The Expense of Seeing: Attaining Trustworthy Multimodal Reasoning Within the Monolithic Paradigm

Karan Goyal
IIIT Delhi, India
karang@iiitd.ac.in

Abstract

The rapid proliferation of Vision-Language Models (VLMs) is often framed as enabling unified multimodal understanding, but rests on an under-examined assumption: that current VLMs faithfully synthesise multimodal data. We argue they often do not, and that this gap reflects a trustworthiness problem in the dominant Vision Encoder-Projector-LLM paradigm. Rather than extracting grounded knowledge from visual inputs, state-of-the-art models frequently exhibit functional blindness, exploiting strong language priors to bypass severe visual representation bottlenecks. We challenge the conventional methodology of multimodal evaluation, which relies on data ablation or new dataset creation and therefore conflates dataset biases with architectural incapacity. We propose an information-theoretic departure: the *Modality Translation Protocol*, which preserves a sample's semantic payload while translating its modality rather than ablating it. From this, we derive three metrics—the Toll (*ToS*), Curse (*CoS*), and Fallacy (*FoS*) of Seeing—culminating in the *Semantic Sufficiency Criterion (SSC)*, a single condition that detects, quantifies, and localises multimodal failures to the visual encoder or the projection head. We further hypothesise a *Divergence Law of Multimodal Scaling*: as language backbones grow, the penalty of the visual bottleneck may widen rather than close. Examining the closest-matching existing benchmarks, motivates our central call: the construction, through *Semantic Equivalence Engineering*, of a large, diverse dataset whose modalities are isomorphic by construction. We frame this not as an experiment awaiting compute, but as a mission for the community: to move beyond "multimodal gain" and build the substrate on which faithful multimodal architectures can be designed.

1 The Illusion of Multimodal Synthesis

The trajectory of multimodal Natural Language Processing has reached a critical inflection point. The frontier of NLP is no longer the modelling of isolated text corpora; it is the construction of unified models that ingest and synthesise disparate, high-dimensional information from text, images, and structured data to perform faithful cross-modal reasoning and decision-making. At the core of this frontier sits the Vision-Language Model (VLM), predominantly governed by the monolithic Vision Encoder-Projector-LLM architectural paradigm (Bordes et al., 2024).

The prevailing assumption within the NLP and multimodal learning community is that these models natively integrate visual and textual streams to execute Compositional Visual Reasoning (CVR) (Ke et al., 2026). Yet, as VLMs are increasingly deployed in high-stakes applications ranging from autonomous medical diagnostics (Hartsock and Rasool, 2024; Zhong et al., 2025) to financial time-series forecasting (Khezresmaeilzadeh et al., 2025), a notable epistemic fragility has been documented. Highly parameterised, state-of-the-art models frequently achieve superficial benchmark supremacy by largely ignoring the visual input. Instead, they exhibit a modern *Clever Hans* effect, executing complex statistical guessing via deeply ingrained text priors housed within their massive Large Language Model (LLM) backbones (Choi et al., 2024; Chen et al., 2024, 2026). A latest work, BabyVision (Chen et al., 2026), shows that SOTA VLMs consistently fail on basic visual tasks that humans, even 3-year-olds, can solve effortlessly. It released a dataset benchmark designed to assess core visual abilities independent of linguistic knowledge for VLMs. MMVP (Tong et al., 2024) released a dataset benchmark to probe visual limitations and found that models fail to distinguish images with clear perceptual differences. ConMe (Huang et al., 2024) released compositional reasoning benchmark to produce 'hard CR Q&A'.

Recent initiatives within the representation learning community, such as MathVerse (Zhang et al.,

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 5604–5614
November 6-10, 2026 ©2026 Association for Computational Linguistics

2024), SeePhys (Xiang et al., 2025) and MMStar (Chen et al., 2024), have attempted to uncover and quantify this phenomenon. MathVerse introduced problem versions with varying visual-textual information balance and observed that some VLMs achieve higher accuracy when visual input was removed entirely. SeePhys extended this to Physics, distinguishing "vision-essential" from "vision-optional" problems. MMStar proposed heuristic metrics like Multimodal Gain and Multimodal Leakage, and a manually vetted vision-indispensable dataset for multimodal assessment. However, we assert that these approaches have a methodological limitation. By ablating (removing) data to test models, they successfully expose *dataset biases* but fundamentally fail to isolate *architectural representation bottlenecks*. We cannot map the limits of a model's information extraction and processing prowess by measuring what happens when information is artificially deleted.

In an era where modern reasoning engines achieve increasingly strong symbolic logic, a difficult question arises: ***What if vision is no longer a value-add for multimodal reasoning in its present form, but an active architectural liability?*** This paper introduces a framework to systematically diagnose, quantify, and address these integration failures. We shift the paradigm from observing macroscopic, dataset-induced heuristics to establishing principled, sample-level diagnostic criteria for trustworthy multimodal evaluation. **We propose the field shift from measuring additive Multimodal Gain or creating new dataset benchmarks towards diagnosing the *Expense of Seeing*** — a step toward constructing genuinely faithful multimodal models within the monolithic paradigm.

