

DECODE: Overcoming the Perception Bottleneck in Small Vision-Language Models for Computer Science Education

Martina Ianaro(✉) and Maurizio Gabbrielli

University of Bologna, Bologna, Italy
{martina.ianaro3,maurizio.gabbrielli}@unibo.it

Abstract. Small Vision-Language Models (SVLMs) represent a critical frontier for Edge AI in Computer Science Education (CSEd), offering the efficiency and accessibility required for decentralized learning environments. However, the effectiveness of these models is currently undermined by a representational collapse of the vision encoder when processing high-fidelity technical diagrams, pseudocode snippets, and pointer representations. This paper introduces **DECODE**, a training-free framework designed to mitigate this *Perception Bottleneck* through Visual Structured Extraction (VSE), with direct application to university-level programming education. By decoupling visual perception from logical inference, VSE offloads the interpretation of instructional details onto a symbolic representation, allowing the student model to focus exclusively on reasoning. Our evaluation across diverse programming benchmarks reveals that the vast majority of model failures stem not from a lack of logical capacity, but from a fundamental failure in visual grounding. By implementing the DECODE protocol, we achieve a substantial parameter efficiency gain, where compact models match or exceed the reasoning performance of significantly larger architectures. This approach effectively reclaims the latent reasoning capacity of SVLMs without the computational overhead of fine-tuning. By providing a scalable and efficient path toward high-quality pedagogical support, DECODE represents a foundational step in democratizing AI tutoring across global STEM disciplines.

Keywords: Multimodal RAG · Small Vision-Language Models · Perception Bottleneck · Computer Science Education · Visual Semantic Extraction · Cross-modal Retrieval

1 Introduction

The rapid advancement of *Downscaling Intelligence* has catalyzed the development of SVLMs capable of running on consumer-grade hardware. In CSEd, these models offer a transformative opportunity for personalized, privacy-preserving

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 610–626, 2027.
https://doi.org/10.1007/978-3-032-30494-0_38

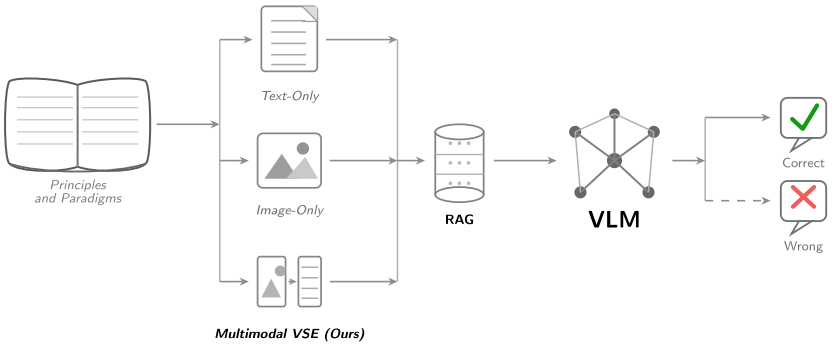


Fig. 1. Comparative architectural analysis of multimodal reasoning configurations. Starting from the textbook, the pipeline extracts multimodal data (Image-Only and Text-Only) alongside our Visual Structured Extraction (Multimodal VSE). Each modality is injected into the RAG knowledge base, which feeds the SVLM at inference time. Model outputs are assessed as correct or incorrect.

tutoring systems that can interpret both pedagogical text and technical illustrations [22,34]. However, model downscaling involves an unequal capacity regression with respect to the nature of the input: while the textual reasoning channel undergoes a moderate reduction, the visual channel suffers a drastic regression [7], ultimately inducing a *representational collapse* of the vision encoder. This phenomenon, which we term the Perception Bottleneck, is particularly consequential in CSEd, where diagrams such as memory stack and heap models, pointer chains, and finite automata are strictly relational and symbolic. According to *Dual Coding Theory* [4], mastering these abstractions requires the simultaneous processing of verbal and visual representations, such that a single perceptual error, for instance, misinterpreting a pointer arrow or a memory address label, can invalidate an entire logical inference chain, even when the model possesses the underlying domain knowledge to solve the problem. To address this, we present DECODE (*Decoupled Encoding of COntextual Diagram sEmantics*), a training-free implementation of the *Extract+Think* paradigm [7] designed to overcome the Perception Bottleneck in CSEd, see Fig. 1. Rather than forcing a constrained vision encoder to parse high-density technical symbols, DECODE employs a teacher model (GPT-4o) as an offline visual extraction oracle to transform raw diagrams into structured Instruction-Relevant Visual Details (IRVD), allowing the SVLM to focus its limited parameter count exclusively on the reasoning phase of the task. Our study utilizes a corpus derived from a foundational programming textbook [9], comprising 619 segments. We test our approach using PROG-VQA, a 1,000-item synthetic benchmark, and 50-VQA, a human-validated gold set designed to ensure ecological validity. Our results demonstrate that by isolating the perception loss, DECODE allows compact SVLMs to approach or exceed the performance of significantly larger archi-

tectures, confirming that the intelligence of small models is often masked by faulty perception. Our contributions are threefold:

1. We formalize the perception bottleneck in CSEd, demonstrating that CLIP-based retrieval [32] achieves an Image Retrieval Rate of 0.00 on formal diagrams, and that the vast majority of SVLM failures are perceptual rather than logical in nature.
2. We introduce DECODE, a scalable VSE pipeline that recovers cross-modal alignment without any model fine-tuning, offering a practical solution for resource-constrained educational environments.
3. We characterize a *Reasoning Threshold*, identifying the minimum model capacity required for successful perception–reasoning decoupling, and demonstrate that compact models can match or exceed the reasoning performance of significantly larger architectures.

