# From Visual Evidence to Reasoning: A Survey of Reasoning in Vision-Language Models

**Mobina Poulaei**

**Amir Qeysarbeigi**

**Masih Shalchian**

**Omid Ghahroodi**

**Hanieh Sartipi**

**Shaygan Adim**

## Abstract

Vision–language models (VLMs) have rapidly evolved from image captioners into multimodal assistants and, more recently, into models expected to perform explicit visual reasoning. Yet their ability to reason over images remains fragile: models that answer ordinary visual questions fluently can still fail on spatial relations, diagrammatic relations, abstract visual rules, and interactive interface tasks. Rather than treating these directions as isolated sub-fields, this survey develops a unified evidence-to-answer framework that maps visual reasoning from perception and grounding through intermediate representation and reasoning to verification. By reading these sub-fields against one another, we identify a recurring failure pattern: models may lack, misground, or fail to verify the visual evidence needed by subsequent reasoning. This perspective connects seemingly different remedies, including structured intermediate representations, explicit evidential reasoning, and perception-aware rewards, as interventions at different stages of the same pipeline. Across the surveyed works, several benchmark studies indicate that scaling improves recognition more reliably than relational or compositional reasoning, while methods that explicitly supervise visual grounding or intermediate evidence can produce stronger gains on targeted reasoning tasks. Recent reinforcement-learning approaches further illustrate the importance of reward design: answer-level rewards can improve final responses without necessarily improving visual perception, whereas perception-aware, self-rewarded, and process-level objectives provide more direct supervision for intermediate visual behavior. We synthesize benchmark comparisons, contrast design trade-offs, and identify open problems including 3D and egocentric understanding, faithful reasoning traces, shortcut-resistant evaluation, reward hacking, and high-confidence hallucination. Overall, the survey suggests that improving visual reasoning requires not only stronger reasoning capabilities, but also reliable mechanisms for acquiring, grounding, and verifying the visual evidence on which reasoning depends.

## 1 Introduction

Vision–language models (VLMs) couple a visual encoder with a large language model (LLM) to answer questions, follow instructions, describe images, and act in visual environments. The release of instruction-tuned multimodal assistants such as LLaVA (Liu et al., 2023) demonstrated that a lightweight projection between a frozen CLIP-style vision encoder and an open LLM, trained on machine-generated instruction data, could produce fluent multimodal dialogue. That success changed the field: the central question is no longer only whether a model can talk about images, but whether it can genuinely *reason* over what it sees.

The distinction is crucial. Many captioning and visual question answering tasks can be solved through object recognition, memorized facts, or language priors. By contrast, visual reasoning requires the model to bind objects to attributes, track relations, infer hidden rules, compare metric or topological quantities, plan a sequence of actions, and sometimes verify its own intermediate claims. A growing benchmark literature shows that frontier VLMs remain brittle on these capabilities: they struggle with left/right, distance, and perspective; they collapse when compositional shortcuts are removed; and they often perform poorly on diagram, chart, or Raven-style tasks even when they appear strong on general visual benchmarks.

Excellent surveys of multimodal reasoning already exist: Li et al. (2025c) map the evolution from modular pipelines to language-centric reasoning models, and Wang et al. (2025c) systematize multimodal chain-of-thought techniques. These surveys are organized primarily by subfield or technique. Our aim is complementary rather than competing: we ask how the seemingly separate directions–spatial grounding, compositional binding, structured intermediate representations, chain-of-thought in its explicit, latent, and imagined forms, reinforcement learning, and evaluation–fit together as interventions at different points of a single *visual reasoning pipeline*:

$$\text{image} \xrightarrow{\text{perception}} \xrightarrow{\text{grounding / binding}} \xrightarrow{\text{intermediate representation}} \xrightarrow{\text{reasoning}} \xrightarrow{\text{verification / action}} \text{answer}, \quad (1)$$