2 Challenging Existing Assumptions: The Crisis in Evaluation

To build trustworthy multimodal NLP systems, our evaluation metrics must isolate the *source* of a model's predictive power. The standard approach of evaluating multimodal capabilities currently relies heavily on the paradigm of data ablation.

2.1 The Flaws of Multimodal Gain and Leakage

Consider the formulation of Multimodal Gain (MG), which measures the difference in accuracy when a model is given both vision and text (S_v) versus text alone (S_{wv}): $MG = S_v - S_{wv}$. Similarly, Multimodal Leakage (ML) assesses leakage by comparing the VLM's text-only performance against its underlying base LLM (S_t): $ML = \max(0,\ S_{wv} - S_t)$.

We assert that these metrics are limited for evaluating modern multimodal models.

1. **Biased Estimators:** ML is limited as a global estimator. By utilising a $\max$ function, it routinely fails to account for destructive interference scenarios where the multimodal alignment training process degrades the base LLM's inherent reasoning capabilities ($S_{wv} < S_t$).
2. **The Ablation Fallacy:** MG does not measure faithful integration; it measures the leverage of an additional signal under conditions of artificial starvation. From an information-theoretic perspective, if we deprive a model of required information (by deleting the image), any subsequent failure cannot be definitively attributed to an architectural inability to process vision. Measuring what happens when information is removed does not directly characterise a model's capacity to extract information when it is present.

2.2 The Necessity of a Paradigm Shift

To build trustworthy systems capable of synthesising powerful combinations of visual and textual data, we must move beyond data ablation and the race to new dataset creation. We must isolate *architectural bottleneck* from *dataset bias*. If the research community continues to use ablative metrics, we risk deploying models that rely on language priors rather than grounded visual evidence, leading to high-cost failures in real-world applications.

2.3 Existing Evidence: Documented Failures of Visual Reliance

Although our protocol is the methodological contribution of this paper, the underlying phenomena that motivate it are already empirically documented across multiple recent benchmarks. We marshal this evidence here to ground the framework in concrete numbers rather than abstract claims.

MMStar reports that GeminiPro achieves 42.7% on MMMU *without any visual input*, outperforming the random-choice baseline by approximately 24% averaged across six benchmarks. Sphinx-X-MoE reaches 43.6% on MMMU without accessing images — more than double its 17.9% pure-LLM-backbone score, indicating that vision-language

training partly inflates text-only performance via memorisation rather than visual grounding.

MathVerse provides perhaps the sharpest finding: Qwen-VL-Max and InternLM-XComposer2 achieve over 5% *higher* accuracy when visual input is removed entirely, i.e., visual input is not merely unhelpful but actively harmful to these models.

SeePhys extends this to vision-essential physics problems and reports that even Gemini-2.5-Pro and o4-mini fall below 55% accuracy, with several leading models below 45%. Most strikingly, DeepSeek-R1 (a text-only LLM) achieves 42.2%, approximately matching o3-mini with multimodal access at 40.3%. A text-only LLM matching a multimodal model on a benchmark explicitly designed to require visual reasoning is direct evidence of the scaling pattern we formalise as the *Divergence Law* in Section 7.

BabyVision demonstrates that on basic visual primitives solvable by three-year-old children, the strongest model (Gemini3-Pro-Preview) scores 49.7% against an adult baseline of 94.1%, and falling below six-year-old human performance by around 20%. MMVP shows most MLLMs perform below random chance on perceptually distinct image pairs that CLIP encodes as similar, exposing a fundamental encoding bottleneck rather than a reasoning failure.

Three patterns recur across these findings: (i) text-only baselines approach or exceed multimodal performance on tasks designed to require vision; (ii) scaling the LLM backbone tends to amplify rather than reduce reliance on text priors; (iii) the visual encoder itself, not just the integration layer, exhibits systematic limitations. These three patterns are precisely what ToS, CoS, and the sign of FoS are designed to disentangle.

2.4 Operationalising the Paradigm Shift: A New Diagnostic Agenda

To successfully transition from observing dataset-induced heuristics to diagnosing fundamental architectural bottlenecks, the NLP and multimodal learning community must align around a new empirical standard. Based on the necessity of preserving semantic equivalence, we propose that the future evaluation of multimodal models must be anchored by six operational, highly testable research questions:

- **[RQ1] The Baseline Penalty:** Do current VLM architectures incur a systematic, quantifiable performance penalty when extracting knowledge from visual inputs compared to processing equivalent (and potentially even lossy) symbolic textual representations?
- **[RQ2] The Architectural Origin:** Can we mathematically distinguish and isolate whether a model's inefficiency originates in the visual encoder (an incapacity to read visual features) or the cross-modal projection head (an incapacity to fuse separate semantic streams)?
- **[RQ3] Semantic Asymmetry:** Do VLMs exhibit semantic inconsistency across modalities? If provided with equivalent information in symbolic textual and symbolic visual forms, does the architecture asymmetrically penalise the act of "seeing" rather than "reading"?
- **[RQ4] The Scaling Paradox:** Within a fixed architectural family, what does drastically increasing parameter scale actually achieve? Does scaling the underlying language engine alleviate the visual bottleneck, or paradoxically exacerbate it?
- **[RQ5] The Universal Constraint:** Can we design a singular, mathematically sound criterion that detects, quantifies, and localises these multimodal failures across any given architecture?
- **[RQ6] Dataset Agnosticism:** Can a diagnostic toolkit definitively prove that an integration failure is caused by an architectural bottleneck rather than dataset bias, eliminating the field's reliance on data ablation and specially vetted "vision-indispensable" benchmarks?