2 Related Work

The Perception Bottleneck in Model Downscaling. A substantial body of work has investigated the failure modes of Vision-Language Models on tasks requiring visual grounding [3,23,33,40,43]. A recurring finding is the tendency of models to bypass visual content entirely, inferring answers from linguistic priors rather than genuinely interpreting the diagram [7,43]. Through a decoupled evaluation framework, Endo et al. [7] demonstrate that parameter reduction disproportionately degrades low-level perceptual capabilities relative to linguistic reasoning, ultimately inducing a *representational collapse* of the visual encoder. While this effect has been documented on scientific figures and mathematical diagrams [40,42], the behavior of SVLMs on formal programming diagrams remains an open empirical question that this work directly addresses. Crucially, Endo et al. [7] propose *visual extraction tuning* as a remediation strategy, which however requires additional supervised training on task-specific data. DECODE, by contrast, achieves the same decoupling in a training-free manner, substituting fine-tuning with an offline VSE protocol driven by a teacher model, making it directly deployable on resource-constrained hardware without any gradient update.

Symbolic Complexity and Dual Coding in CSEd. The perceptual bottleneck is particularly consequential in CSEd, where visual representations typically consist of technical diagrams, pointer chains, and memory layouts [9] whose semantics are relational and strictly domain-specific [4,6,35]. Dual Coding Theory [4] highlights the critical role of synchronizing verbal and visual representational channels, given that subject materials are deeply supported by instructional figures. Textual explanation alone is insufficient for students to construct accurate mental models of theoretical programming concepts [35]. This structural specificity renders general-purpose encoders like CLIP [32] ill-suited for the task, since a single perceptual error, such as misinterpreting a pointer arrow or a memory address label, invalidates the entire reasoning chain [23,40]. Zhang

et al. [40] further confirm this by showing that most multimodal models rely on textual priors rather than genuinely interpreting diagrams, with some models scoring higher *without* visual input at all. Unlike the mathematical diagrams studied in MathVerse [40], CS-specific diagrams such as pointer chains, memory maps, and abstract machines are non-redundant: no textual description accompanies them in the source material [9], making linguistic shortcutting impossible and visual grounding strictly necessary. Our empirical finding that CLIP-based retrieval achieves an Image Retrieval Rate of 0.00 on such diagrams provides direct quantitative support for this framing, and constitutes the core motivation for our VSE-based approach, which is specifically designed to operate on this class of non-redundant, strictly symbolic diagrams.

Multimodal Retrieval. The *Extract+Think* paradigm proposes decoupling perceptual grounding from logical inference as a principled response to the bottleneck described above [7,42]. Retrieval-Augmented Generation (RAG) has emerged as a dominant strategy for grounding model outputs in domain-specific knowledge without retraining [8,37–39], with a consistent finding across recent analyses [23,33] that textual descriptions of visual content outperform raw image embeddings in retrieval, suggesting that the challenge lies in how information is represented rather than in the generative capacity of the model [1,14]. Nowak et al. [23] confirm this across chart datasets but require a large vision encoder and task-specific training; Riedler and Langer [33] corroborate it in industrial settings, noting that image retrieval remains the primary failure point; Bharadwaj and Jain [1] further show that RAG benefits smaller models more than larger ones, a pattern we deliberately exploit through our reasoning threshold characterization. Unlike these approaches, which either operate on non-symbolic visual domains or require task-specific training and large vision encoders, DECODE addresses the perception bottleneck in the CSEd domain in a fully training-free manner, combining offline VSE extraction with a local, privacy-preserving inference stack that is directly deployable on resource-constrained hardware without any gradient update.

Ethics, Privacy, and the Case for Edge AI. The deployment of AI tutors raises critical concerns regarding student privacy and digital sovereignty [11–13,15,21]. Cloud-based inference routes sensitive educational data through third-party infrastructure [2,5], a concern that is particularly acute in low-resource settings where cloud reliance risks exacerbating the digital divide [10,17]. Edge AI and on-device inference represent a viable alternative [16,22,30,34]: by executing inference locally via OLLAMA [27], sessions remain private and fully functional without connectivity [34]. DECODE directly addresses this gap by enabling SVLMs to effectively process the complex formal diagrams essential to CSEd, providing high-quality pedagogical support that is simultaneously privacy-preserving and accessible on low-resource devices worldwide.

3 The PROG-VQA Dataset: Beyond Aesthetic Perception

The knowledge base in this project is built from the university course textbook *Programming Languages: Principles and Paradigms* [9]. The choice is strictly related to the goal of our work: building an educational AI tutor tailored to the students of the course, capable of answering with the specific terminology used and explained during lessons. To maintain the multimodal aspect, central in our work and in this domain, we extracted text and images from the book. The images are highly informative and technically focused.

3.1 The Challenge of CS-Specific Diagrams

The challenge lies in depicting these images as knowledge base entries. Here, the problem acquires the name of the perception bottleneck identified by [7]. The course concepts, apart from the visual component, could be covered by SVLMs due to their capacity to *Think*, which is not subject to the bottleneck. The problem remains limited to visual representation, which, if misinterpreted, entails a significant lack in the quality of answers, leaving them incomplete in terms of technical details. The *Extract* phase therefore suffers from *Representational Collapse* when faced with the high-density visual signals typical of CS formalisms.