with training (SFT and RL) shaping every stage and evaluation probing them. This framing is a survey-level abstraction, not a claim made by any single paper, but it lets us pose five questions that recur throughout the reviewed literature: (i) *where is the bottleneck*–in the reasoning engine or in the evidence supplied to it; (ii) *what should the intermediate representation be*–raw visual tokens, regions and depth, symbolic structure, captions, tools, or latent states; (iii) *where does language reasoning help, and where does it fail*–in reasoning over grounded visual evidence, but not necessarily in acquiring or grounding that evidence; (iv) *how should reasoning behavior be trained*–and what do rewards actually teach; and (v) *how can we verify* that a model's reasoning is grounded in the image rather than in language priors. Figure 2 previews this framework and maps the surveyed works onto it, and Figure 3 traces how the field arrived at it.

**Contributions.** In summary, this survey makes the following contributions:

- **A unifying pipeline framework.** We organize spatial reasoning, compositional binding, structured intermediate representations, chain-of-thought (explicit, latent, and imagined), reinforcement learning, and evaluation as interventions at different stages of a single visual reasoning pipeline (Eq. 1). This provides a functional, evidence-centric blueprint that complements prior surveys organized primarily by historical roadmap, specialized techniques, or capability layers.
- **A cross-category synthesis of shared bottlenecks and remedies.** We identify unreliable or ungrounded intermediate visual evidence as a recurring failure pattern across spatial, compositional, chart, diagrammatic, and OCR settings, and that the corresponding remedies (structured intermediate representations, evidential reasoning steps, and perception-aware rewards) are variations of a single design principle expressed in different task languages.
- **Quantitative and qualitative benchmark comparison.** We synthesize human–model performance gaps and design trade-offs across representative benchmarks (Section 5, Figure 9, Table 3), offering a consolidated view of where current VLMs fail most severely.
- **An agenda of open problems.** We identify concrete open directions, including 3D/egocentric understanding, faithful (rather than merely formatted) reasoning traces, shortcut-resistant evaluation, reward hacking, and high-certainty hallucination, that follow directly from the cross-category synthesis.

**How this survey differs.** Rather than organizing the literature primarily by historical architectures Li et al. (2025c), prompting techniques Wang et al. (2025c), image-mediated tool paradigms Su et al. (2025), or perception-versus-cognition capability layers Zhou et al. (2025), we analyze visual reasoning through a unified, *evidence-centric* pipeline framework. While existing reviews evaluate isolated techniques or specific interaction paradigms, our cross-category approach examines precisely where interventions occur across the visual reasoning lifecycle—from perception and grounding to intermediate representation, reasoning,

| Survey | Core Scope & Focus | Primary Organizing Principle | Distinguishing Feature |
|---|---|---|---|
| Li et al. (2025c) | Evolution of VLM architectures and large multimodal reasoners. | Historical and architectural (modular to language-centric). | Comprehensive timeline of architectural shifts. |
| Wang et al. (2025c) | Multimodal Chain-of-Thought (M-CoT) prompting and fine-tuning. | Technique-centric (prompting vs. training strategies). | Deep dive into explicit reasoning trace generation. |
| Su et al. (2025) | Tool-augmented and image-mediated visual reasoning. | Paradigm-centric ("thinking with images"). | Focus on visual outputs as intermediate reasoning steps. |
| Zhou et al. (2025) | Interactive reasoning and visual dialogue. | Capability layers (perception vs. cognition). | Separation of basic recognition from complex interaction. |
| **Ours** | Bottlenecks in intermediate visual evidence across domains. | Evidence-centric (the visual reasoning pipeline). | Integrates spatial, compositional, RL, and CoT domains to reveal that preserving intermediate visual evidence is a shared remedy. |

Table 1: Comparison of recent surveys on multimodal reasoning and vision–language models. While prior surveys typically organize the literature by specific techniques or model capabilities, our survey introduces an *evidence-centric* pipeline framework to identify shared failure modes and remedies across diverse reasoning domains.

and verification. By synthesizing insights across spatial, compositional, RL, and CoT domains, this survey demonstrates that preserving and stabilizing intermediate visual evidence serves as a universal, shared remedy for diverse reasoning failure modes. Section 4 develops these pipeline connections within individual reasoning categories, while Section 5 synthesizes the resulting cross-category insights to identify shared evidence bottlenecks and remedies. Figure 1 summarizes this overall structure.