3 The Modality Translation Protocol & High-Stakes Case Studies

We propose a new methodological approach: The Modality Translation Protocol. Instead of deleting information to test a model, this protocol preserves the exact semantic payload of a data sample while translating its modality across different representation states.

Let $S(\cdot)$ denote the primary evaluation metric (e.g., Accuracy, Exact Match) of a model $\mathcal{M}$ on a given task. For any single multimodal data sample, we define three distinct modulations:

1. S_{Full} **(Standard VLM):** Evaluated with standard visual input V and textual input T.

$$S_{Full} = S(\mathcal{M}(V, T)) \tag{1}$$

2. S_{SymT} **(Symbolic Text Ceiling):** The visual input V is replaced by a task-sufficient symbolic text representation V_{label}. This measures the

LLM's task-relevant reasoning capacity given symbolic access to the same evidence the visual modality is meant to convey.

$$S_{SymT} = S(\mathcal{M}(\emptyset, T + V_{label})) \quad (2)$$

3. S_{SymV} **(Symbolic Vision):** The textual question T is rendered perfectly as text-within-an-image T_{img}, forcing the model to read solely via its visual encoding pipeline without discrete text tokens.

$$S_{SymV} = S(\mathcal{M}(V + T_{img}, \emptyset)) \quad (3)$$

We illustrate how this protocol exposes architectural failures across three critical high-stakes domains.

- **Case Study 1: Financial Time-Series Mining:** A VLM analyses a candlestick chart to predict a breakout. S_{SymT} replaces the chart with perfect OHLC tabular text. If S_{SymT} yields 95% accuracy but S_{Full} yields 60%, the underlying LLM demonstrates competence at financial reasoning, but the visual encoder actively bottlenecks knowledge extraction.
- **Case Study 2: Trustworthy Medical Diagnostics:** A VLM evaluates a chest X-Ray, prompted with clinical notes: *"Patient has a 30-year history of smoking."* Driven by text priors, S_{Full} predicts cancer. S_{SymT} replaces the image with ground-truth symbolic findings: *"Clear lungs."* If S_{SymT} correctly predicts "Healthy" but S_{Full} hallucinates cancer, we expose a serious text-prior override. We emphasise that "Clear lungs" is a findings-level observation in the standard radiology-reporting sense, distinct from the diagnostic conclusion the model produces.
- **Case Study 3: Molecular Graph Mining for Drug Discovery:** A VLM screens a 2D molecular structure for toxicity based on visual topology and textual properties. S_{SymV} removes the text prompt, rendering the text directly into the 2D molecule image. If S_{Full} underperforms S_{SymV}, the projection head cannot align continuous visual coordinate spaces with discrete token spaces. And if it outperforms, then it indicates inefficiency in visual encoding.

3.1 Constructing V_{label}: Task-Conditional Sufficiency

The protocol's diagnostic power rests on a precise construction of V_{label}. We do not require V_{label} to be a strictly lossless information-theoretic translation of V, which is generally unachievable: a chest X-ray contains continuous pixel-level information that no finite symbolic description fully captures. Instead, we define V_{label} as *task-sufficient*: it preserves all task-relevant discriminative information that an idealised observer could extract from V to answer the task. Under this definition, S_{SymT} characterises the LLM's reasoning capacity on the task given symbolic access to the same evidence, and $ToS > 0$ admits a clean interpretation: *given equivalent task-relevant information, the model performs worse when that information is delivered visually than symbolically.*

This formalisation also clarifies how V_{label} should be constructed in practice. In a large and important class of multimodal tasks, the visual modality is itself a rendering of underlying structured data; here V_{label} recovers the pre-rendering symbolic form. The three case studies above instantiate this pattern: OHLC tabular text underlies the candlestick chart; ground-truth radiological findings underlie the X-ray label; SMILES strings or atom-bond lists underlie the molecular diagram. Where no pre-rendering form exists, V_{label} is constructed via oracle annotation against the task's ground truth (expert annotation, structured database lookup, or symbolic extraction pipelines).

This scoping is a feature, not a concession. The framework applies cleanly to the broad class of tasks central to scientific document understanding, structured visual reasoning, chart and diagram comprehension, annotated medical imaging, molecular and graph-based reasoning, and document intelligence — precisely the high-stakes settings where trustworthy multimodal evaluation matters most. Open-ended perceptual tasks (e.g., describing the mood of a photograph) fall outside this scope by design, and we make no claim about them.