3.2 Benchmark Composition

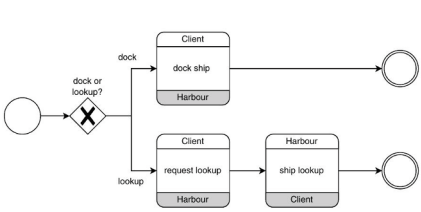
We curated the Prog-VQA benchmark, composed of two evaluation sets: the 50-VQA human-validated dataset, comprising 50 high-complexity diagrams extracted directly from the textbook [9] and manually annotated, and the Synthetic Augmentation Prog-VQA, composed of 1,000 visual question-answering pairs generated to ensure broad coverage of difficulty levels and diagram types.

4 Methodology: The DECODE Framework

The Decode framework is a training-free pipeline designed to mitigate the representational collapse of small vision encoders by decoupling multimodal perception from logical inference. Following the *Extract+Think* paradigm [7], the framework offloads the visual grounding of high-fidelity diagrams to a high-capacity teacher model, providing a clear symbolic representation to a compact student model and thereby bypassing the perception bottleneck that typically inhibits performance in Edge AI applications [22,34].

Data acquisition is structured into two distinct stages:

– *Stage 1: Structural Textual Segmentation.* Textual content is segmented using `PyPDFLoader` [18], preserving the logical structure of paragraphs alongside page-level metadata.

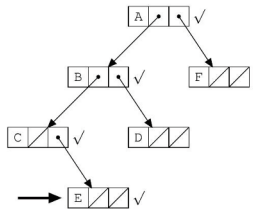


Q: How does the guarded choice work in this schema ?

✓ **DECODE:** In this abstract machine diagram, the guarded has to choose between `dock` or `lookupShip` where if the client selects `dock`, the interaction terminates; otherwise, `lookupShip` triggers a response.

× **Hallucination:** The guarded choice is a standard Erlang process management tool. It uses guard clauses to manage incoming requests in the VM.

Fails to consider the current image and visual branching structure in it.



Q: What is the value obtained by `*p`?

✓ **DECODE:** Looking at the memory map, `p` at `0x02` points to `0x01`. Since `x` is `10` at that address, dereferencing `p` results in **10**.

× **Perceptual Failure:** The pointer `p` seems to reference an unspecified location.

Fails to follow the pointer edge due to a lack in visual capability.

Fig. 2. Case studies on Programming Language Pedagogy visual queries. The **Baseline** (SVLM + CLIP) exhibits *Representational Collapse*, failing to resolve the relational structures of the abstract machine diagram (left) and the pointer chain (right). Our **DECODE** (SVLM + VSE) overcomes the *Perception Bottleneck* via the **Extract+Think** paradigm: on the left, it correctly identifies the branching logic; on the right, it successfully dereferences the pointer by reading the memory map.

– *Stage 2: Visual Element Isolation.* Visual extraction is performed using the `Unstructured` library [36], which isolates all figures present in the textbook and attaches metadata encoding their placement and positioning within the pedagogical flow.

The pipeline produced 619 segments, comprising 541 narrative text chunks and 78 visual elements. Although visual elements constitute only 12.6% of the corpus, they represent the semantic core of the domain, containing the high-density relational data, such as pointer addresses and memory layout structures, where SVLMs typically fail.

4.1 RAG System Architecture

We implemented a modular RAG architecture using a local Qdrant instance [31]. To isolate the impact of visual information on model performance, we constructed three distinct collections.

Decode—Multimodal VSE Collection. This collection implements our training-free extraction protocol. Unlike standard image captioning, which often produces generic descriptions, VSE is engineered to capture the specific formalisms of technical diagrams, including memory addresses, pointer relationships, and execution flows in code structures [23,33]. Visual descriptions were generated offline by GPT-4o acting as a zero-shot visual oracle, transforming raw pixels into structured VSE descriptors that serve as IRVD. As illustrated

in Fig. 2, symbolic relations are translated into text, for example representing a pointer p at address `0x02` pointing to x at `0x01`, allowing the SVLM to bypass its internal vision encoder entirely. Formally, the composite representation c_i for chunk i is:

$$c_i = \text{Enc}_{\text{MiniLM}}(d_i \,\|\, v_i), \quad c_i \in \mathbb{R}^{384} \tag{1}$$

where d_i is the text chunk, v_i the VSE descriptor, and $\|$ denotes concatenation. By providing the student model with a clear symbolic script of the diagram, we transform a complex multimodal reasoning task into a focused logical inference problem, allowing it to leverage its inherent linguistic and logical capabilities, which are disproportionately stronger than its visual perception [7,43].

Baseline Configurations. DECODE is evaluated against two alternative data representation strategies:

- *CLIP RAG (Native):* Visual content is embedded directly into the RAG system using *CLIP ViT-B/32* [32], with context retrieved based on visual similarity. This configuration maintains a purely multimodal architecture and serves as the primary reference point for quantifying the perceptual gap, as it represents the most common baseline in multimodal RAG literature [23,33].
- *Text-Only RAG:* Only the narrative textual segments from Stage 1 are encoded, excluding all visual metadata. This configuration isolates the model's internal linguistic priors, allowing us to distinguish failures due to missing visual grounding from those attributable to insufficient domain knowledge [1,8].