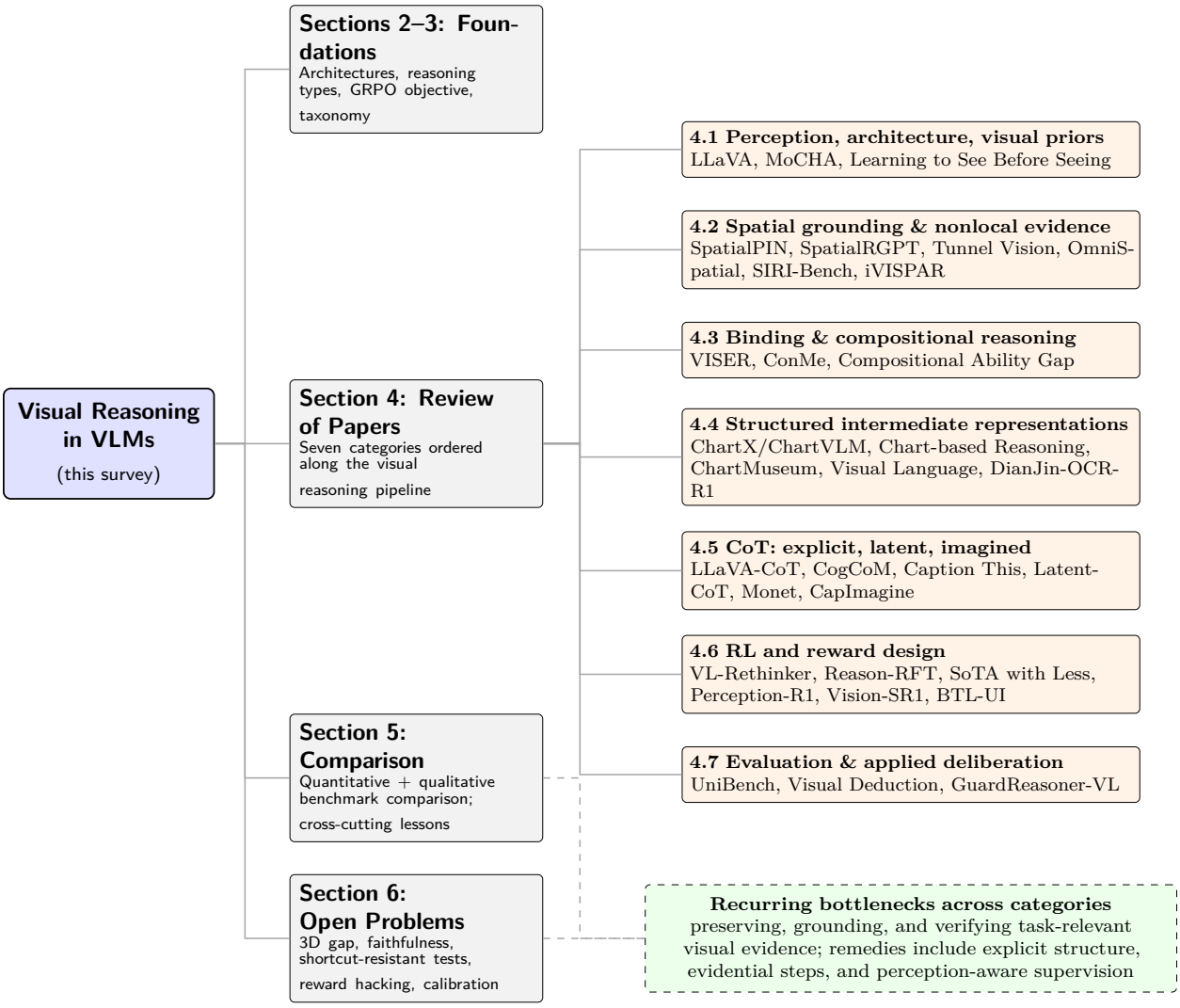


Figure 1: Structural overview of this survey as a mind map. The root node (left) is the survey topic; the middle column lists its major parts (Sections 2–3: background and taxonomy; Section 4: the paper review; Section 5: cross-category comparison; Section 6: open problems); and the right column expands Section 4 into its seven categories, ordered along the visual reasoning pipeline of Figure 2, each listing the works reviewed there. The dashed green box states the unifying thread that Sections 5 and 6 develop across all categories: the recurring bottleneck is unverified intermediate visual evidence, and the recurring remedies are explicit structure, evidential reasoning steps, and perception-level rewards.

## 2 Background

**VLM architectures.** Vision–Language Models (VLMs) have evolved through several architectural paradigms, differing in how visual and textual information are represented and fused. Early models such as CLIP (Radford et al., 2021) adopted a *dual-encoder* architecture, where independent image and text encoders are jointly trained using contrastive learning to align their embeddings in a shared semantic space. This design enables efficient cross-modal retrieval and zero-shot recognition but performs only shallow interaction between modalities, making it insufficient for complex multimodal reasoning. GLIP (Li et al., 2022a) extends this paradigm to grounded vision tasks by jointly aligning textual queries with object-level visual representations for open-vocabulary detection.

To enable richer cross-modal interactions, *fusion* architectures perform multimodal fusion within the network using cross-attention or multimodal transformer layers. Representative models include CoCa (Yu et al., 2022), which combines contrastive and captioning objectives in a unified encoder–decoder framework, allowing it to excel in both representation learning and text generation.