3.2 A Worked Example: Why Existing Benchmarks Fall Short

To examine whether the protocol can be instantiated on existing data, consider MathVerse, the benchmark whose design comes closest, as it releases multiple problem versions that vary the balance of information between text and image. Two of its versions correspond to our modulations: the *Vision Only* version, in which the problem statement is rendered *into* the diagram with no separate text input, realises S_{SymV}; and the *Text Only* version, in which the problem is stated fully in

text with no diagram, would realise S_{SymT}. Crucially, none of MathVerse's versions corresponds to a clean S_{Full}: the standard dual-stream operating point, in which the genuine image V and the genuine textual question T each carry their natural, non-duplicated share of the payload, is simply not among them. MathVerse's other versions (*Text Dominant*, *Text Lite*, *Vision Intensive*, *Vision Dominant*) either duplicate information across modalities or migrate conditions between them, and so none isolates this canonical point.

This near-miss is itself instructive, and it reflects a difference in design principle. MathVerse's versions trace an *ablation and migration gradient*: they vary how much information is present and in which modality it resides. Our protocol does the opposite — it holds the task-relevant payload fixed and varies only the *modality of delivery*. Because of this difference, the MathVerse versions do not include the three operating points the protocol requires; they sample a different axis. The situation is sharper still in practice: the public MathVerse release does not include the *Text Only* version at all, providing five of the six versions described. As released, it therefore realises only one of our three modulations (S_{SymV}, via *Vision Only*). Since each of ToS, CoS, and FoS is a difference between two modulations, and the release supplies at most one of the two each requires, *none* of the three metrics can be computed from MathVerse as publicly available.

This is not a shortcoming of MathVerse, which was not built to our specification; it is evidence about the state of the field. The single benchmark whose design comes closest to the protocol still defines only two of the three required operating points, releases data for only one, and lies within a single narrow niche (visual mathematics). The other benchmarks fare no better: SeePhys tags problems as vision-essential but never renders the question into the image in the data release, so it admits no S_{SymV}; MMStar's per-sample data is ablation-based (text-removed), precisely the methodology we argue against, and likewise yields no S_{SymV} or S_{SymT}. These benchmarks motivate the protocol; none instantiates it. Diagnosing the Expense of Seeing across the breadth of SOTA capabilities — and, more importantly, using it to *design* faithful models — requires at least one large, diverse dataset whose $(V, V_{label}, T, T_{img})$ tuples are isomorphic by construction and which supplies all three operating points by design. No such dataset exists today. Building it is the mission we set out in Section 8.

4 True Quantifiers of Visual Reception

Utilising the Modality Translation Protocol, we define three novel metrics that act as quantifiable indicators of multimodal knowledge bottlenecks. These metrics move the community beyond asking *whether* a model works, to diagnosing *why, how much, and where* multimodal reasoning breaks down.

4.1 ToS: Toll of Seeing

The actual expense the VLM bears to process the visual modality, operationalising the baseline penalty of integration:

$$ToS = S_{SymT} - S_{Full} \tag{4}$$

Diagnostic Interpretation [RQ1]: Ideally, $ToS \leq 0$. If $ToS > 0$, we identify an architectural inefficiency in visual encoding and/or integration. The model incurs a systematic performance penalty when processing visual input compared to its equivalent (and potentially even lossy) symbolic textual representation. Vision acts as a toll on the LLM's inherent reasoning capacity.

4.2 CoS: Curse of Seeing

The asymmetric penalty of processing information across different modalities:

$$CoS = S_{SymT} - S_{SymV} \tag{5}$$

Diagnostic Interpretation [RQ3]: Ideally, $CoS \leq 0$. If $CoS > 0$, the architecture exhibits semantic inconsistency. It reveals an asymmetric penalisation of *seeing* rather than *reading* equivalent information (and potentially even lossy). A faithful multimodal model should treat semantically equivalent inputs symmetrically; $CoS > 0$ indicates the model is systematically biased against non-textual knowledge extraction.

4.3 FoS: Fallacy of Seeing

The third metric of our diagnostic resolution, distinguishing the exact origin of the architectural bottleneck:

$$FoS = S_{Full} - S_{SymV} \tag{6}$$

The asymmetry between S_{Full} (dual-stream) and S_{SymV} (single-stream) is intentional: it is precisely this contrast that enables the sign of FoS to localise failure to encoder versus projector, since the

two conditions hold the semantic payload constant while varying how the model must fuse it.

Diagnostic Interpretation [RQ2]: Mathematically, $FoS \equiv CoS - ToS$. However, we must explicitly define and evaluate FoS separately because it diagnoses a distinct failure mode. Ideally, $FoS = 0$. Humans process S_{Full} and S_{SymV} equally well without loss of fidelity. If $FoS \neq 0$, the architecture fails to process the same lossless semantic payload consistently across these conditions. Crucially, the sign of FoS reveals two distinct, mutually exclusive failure modes:

- **The Positive Collapse Mode ($FoS > 0$):** Indicates an inefficiency in visual encoding. The model struggles to read and extract text when it is rendered purely as an image, indicating the vision encoder (e.g., the ViT) lacks the granular spatial resolution to extract symbolic features.
- **The Negative Collapse Mode ($FoS < 0$):** Indicates an inefficiency in visual integration. The model performs paradoxically better when forced into a single visual modality (S_{SymV}) than when handling separate visual and textual streams (S_{Full}). This isolates the failure to the cross-modal projection head, indicating it cannot meaningfully fuse separate modalities in the latent space.