Because DECODE is entirely training-free, it circumvents the substantial computational costs and data requirements of fine-tuning or domain-specific pretraining, making it a scalable and accessible solution for deploying high-quality multimodal educational support on resource-constrained devices where privacy and low latency are critical [16,22,34].

5 Experimental Setup: Isolating the Perception Loss

The objective of our experimental design is to empirically isolate the perception bottleneck and quantify the extent to which representational collapse in vision encoders inhibits the reasoning potential of small-scale models. By adopting a training-free approach, we aim to demonstrate that compact Edge AI architectures can achieve performance parity with much larger models when visual grounding is properly structured through VSE [7,22,34].

5.1 Model Selection and Downscaling Strategy

To assess the scalability and architectural resilience of DECODE, we selected four representative SVLMs stratified by parameter count, examining how the

Table 1. Overall performance and efficiency metrics for human-graded and synthetic technical queries. The modalities comparison over models evaluated over our two benchmarks: Human Gold is the 50-VQA human validated benchmark and the 1K synthetic PROG-VQA benchmark. The names of the abbreviated metrics in the table are as follows: BS = BERTScore F1; R-L = ROUGE-L; Lat = avg. latency (s). † perception collapse of Moondream on visual items.

Strategy	Model	Human Gold (50)		PROG-VQA (1k)		Efficiency
		BS	R-L	BS	R-L	Lat
DECODE (Ours)	LLaVA-7B [24]	**.82**	**.25**	**.83**	**.31**	1.83
	LLaVA-Phi3 [25]	.79	.21	.79	.23	1.78
	Moondream† [26]	.55	.14	.66	.18	0.85
	Qwen2.5-7B [28]	.80	.25	.79	.23	4.40
CLIP RAG	LLaVA-7B [24]	.81	.24	.80	.24	1.64
	LLaVA-Phi3 [25]	.80	.22	.79	.21	1.72
	Moondream [26]	.79	.20	.73	.18	0.83
	Qwen2.5-7B [28]	.78	.21	.76	.17	3.15
Text-Only RAG	LLaVA-7B [24]	.82	.26	.82	.29	1.66
	LLaVA-Phi3 [25]	.81	.23	.80	.23	1.85
	Moondream [26]	.80	.21	.76	.21	0.82
	Qwen2.5-7B [28]	.80	.24	.78	.22	2.81

† Moondream: empty/nonsensical outputs on 34% (Human Gold) and 26% (PROG-VQA) of visual items under DECODE, deflating the average BS.

perception bottleneck manifests across the size spectrum. All models are served locally via OLLAMA [27], ensuring data sovereignty and reproducible latency, both of which are critical constraints for educational edge deployment [22,34].

- **Lower-Bound SVLMs:** MOONDREAM2 [26] (1.6B) serves as the extreme-efficiency lower bound, optimized for high-efficiency Edge AI deployment, while LLAVA-PHI3 [25] (3.8B) represents the *downscaling sweet spot*, combining mid-range multimodal reasoning with a compact footprint.
- **Upper-Bound References:** LLAVA V1.5 [24] (7B) and QWEN 2.5 [28] (7B) serve as high-capacity baselines to assess whether larger vision encoders (CLIP-ViT-L/14) or higher textual knowledge density can autonomously resolve the *representational collapse* without external grounding support.

The structured visual metadata provided by DECODE is hypothesized to enable the 1.6B model to approach the performance of the 7B baselines, validating perception-reasoning decoupling as a parameter-efficient strategy for Edge AI deployment.

5.2 Implementation Details

The entire inference stack runs locally to ensure data sovereignty and reproducibility [22,34]. All four VLMs are served via OLLAMA [27] on a single

Table 2. Ablative analysis performance over BERTScore F1 metric. The abbreviations in the table stand for: H50 for Human Gold 50-VQA and PV for PROG-VQA.

		By Difficulty						**By Type**			
		Easy		Medium		Hard		Visual (V)		Textual (T)	
Strategy	**Model**	H50	PV	H50	PV	H50	PV	H50	PV	H50	PV
DECODE	L7B	.83	.82	**.82**	**.83**	**.81**	**.82**	**.82**	**.79**	**.82**	**.84**
(Ours)	LP3	.83	.79	.78	.80	.79	.79	.79	.76	.81	.81
	M1.6‡	.80	.66	.48	.67	.57	.65	.52	.38	.79	.78
CLIP RAG	L7B	.82	.79	.82	.81	.80	.80	.81	.78	.80	.81
	LP3	.78	.77	.80	.80	.80	.79	.80	.78	.79	.79
	M1.6	.79	.67	.79	.74	.78	.76	.79	.63	.77	.77

L7B = LLaVA-7B [24]; LP3 = LLaVA-Phi3 [25]; M1.6 = Moondream2 [26].
‡ Perception collapse: Moondream performance drops under DECODE due to SVLM image-processing failures.

consumer-grade machine, without any cloud dependency.[1] A unified generation configuration was applied across all runs, with a near-deterministic decoding strategy and a fixed random seed to guarantee reproducibility; no model-specific prompt tuning was applied, so all differences in output are attributable solely to the retrieval strategy.[2] PDF ingestion follows the two-stage protocol described in Sect. 4: narrative text is segmented via *PyPDFLoader* [18] (LangChain), while figures are isolated with the *Unstructured* library [36], preserving page-level metadata and positional context. For the DECODE collection, VSE descriptors were generated *offline* by GPT-4o [29] as a zero-shot visual oracle, incurring no latency at inference time. The vector store is managed via a local QDRANT [31] instance. Responses are finally assessed using *BERTScore F1* [41][3] for semantic alignment and *ROUGE-L* for surface-level generation quality, as detailed in Sect. 5.5.

