric composition, limited palette, minimal background detail).

Annotators selected the best candidate only if at least one image satisfied both criteria. This was *not* a forced-choice task: if all four candidates failed to preserve the intended meaning or violated the stylistic constraints, annotators rejected the entire batch.

Regeneration loop. Rejected batches triggered a new generation cycle. Four fresh candidates were produced and submitted for re-annotation under the same protocol. This generation–verification loop continued until one candidate passed the semantic and stylistic checks.

Stopping criterion. An instance was finalized only when at least one candidate was judged acceptable under both criteria and selected as the best rendering for that noun compound and target sense.

Statistics. Across the full dataset (200 test instances $\times$ 5 image categories = 1,000 items), the first-round acceptance rate was 78.4%. The remaining 21.6% required one regeneration cycle; fewer than 3% required two or more cycles. Inter-annotator agreement (Fleiss' κ) on the accept/reject decision was 0.81, indicating almost perfect agreement.

C Qualitative Failure Cases

We provide representative examples for the two main failure types discussed in Section 6.3. In each case, we compare model behavior under high-fidelity and iconographic conditions and describe the observed error pattern.

C.1 Failure Type I: Hyper-fidelity Literal Pull

This failure occurs when visually rich images over-emphasize constituent objects, textures, or scene details, causing the model to prefer a literal interpretation over the intended idiomatic one.

Example 1: *Eye Candy*. In the high-fidelity condition, the image contains visually salient object-level cues associated with literal *eyes* and entertainment-related artifacts. Several models anchor on these literal objects and assign higher semantic fit to the literal rendering than to the idiomatic one. Under iconographic abstraction, these distractor cues are suppressed, and stronger generative models show a substantial reduction in signed literal bias.

Example 2: *Night Owl*. In the high-fidelity condition, the presence of an owl-like figure and nighttime context attracts the model toward a literal interpretation. After simplification into a schematic representation of the intended idiomatic meaning, generative models reduce this literal preference, although some discriminative models still retain a positive bias toward the literal reading.

C.2 Failure Type II: Semantic Drift under Partial Abstraction

This failure occurs when a model no longer commits to a fully literal reading, but still fails to align with the intended idiomatic meaning. Instead, it drifts toward visually or semantically adjacent alternatives.

Example 1: *Paper Tiger*. After iconographic abstraction, the model no longer focuses on photorealistic animal detail, but still overweights surface-level cues associated with *paper* or *tiger* independently. As a result, it selects a semantically nearby but incorrect candidate rather than the intended idiomatic depiction.

Example 2: *Cold Shoulder*. In this case, abstraction reduces the influence of high-fidelity scene detail, yet the model remains attracted to concrete visual elements associated with bodily posture or temperature. The resulting prediction reflects partial semantic overlap rather than the conventional idiomatic meaning.

Summary. Across these cases, the qualitative evidence supports the quantitative results in Table 1. High-fidelity images tend to amplify literal visual attraction, while iconographic abstraction reduces this pull. However, simplification alone does not fully solve idiomatic grounding: weaker models, especially discriminative encoders, may still drift toward semantically related but incorrect interpretations.

D External Validation and Statistical Details

Human correlation study. To validate that Δ captures a meaningful underlying phenomenon, we conducted a blinded human evaluation. We randomly sampled 100 noun compounds (200 images) from our dataset. Three independent annotators rated the idiomatic alignment of each image on a 1–5 Likert scale, strictly blinded to model outputs, architectures, and automated Δ scores.

Inter-annotator agreement was substantial (Fleiss' $\kappa = 0.76$). We computed "Human Δ" as the absolute difference in mean human ratings between literal and idiomatic images for each NC. Spearman rank correlations between Human Δ and automated Δ: Discriminative/Cosine ($\rho = 0.64$, $p < 0.001$), Generative/Log-Prob ($\rho = 0.69$, $p < 0.001$), and Proprietary/Confidence-Score ($\rho = 0.73$, $p < 0.001$).

Confidence score vs. choice-frequency validation. GPT-5: $\rho = 0.82$; Claude 4.5 Sonnet: $\rho = 0.79$. Both signals capture related but not identical aspects of preference, confirming the confidence-based Δ values in Table 1.

E Examples

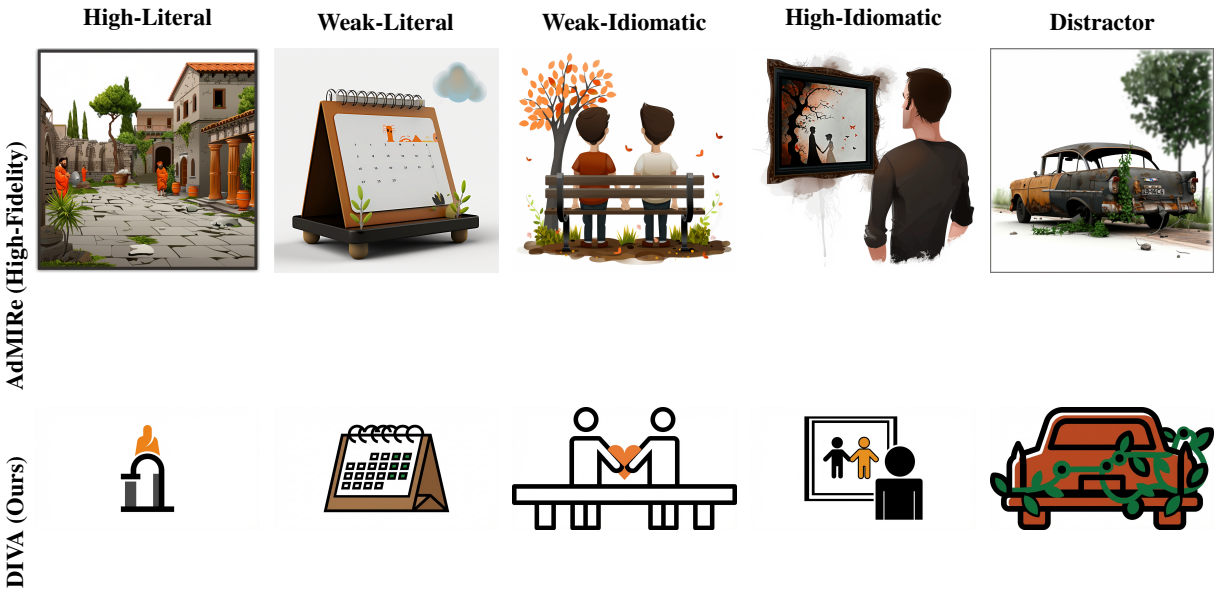

Figure 4: **Iconographic Abstraction in Action (AdMIRe vs. DIVA).** Top Row: The original high-fidelity images from AdMIRe, where high-frequency texture creates "semiotic noise." Bottom Row: Our corresponding DIVA icons. By systematically simplifying the images across the full semantic spectrum (from Literal to Idiomatic), DIVA provides a clean, structure-aware testbed for multimodal reasoning.