5 Semantic Sufficiency Criterion (SSC)

Together, these metrics establish a mandatory mathematical condition for semantically grounded, faithful multimodal modelling. We define the Semantic Sufficiency Criterion (SSC):

$$SSC : \max(ToS, CoS, |FoS|) = 0 \quad (7)$$

5.1 A Diagnostic Constraint, Not an Immediate Goal

Crucially, the SSC is not treated as an immediately achievable performance target for current models, but as a stringent diagnostic constraint **[RQ5]**. Violations of the SSC (where $SSC > 0$) quantify the exact magnitude and location of a VLM's failure. The application of the absolute value $|FoS|$ is necessary to ensure that both the positive (encoding) and negative (integration) failure modes are captured and audited.

5.2 Universal Dataset Applicability

Because the protocol never ablates either signal (in contrast to MMStar), failures detected by SSC can be attributed to *architectural bottlenecks* rather than dataset-induced artifacts. Consequently, this diagnostic toolkit provides a great advantage for the community: it can be applied to any regular dataset **[RQ6]**. We no longer need to rely on specially vetted datasets to test models. The SSC identifies and quantifies architectural violations across multimodal tasks.

6 Discussing FoS as a Third Metric and SSC

We note that $FoS = CoS - ToS$ makes it seem like FoS is redundant, and we agree this in numerical terms and have already stated in Section 4.3. But we defined FoS separately for two reasons.

Firstly, since this work is about trustworthy multimodal reasoning, we believe that presentation matters as it informs interpretation. The arithmetic ($CoS - ToS$) has no interpretive content, whereas the direct comparison ($S_{Full} - S_{SymV}$) is grounded in human perceptual symmetry, i.e., humans process S_{Full} and S_{SymV} equivalently without loss of fidelity. The output is same but only the latter framing carries the diagnostic meaning the paper's mission requires.

Secondly, FoS as ($S_{Full} - S_{SymV}$) is operationally independent of S_{SymT}: it needs only two measurements — the dual-stream and single-stream visual conditions — to answer $RQ2$ (localising failure to encoder vs. projector). Recovering the same quantity via $CoS - ToS$ would require also measuring S_{SymT}, i.e., constructing V_{label}, which is precisely what makes the closest-matching existing benchmark unable to compute the metric (§3.2). FoS in our formulation is the minimally-sufficient quantity for $RQ2$, whereas the algebraic reroute is not.

Importantly, one may overlook to characterise SSC as "really just one constraint." This is not the case: $\max(ToS, CoS, |FoS|) = 0$ is equivalent to the conjunction $ToS \leq 0$ AND $CoS \leq 0$ AND $CoS = ToS$, and no one of these three implies the other two. The SSC therefore imposes three algebraically independent conditions on the underlying triple $(S_{Full}, S_{SymT}, S_{SymV})$, whether written using FoS or not. The framing that these conditions can be stated using only ToS and CoS (via $CoS = ToS$ in place of $|FoS| = 0$) is correct — but that is a statement about symbol choice, not about the number of constraints. The reasons for retaining FoS as a separately-defined metric —

perceptual grounding and operational minimality for $RQ2$ — also remain in the SSC definition as stated above.

7 The Divergence Law of Multimodal Scaling

The dominant orthodoxy in scaling multimodal models dictates a simple heuristic: scaling the compute and parameter counts inherently resolves multimodal alignment challenges. The industry operates on the foundational assumption that concatenating progressively larger Vision Transformers (ViTs) to progressively larger Large Language Models (LLMs) will organically yield faithful multimodal synthesis. **We argue this assumption is unlikely to hold.**

By operationalising our diagnostic agenda, specifically addressing **[RQ4]**, we posit that the current architectural paradigm is structurally limited in achieving true multimodal reasoning. Scale-driven alignment may mask, rather than resolve, a structural failure.

7.1 The Mechanics of the Representation Bottleneck

To understand why scaling fails, we must examine the information-theoretic capacity mismatch inherent in the Vision Encoder-Projector-LLM paradigm. Visual data manifolds (e.g., the topology of a molecular graph or the high-frequency pixel variations in a medical scan) are continuous, high-dimensional, and dense. Conversely, textual token spaces are discrete, sequential, and highly compressed.

Current architectures force the entirety of the visual manifold to pass through a narrow, fixed-capacity cross-attention or projection head to be translated into text-like embeddings. As the LLM backbone scales, its capacity to execute complex symbolic logic and leverage statistical priors outpaces the projection head's ability to faithfully translate the visual manifold. We term this structural choke point the *information compression penalty*. Regardless of the reasoning engine's capacity, its visual pipeline remains a low-bandwidth, lossy conduit.