5.3 Comparative RAG Configurations

The core of our diagnostic strategy is a perception-reasoning ablation designed to detect representational collapse [7,43]. The three configurations isolate where the failure occurs:

1. **Native Multimodal RAG:** The SVLM receives raw pixels alongside the query, with the model's internal encoder responsible for parsing complex CS diagrams. We expect this to be the primary site of the bottleneck, as the integrated *Extract+Think* capacity is subject to representational collapse [23, 40].

[1] All experiments were conducted on a single NVIDIA GPU with 16 GB VRAM.
[2] Specifically: `temperature` = 0.1, `top_p` = 0.9, `num_predict` = 512 tokens, `seed` = 42.
[3] `distilbert-base-uncased`, via the `bert-score` Python library.

Table 3. In-depth Analysis: Error Distribution and Cross-domain Generalization. Part I: Error shift on the 50-item Gold Set between Native CLIP and Decode; Part II: Zero-shot performance across public benchmarks on the top line and between SVLMs architectures on the bottom line, comparing Image Only and DECODE.

Part I: Error Distribution (Human Gold Set)					
	Native CLIP		**Decode**		**Delta**
Error Category	Count	%	Count	%	
Perceptual Failure	38	76%	2	4%	−72%
Linguistic Hallucination	6	12%	3	6%	−6%
Reasoning Flaw	4	8%	2	4%	−4%
Total Failures	**48/50**	**96%**	**7/50**	**14%**	**+82% Acc.**
Part II: Zero-Shot Generalization and Scalability					
	Image Only		**Decode**		
Benchmark / Model	BS	Acc	**BS**	**Acc**	**ΔBS**
ChartQA [20]	.58	.03	**.64**	**.06**	+0.06
MMStar [3]	.63	.03	**.66**	**.11**	+0.03
ScienceQA [19]	.68	.04	**.69**	**.03**	+0.01
LLaVA-7B [24]	.60	.04	**.65**	**.10**	+0.05
Qwen2.5-7B [28]	.58	.01	**.62**	**.06**	+0.04
LLaVA-Phi3 [25]	.64	.11	.63	.08	(−0.01)

BS = BERTScore F1; Acc = Accuracy; ΔBS = Gain over Baseline.

2. **Text-Only RAG:** The model receives only narrative text segments, isolating its *pure reasoning* phase without any visual interference, thereby establishing the linguistic prior as a lower bound [1,8].
3. **Decode:** Following the protocol in [7], raw pixels are replaced with IRVD generated by the teacher model. By providing pre-extracted perception, we measure whether the SVLM can reclaim the reasoning performance of larger architectures, quantifying the *latent reasoning delta* achieved solely by resolving the perception bottleneck.

5.4 Cross-Domain Generalization Benchmarks

To evaluate the robustness of Decode beyond CSEd and address concerns about generalizability, we extend our analysis to three established multimodal benchmarks that require structured visual reasoning: ScienceQA [19], covering multimodal scientific reasoning where diagrams in the Natural Science domain require precise attribute extraction; ChartQA [20], involving complex logical reasoning over visual charts where SVLMs typically struggle with coordinate and value perception; and MMStar [3], specifically designed to probe whether vision-language models are genuinely interpreting visual inputs or merely exploiting

linguistic priors, directly addressing our central hypothesis on the perception bottleneck.

5.5 Evaluation Metrics and Methodology

ROUGE-L is computed to assess surface-level generation quality, while *BERTScore F1* [41] measures semantic alignment, penalizing hallucinations without penalizing correct paraphrases in open-ended VQA settings. As illustrated by the decoupling analysis in Fig. 4, the evaluation focuses on *Perceptual Fidelity*: the ability to correctly identify relational and symbolic elements such as pointer dereferencing. To isolate this, we specifically categorize instances of *Perceptual Failure* (see Table 3, Part I), defined as cases where the model's logical reasoning is sound but its visual input is corrupted by the perception bottleneck.

5.6 Ablation: Task Complexity and Visual Gap

We conduct a structured ablation along two dimensions. **Task Complexity:** Questions are categorized by the logical density of their associated visual data into *Easy*, *Medium*, and *Hard* levels. The bottleneck is hypothesized to be most pronounced on *Hard* queries, such as those involving multi-level pointer chains, where a single perceptual error cascades through the entire reasoning chain [23,40]. **Visual-Requirement Gap (V vs. T):** Questions that strictly require diagram analysis are labeled as visual queries (V), while the remainder are treated as textual (T). This division tests the decoupling hypothesis directly: SVLMs should remain competitive on textual queries regardless of the retrieval strategy, while degrading significantly on visual queries unless supported by DECODE.

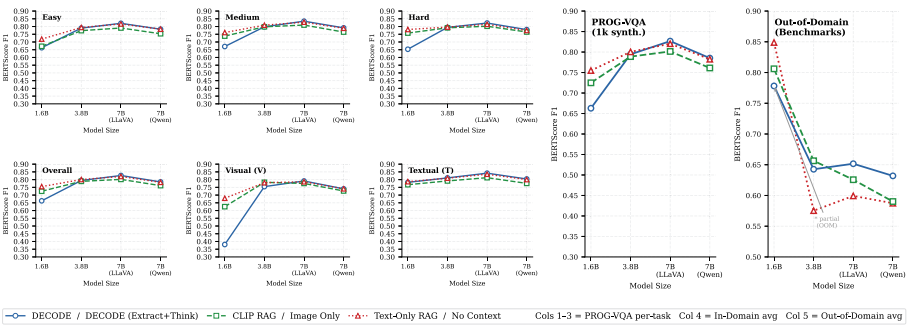


Fig. 3. Quantitative performance across task complexities and domains. Results are shown for three retrieval modalities across model sizes, under Task Complexity ablation (Easy/Medium/Hard) and Visual-Textual type split. The rightmost columns report Out-of-Domain generalization on CHARTQA [20], SCIENCEQA [19], and MMSTAR [3].

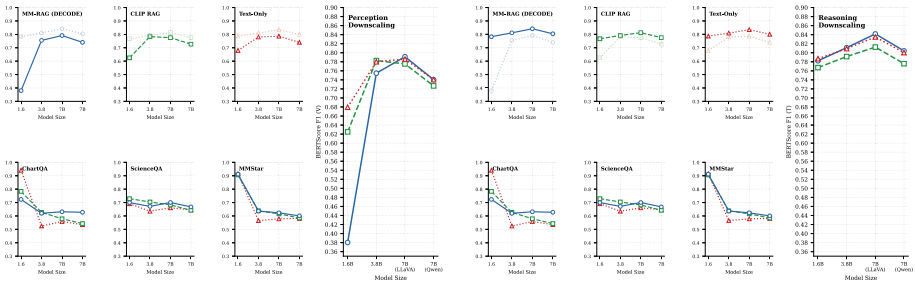


Fig. 4. The **Perception-Reasoning Asymmetry** in Model Downscaling. The left panel (*Perception*) evaluates performance on visually heavy queries only, while the right panel (*Reasoning*) isolates text-based queries, showing that logical capacity remains largely intact regardless of model size. Bottom rows report per-modality performance across out-of-domain benchmarks, confirming that the asymmetry between vision and logic is a domain-general phenomenon in technical visual reasoning.

6 Results and Discussion

As shown in Table 1, DECODE with LLAVA V1.5-7B achieves the highest overall performance, reaching a BERTScore of 0.82 on 50-VQA and 0.83 on PROG-VQA, outperforming all competing configurations. Notably, CLIP RAG consistently underperforms even the Text-Only RAG baseline when the underlying model is held constant, as observed for instance with QWEN 2.5-7B, confirming that raw image embeddings actively harm retrieval quality on formal CS diagrams. MOONDREAM (1.6B), the smallest model evaluated, degrades under DECODE due to information overload: the volume of structured visual details exceeds its reasoning capacity, placing it below the reasoning threshold identified in Sect. 1.

Table 2 provides a fine-grained breakdown by task complexity and query type across both benchmarks. DECODE with LLAVA V1.5-7B demonstrates strong robustness to increasing difficulty, maintaining a *BERTScore* in the range of 0.81–0.83 across Easy, Medium, and Hard queries, whereas CLIP RAG models exhibit a consistent performance drop as task complexity increases. The failure of MOONDREAM under DECODE on Visual-type queries is consistent with the information-overload hypothesis: the model cannot leverage the additional structured context, while CLIP RAG, which provides no symbolic detail, paradoxically avoids this collapse. LLAVA V1.5-7B with DECODE yields the strongest gains specifically on Visual (V) queries, where perceptual grounding is strictly necessary.

Table 3, Part I, reveals the magnitude of the perceptual failure under the Native CLIP baseline: 96% of responses are incorrect, with 76% of errors attributable to perceptual failure rather than reasoning deficits. Under DECODE, perceptual failures drop from 76% to 4%, yielding an overall accuracy gain of 82% points and confirming that the bottleneck is primarily perceptual, not cognitive. Table 3, Part II, reports zero-shot generalization performance on three

public benchmarks. Despite not being designed for these out-of-domain settings, DECODE yields positive Δ *BERTScore* on CHARTQA (+0.06), MMSTAR (+0.03), and SCIENCEQA (+0.01). These results are reflected in accuracy improvements, with scores increasing from 0.03 to 0.06 on CHARTQA and from 0.03 to 0.11 on MMSTAR. These results confirm that the VSE-based approach generalizes beyond CS-specific diagrams to technical visual reasoning more broadly, with the most consistent gains observed on mid-sized model architectures.

Figure 3 corroborates these findings quantitatively. The DECODE curve remains flat or improves as task difficulty increases from Easy to Hard, in contrast to both baselines, which degrade monotonically. Furthermore, the performance advantage of DECODE emerges consistently above a model size of approximately 3.8B parameters, which we identify as the reasoning threshold, below which, as in the case of MOONDREAM (1.6B), the additional structured context cannot be effectively exploited.

Figure 4 isolates the *downscaling asymmetry* directly: visual performance degrades sharply as model size decreases, while text-based reasoning remains largely stable across all model sizes. This asymmetry holds consistently across out-of-domain benchmarks, confirming that the gap between perceptual and reasoning capacity is a domain-general property of small vision-language models on technical visual content, rather than an artefact of the CSEd setting.