7.2 Formulating the Divergence Law

We hypothesise the Divergence Law of Multimodal Scaling to analyse how this penalty manifests as the models grow. As model parameters scale by orders of magnitude, macroscopic benchmark accuracy (S_{Full}) will increase. This rising metric is conventionally taken as evidence of progress. Under the Modality Translation Protocol, however, it may overstate true multimodal capability.

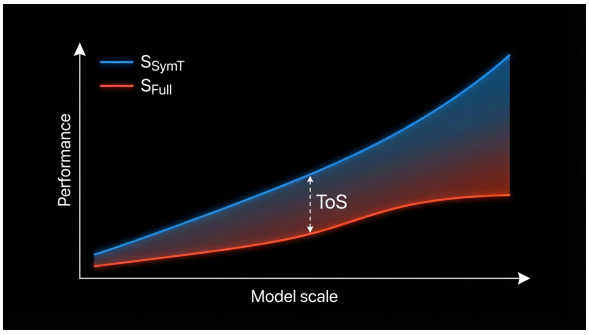


Figure 1: The Divergence Law of Multimodal Scaling. As Model Scale in Parameters (x-axis) increases, the LLM's true reasoning capability (S_{SymT}, logarithmic curve) rises sharply, while the macroscopic benchmark performance (S_{Full}) scales more shallowly. The expanding shaded region between the two curves represents a growing Toll of Seeing (ToS).

Because the visual projection bottleneck cannot scale its representational bandwidth proportionally to the LLM's cognitive leap, the model's true symbolic reasoning ceiling (S_{SymT}) rises at a faster rate compared to S_{Full}. Consequently, we project that the Toll of Seeing will increase proportionally with the model scale, as shown in Figure 1.

7.3 The Illusion of Capability

The implications of this Divergence Law for multimodal models are significant. If ToS widens as scale increases, it implies that the relative cost of the visual modality grows with architecture scale.

Scaling may not resolve multimodal reasoning failures; instead, it can amplify the LLM's ability to exploit text priors, masking the underlying visual bottleneck. As the LLM's text-only capability grows, the incentive to bypass a weaker visual encoder via language priors grows with it. Continued scaling within the current monolithic paradigm is therefore unlikely to yield faithful multimodal synthesis on its own, and may instead amplify reliance on language priors.

8 A Roadmap: From a Diagnostic Constraint to SSC-Guided Architectures

The introduction of the Semantic Sufficiency Criterion (SSC) represents a paradigm shift that

uniquely targets a core mandate of the NLP and multimodal learning community. Historically, the field has relied on *data scavenging*, i.e., scraping massive, noisy repositories of loosely correlated image-text pairs or even the manually intensive image-text pair creation, and relying on the false hope that blind, billion-parameter next token prediction will force emergent alignment.

The Divergence Law of Multimodal Scaling suggests that passive scaling alone is insufficient. However, this does not mean we must abandon the monolithic paradigm (Vision Encoder-Projector-LLM). Instead, we propose that the community elevate the SSC from a passive diagnostic constraint into an *active architectural blueprint and training objective*. We urge the community to pioneer four directions to support this new paradigm, aiming to improve compositional visual reasoning within monolithic VLMs:

8.1 Semantic Equivalence Engineering (SEE)

The central infrastructural mission this paper advances is the construction of a large, diverse dataset built on strict modality isomorphism. As Section 3.2 establishes, even the existing benchmark whose design comes closest supplies only one of the protocol's three operating points and lies within a single niche, so neither faithful model *design* nor rigorous *evaluation* of general-purpose multimodal models is possible without a substrate that does not yet exist: a corpus whose tuples $(V, V_{label}, T, T_{img})$ are guaranteed equivalent in task-relevant mutual information, spanning the breadth of real multimodal capabilities rather than a single niche. SEE is the program to build it. Utilising oracle-driven symbolic extraction — recovering structured pre-rendering forms (tables, findings, graphs, scene descriptions) at scale — the community must engineer the isomorphic tuples that simultaneously (i) supply the strict S_{SymT} baselines required to use SSC as a training objective, and (ii) provide the diverse, non-approximate measurements required to evaluate SOTA models faithfully. This single dataset is therefore the engine for both halves of the agenda: it powers SSC-guided design and SSC-grounded evaluation at once. We frame its construction not as an experiment to be run on present-day tooling, but as a foundational community undertaking — the kind of mission this theme track exists to initiate.

8.2 SSC as an Objective Function

The multimodal learning community should move beyond static benchmark leaderboards and cross-entropy loss. We envision the development of *Information-Theoretic Alignment Optimisation*. Researchers can formulate the Toll (ToS), Curse (CoS), and Fallacy (FoS) of Seeing into dynamic regularisation penalties during the VLM pre-training and alignment phases. By penalising the model when S_{Full} diverges from S_{SymT} or S_{SymV}, the loss landscape can be shaped to discourage over-reliance on text priors and encourage faithful visual feature extraction and alignment. One such way is to compute ToS, CoS and FoS on held-out probe batches at intervals during training and use their divergences from zero as regularisation penalties on the loss landscape — analogous to how contrastive losses or KL-divergence terms enter standard training objectives. The metrics themselves remain inference-time quantities; their use during training is as an auxiliary signal that shapes parameter updates, not as a per-step evaluation of task accuracy.