Qualitative Evidence. Figure 2 anchors these quantitative patterns in concrete failure modes. The Native CLIP baseline fails to follow pointer edges or resolve branching logic, whereas DECODE restores the Dual Coding synchronization necessary for correct mental model formation, effectively transforming a perceptually impaired SVLM into a grounded pedagogical tutor.

7 Conclusion

This paper has addressed the fundamental challenge of deploying SVLMs for CSEd, identifying a critical representational collapse within vision encoders as the primary barrier to effective multimodal pedagogy [7,43]. We have demonstrated that the failure of SVLMs to interpret technical diagrams is not a deficit of latent intelligence, but a failure of visual grounding: visual comprehension degrades catastrophically as models are scaled down, while reasoning capacity remains comparatively intact, confirming a pronounced *downscaling asymmetry*. Our analysis of the textbook-grounded corpus [9] revealed that while visual elements, such as memory models and pointer chains, comprise only 12.6% of the 619 extracted segments, they represent a critical semantic threshold where SVLMs consistently fail. The results from the PROG-VQA benchmark and the human-validated 50-VQA gold set demonstrate that offloading the interpretation of IRVD to a teacher model allows SVLMs to reclaim their latent reasoning capabilities, confirming that *Downscaling Intelligence* can be effectively managed in educational settings by bypassing faulty vision encoders [23,33]. Ultimately,

DECODE restores the pedagogical benefits of Dual Coding [4]: by providing high-fidelity textual descriptors of symbolic relations, it ensures that the verbal and visual channels remain synchronized, even within the constraints of a compact model. This is a crucial step toward the democratization of AI tutoring, ensuring that high-quality, privacy-preserving programming education remains accessible in resource-constrained environments worldwide [16,22,34].

8 Future Work and Limitations

Future work will extend DECODE along three directions: (i) autonomous, training-free extraction agents for multi-layered diagrams without predefined templates; (ii) validation across broader symbolic-heavy STEM disciplines to test the universality of the perception bottleneck [4,35]; and (iii) lightweight on-device vision-language adapters to eliminate the dependency on an external teacher model [16,34]. A further direction concerns evaluating the long-term pedagogical impact of structured visual grounding on students' understanding of abstract programming concepts [6,22].

The primary limitation of this study concerns its scope: the evaluation is grounded in a single programming textbook [9], restricting the direct generalizability of our findings to other curricula and instructional styles. A secondary limitation is the one-time dependency of the VSE phase on GPT-4o [29], which, while executed offline, is not yet fully consistent with a self-contained Edge AI deployment [22,34].

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bharadwaj, A., Jain, K.: Rag-assisted small language models for domain-level reasoning. In: 2025 Eighth International Conference on Image Information Processing (ICIIP), pp. 231–236 (2025). https://doi.org/10.1109/ICIIP68302.2025.11346385
2. Chaparadza, S., Bafana, S.: Promoting student sovereignty and data privacy in an ai-free higher education landscape: challenges and benefits. GAS J. Educ. and Lit. (GASJEL) **2**(9), 1–14 (2025)
3. Chen, L., et al.: Are we on the right way for evaluating large vision-language models? (2024). https://arxiv.org/abs/2403.20330
4. Clark, J., Paivio, A.: Dual coding theory and education. Educ. Psychol. Rev. **3**, 149–210 (09 1991). https://doi.org/10.1007/BF01320076
5. Deckker, D., Sumanasekara, S.: Student-centric ethical frameworks for ai-driven education: participatory consent and data ownership in a global perspective. Int. J. Sci. Res. Arch. **15**(3), 1491–1503 (2025)
6. Denny, P., MacNeil, S., Savelka, J., Porter, L., Luxton-Reilly, A.: Desirable characteristics for ai teaching assistants in programming education. In: Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1., pp. 408–414. ITiCSE 2024, ACM (Jul 2024). https://doi.org/10.1145/3649217.3653574

7. Endo, M., Yeung-Levy, S.: Downscaling intelligence: exploring perception and reasoning bottlenecks in small multimodal models (2025). https://arxiv.org/abs/2511.17487
8. Fan, W., et al.: A survey on rag meeting LLMs: Towards retrieval-augmented large language models (2024). https://arxiv.org/abs/2405.06211
9. Gabbrielli, M., Martini, S.: Programming Languages: Principles and Paradigms. Undergraduate Topics in Computer Science, Springer (2010)
10. Gonzabay-Jiménez, E.F., Vargas-M, A., Velasteguí, H.J., Castro-Chuquiana, R.: Academic performance and inequality: Can generative ai widen the digital divide in higher education? In: International Conference on Emerging Technologies and Educational Innovation, pp. 193–204. Springer (2025)
11. Hamadeh, S., Amin, H.: AI, education and digital sovereignty. In: Frontiers in Education, vol. 10, p. 1677727. Frontiers Media SA (2025)
12. Holmes, W., Porayska-Pomsta, K.: The ethics of artificial intelligence in education. Lontoo: Routledge, pp. 621–653 (2023)
13. Huang, L.: Ethics of artificial intelligence in education: student privacy and data protection. Sci. Insights Educ. Front. **16**(2), 2577–2587 (2023)
14. Kazamel, S., Kocasarı, U., Alıcı, A.: An end-to-end computer vision system for structured information extraction from turkish id card images. In: Computational Science and Its Applications – ICCSA 2025, Part II, pp. 51–64. Springer-Verlag (2025). https://doi.org/10.1007/978-3-031-96997-3_4
15. Khan, W.N.: Ethical challenges of ai in education: balancing innovation with data privacy. AI EDIFY J. **1**(1), 1–13 (2024)
16. Koukaras, C., Stavrinides, S.G., Hatzikraniotis, E., Mitsiaki, M., Koukaras, P., Tjortjis, C.: Navigating the future of education: a review on telecommunications and ai technologies, ethical implications, and equity challenges. Telecom **7**(1), 2 (2026)
17. Landa-Blanco, M.: Artificial intelligence in education: applications and limitations for teachers in low-and middle-income countries. In: Frontiers in Education. vol. 10, p. 1681836. Frontiers Media SA (2025)
18. LangChain: Pypdfloader – langchain document loaders. https://docs.langchain.com/oss/python/integrations/document_loaders/pypdfloader (2024). Accessed 16 Mar 2026
19. Lu, P., et al.: Learn to explain: Multimodal reasoning via thought chains for science question answering (2022). https://arxiv.org/abs/2209.09513
20. Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning (2022). https://arxiv.org/abs/2203.10244
21. Moudgalya, S.K., Swaminathan, S.: Toward data sovereignty: Justice-oriented and community-based ai education. In: Proceedings of the 2024 on RESPECT Annual Conference, pp. 94–99 (2024)
22. Muminova, U., et al.: Real-time content personalization in educational IoT networks using on-device learning. J. Wirel. Mob. Netw., Ubiquit. Comput., Depend. Appl. **16**(2), 793–808 (2025)
23. Nowak, A., Piccinno, F., Altun, Y.: Multimodal chart retrieval: a comparison of text, table and image based approaches. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 5488–5505 (2024)
24. Ollama Team: Llava on ollama library. https://ollama.com/library/llava (2024), 7B multimodal reference model

25. Ollama Team: Llava-phi3 on ollama library. https://ollama.com/library/llava-phi3 (2024), 3.8B parameter model based on Phi-3 Mini
26. Ollama Team: Moondream2 on ollama library. https://ollama.com/library/moondream (2024), ultra-light 1.6B parameter model
27. Ollama Team: Ollama: Run large language models locally. https://ollama.com (2024), framework for running LLMs locally
28. Ollama Team: Qwen2.5 on ollama library. https://ollama.com/library/qwen2.5 (2024), 7B reasoning-oriented model
29. OpenAI: Hello GPT-4o. https://openai.com/index/hello-gpt-4o/ (2024). Accessed 16 Mar 2026
30. Pinho, P.C., da Silva, R.Z., Primo, T.T., Dermeval, D., Bittencourt, I.I., Isotani, S.: Win-win situation: Generative ai in educational constrained environments. In: Simpósio Brasileiro de Informática na Educação (SBIE), pp. 1607–1616. SBC (2025)
31. Qdrant Team: Qdrant: High-performance vector similarity search engine. https://qdrant.tech (2024). Accessed 16 Mar 2026
32. Radford, A., et al.: Learning transferable visual models from natural language supervision (2021). https://arxiv.org/abs/2103.00020
33. Riedler, M., Langer, S.: Beyond text: Optimizing rag with multimodal inputs for industrial applications (2024). https://arxiv.org/abs/2410.21943
34. Shidaganti, G., Raj, V.A., Raman, V.M., Kumaran, S.: Edge computing in educational technology: The power of edge ai for dynamic and personalized learning. Edge of Intelligence: Exploring the Frontiers of AI at the Edge, pp. 121–152 (2025)
35. Sweller, J.: The worked example effect and human cognition. Learn. Instruct. **16**(2), 165–169 (2006). https://doi.org/10.1016/j.learninstruc.2006.02.005, https://www.sciencedirect.com/science/article/pii/S0959475206000193, recent Worked Examples Research: Managing Cognitive Load to Foster Learning and Transfer
36. Unstructured Technologies: Unstructured: Document processing for LLM applications. https://unstructured.io (2024). Accessed 16 Mar 2026
37. Xiong, P., Lacewell, C.W., Gobriel, S., Jain, N.: Context is all you need: Efficient retrieval augmented generation for domain specific AI. In: First Workshop on Scalable Optimization for Efficient and Adaptive Foundation Models (2025). https://openreview.net/forum?id=6bKHoUQWlo
38. Xu, R., et al.: Simrag: self-improving retrieval-augmented generation for adapting large language models to specialized domains. In: Chiruzzo, L., Ritter, A., Wang, L. (eds.) Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 11534–11550. Association for Computational Linguistics, Albuquerque, New Mexico (2025)
39. Yang, M., Ren, Y., Opoku, D.O., Li, R., Ren, P., Xing, C.: DSRAG: A domain-specific retrieval framework based on document-derived multimodal knowledge graph. arXiv preprint arXiv:2509.10467 (2025)
40. Zhang, R., et al.: Mathverse: does your multi-modal LLM truly see the diagrams in visual math problems? In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision - ECCV 2024, pp. 169–186. Springer, Cham (2025)
41. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTscore: Evaluating text generation with BERT (2020). https://arxiv.org/abs/1904.09675

42. Zhang, Z., et al.: Diagram formalization enhanced multi-modal geometry problem solver. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE (2025)
43. Zhou, Y., et al.: Scientists' first exam: Probing cognitive abilities of MLLM via perception, understanding, and reasoning. arXiv preprint arXiv:2506.10521 (2025)