8.3 Architecting the Faithful Monolithic Paradigm

By utilising FoS as a localised gradient signal, we can address the information compression penalty inherent in monolithic VLMs. The framework in fact prescribes a two-mode architectural response driven by the sign of FoS. For example, when Negative Collapse Mode ($FoS < 0$) is detected during training, it signals integration failure. Training systems can use this real-time feedback to dynamically route, expand, or regularise the bandwidth of the projection head. And when $FoS > 0$ (encoding failure, vision encoder), the actionable response is different — higher input resolution, finer patch tokenisation, or richer visual token grids — since the deficit lies in the encoder's capacity to extract symbolic features rather than in cross-modal fusion. By using the SSC to guide architectural decisions during training, compositional visual reasoning may be achievable within the monolithic paradigm, aligning continuous visual manifold with discrete token space.

One practical concern may seem to arise that in standard VLM pipelines, the ViT is often pre-trained and frozen while the projection head and LLM are updated jointly. This is precisely the setting in which our proposal is most naturally applied.

The trainable parameters, i.e., projection head and LLM adapters, are where the $FoS < 0$ localisation points, and FoS-guided interventions (routing, bandwidth expansion, regularisation) act on those parameters directly through the standard training loss. The $FoS > 0$ mode (encoder failure) is addressable within the same pipeline through changes to encoder configuration prior to training, e.g., input resolution, patch granularity, or the choice of pretrained encoder itself — none of which requires jointly training the encoder alongside the LLM. Both modes are therefore compatible with the standard training pipeline, although completely new training strategies can be conceptualised to reap maximum benefit.

8.4 Dynamic SSC Auditing Architectures

We envision the development of Dynamic SSC Auditing Engines—autonomous, data modulating systems deployed in production. These systems will continuously perturb incoming visual data streams, translating them into S_{SymT} and S_{SymV} on the fly to monitor a deployed VLM's Expense of Seeing in real time. This ensures that a model's knowledge extraction remains faithful under distributional shifts in a high-stakes inference scenario.

9 If Successful: Implications for Faithful Multimodal Models

If the multimodal NLP community embraces the "Expense of Seeing" framework, it could meaningfully shift practice in four areas:

9.1 Redefining State-of-the-Art Benchmarking

We will witness a clearer separation between models that genuinely use visual input and those that do not. By operationalising the SSC, the community will systematically identify and deprioritise models that disguise textual statistical guessing as visual reasoning. We will no longer celebrate a model that achieves 80% visual accuracy if its symbolic text ceiling (S_{SymT}) is 95%. The resulting 15% *Toll of Seeing* will be recognised as a representation bottleneck, rather than celebrated as a benchmark win.

9.2 Achieving True CVR Capability via Monolithic Paradigm

A successful outcome would be reliable compositional visual reasoning within the monolithic paradigm. By enforcing SSC constraints during the training loop, monolithic architectures may be designed such that visual and textual representations share a more symmetric latent space. This would suggest the monolithic paradigm is not inherently flawed; rather, the historical reliance on metrics like accuracy and unconstrained, ablative training data was blinding it.

9.3 Implications for Regulating Multimodal AI

As we push toward foundation multimodal models entrusted with high-stakes decisions (e.g., autonomous medical diagnostics, algorithmic trading, and response management systems), trust cannot be heuristic; it must be provable. The SSC could potentially inform regulatory standards for Trustworthy Multimodal AI. One could imagine future regulatory frameworks requiring developers to demonstrate that $\max(ToS, CoS, |FoS|) \approx 0$, indicating that the model extracts truth faithfully without asymmetric modality bias.

9.4 Rethinking Scaling Priorities

This framework offers an alternative to purely compute-driven scaling that currently dominates the AI industry. We will shift the current infrastructure race away from merely maximising parameter counts and next token prediction. Success means the community broadens how it measures multimodal capability: true capability will no longer be measured by how much data a model can ingest, but by its modality symmetry.

By transitioning from asking *"Are models working?"* to establishing a rigorous, trustworthy diagnostic constraint that directly guides architectural training, the community will lay a foundation for evaluating multimodal models more rigorously.

Limitations

This paper proposes a diagnostic framework rather than an empirical study, and we note four limitations that scope the contribution and inform future work.

Empirical validation. We do not provide a large-scale empirical campaign that computes ToS, CoS, and FoS across many VLMs and tasks. The metrics are introduced and motivated through existing reported findings (Section 2.3) and a structural analysis of MathVerse (Section 3.2). We stress that no existing dataset, including the closest-matching one, can supply these metrics: as we show, even MathVerse defines only two of the three operating points, releases data for only one (S_{SymV}),

and covers a single niche (visual mathematics), so none of ToS, CoS, or FoS is computable from public data. A general empirical program — one that evaluates SOTA general-purpose models or guides their design — cannot be run on present-day data, because it is blocked on a large, diverse, isomorphically-constructed dataset that does not yet exist. Building that dataset is the Semantic Equivalence Engineering mission of Section 8, and is the principal contribution this framework-driven position paper sets out, rather than a measurement deferred to a follow-up. The framework is thus deliberately presented as a mission for the community, not an experiment awaiting compute.

Scope of V_{label}. The construction of V_{label} requires a task-sufficient symbolic form, which is straightforward when the visual modality is a rendering of underlying structured data (charts, molecular diagrams, annotated medical imaging) but more demanding for purely perceptual tasks where no pre-rendering symbolic form exists. We scope the framework's clean applicability to the former class; open-ended perceptual tasks are not covered by our claims.

Divergence Law as hypothesis. The Divergence Law of Multimodal Scaling is a hypothesis with mechanistic motivation, not a theorem. Confirming or refuting it requires scaling-curve experiments across families of models that hold the projection head and visual encoder topology fixed while varying the LLM backbone, which neither current open releases nor academic compute budgets readily support.

S_{SymV} and rendered-text reading. When T is rendered as T_{img}, S_{SymV} measures the vision pipeline's ability to read rendered symbolic content. We treat this as part of visual encoding (since modern ViTs handle rendered text within the same machinery as other visual features) rather than as a separate OCR confound, but readers should note that S_{SymV} on text-heavy renderings probes the encoder's symbolic extraction capacity specifically. Pure pictorial T_{img} rendering (e.g., iconographic encoding) would be a complementary test that we do not pursue here.

Acknowledgement

To Dikshant Kukreja, for the framing of ToS. We used Claude as our AI assistant for polishing the manuscript.

References

Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, Mark Ibrahim, Melissa Hall, Yunyang Xiong, Jonathan Lebensold, Candace Ross, Srihari Jayakumar, Chuan Guo, Diane Bouchacourt, Haider Al-Tahan, and 22 others. 2024. An introduction to vision-language modeling. *Preprint*, arXiv:2405.17247.

Liang Chen, Weichu Xie, Yiyan Liang, Hongfeng He, Hans Zhao, Zhibo Yang, Zhiqi Huang, Haoning Wu, Haoyu Lu, Y. charles, Yiping Bao, Yuantao Fan, Guopeng Li, Haiyang Shen, Xuanzhong Chen, Wendong Xu, Shuzheng Si, Zefan Cai, Wenhao Chai, and 11 others. 2026. Babyvision: Visual reasoning beyond language. *Preprint*, arXiv:2601.06521.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. Are we on the right way for evaluating large vision-language models? In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Dasol Choi, Guijin Son, Soo Yong Kim, Gio Paik, and Seunghyeok Hong. 2024. Improving fine-grained visual understanding in vlms through text-only training. *Preprint*, arXiv:2412.12940.

Iryna Hartsock and Ghulam Rasool. 2024. Vision-language models for medical report generation and visual question answering: A review. *Frontiers in artificial intelligence*, 7:1430984.

Irene Huang, Wei Lin, Muhammad Jehanzeb Mirza, Jacob A Hansen, Sivan Doveh, Victor Ion Butoi, Roei Herzig, Assaf Arbelle, Hilde Kuehne, Trevor Darrell, Chuang Gan, Aude Oliva, Rogerio Feris, and Leonid Karlinsky. 2024. Conme: Rethinking evaluation of compositional reasoning for modern vlms. *Advances in Neural Information Processing Systems*, 37:22927–22946.

Fucai Ke, Joy Hsu, Zhixi Cai, Zixian Ma, Xin Zheng, Xindi Wu, Sukai Huang, Weiqing Wang, Pari Delir Haghighi, Gholamreza Haffari, Ranjay Krishna, Jiajun Wu, and Hamid Rezatofighi. 2026. Explain before you answer: A survey on compositional visual reasoning. *Preprint*, arXiv:2508.17298.

Tina Khezresmaeilzadeh, Parsa Razmara, Mohammad Erfan Sadeghi, Seyedarmin Azizi, and Erfan Baghaei Potraghloo. 2025. Morfi: Mutimodal zero-shot reasoning for financial time-series inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4236–4245.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *Preprint*, arXiv:2401.06209.

Kun Xiang, Heng Li, Terry Jingchen Zhang, Yinya Huang, Zirong Liu, Peixin Qu, Jixi He, Jiaqi Chen, Yu-Jie Yuan, Jianhua Han, Hang Xu, Hanhui Li, Mrinmaya Sachan, and Xiaodan Liang. 2025. Seephys: Does seeing help thinking? – benchmarking vision-based physics reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. 2024. MATHVERSE: Does your multi-modal llm truly see the diagrams in visual math problems? In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VIII*, pages 169–186, Berlin, Heidelberg. Springer.

Zhusi Zhong, Yuli Wang, Jing Wu, Wen-Chi Hsu, Vin Somasundaram, Lulu Bi, Shreyas Kulkarni, Zhuoqi Ma, Scott Collins, Grayson Baird, and 1 others. 2025. Vision-language model for report generation and outcome prediction in ct pulmonary angiogram. *NPJ digital medicine*, 8(1):432.