Recent advances have been driven by *LLM-based* architectures, where a pretrained vision encoder is connected to a large language model (LLM) through a lightweight adapter. These models typically consist of three components: (1) a pretrained vision encoder, often a CLIP-style Vision Transformer; (2) a connector, such as a linear projection, MLP, Q-Former, or cross-attention resampler, that maps visual features into the LLM embedding space; and (3) a pretrained LLM that jointly processes visual and textual tokens. Flamingo (Alayrac et al., 2022) introduced gated cross-attention layers that allow frozen LLMs to attend to visual features while preserving their language capabilities. BLIP-2 (Li et al., 2023) proposed the Q-Former, a lightweight querying transformer that efficiently bridges frozen vision encoders and frozen LLMs, substantially reducing the amount of trainable parameters required for multimodal adaptation. LLaVA further simplified this paradigm by replacing the Q-Former with a single trainable linear projection between a frozen CLIP ViT-L/14 encoder and the Vicuna language model. Its two-stage training strategy—feature alignment followed by multimodal instruction tuning—became a practical blueprint for many subsequent open-source VLMs. While these architectures have significantly improved multimodal instruction following and general visual understanding, recent benchmarks demonstrate that architectural improvements alone are insufficient to achieve robust visual reasoning, motivating specialized reasoning datasets, training strategies, and evaluation benchmarks.

Abstracting over these designs, every LLM-based VLM implements the same information flow. An image $I$ is encoded into visual features $V = f_\phi(I)$, a connector maps them into the language model's embedding space, $Z = g_\psi(V)$, and the language model generates a response conditioned on this grounded representation and the query $q$:

$$y \sim \pi_\theta\big(\cdot \mid Z, q\big), \qquad Z = g_\psi\big(f_\phi(I)\big). \tag{2}$$

We use this abstraction throughout: whatever reasoning $\pi_\theta$ is capable of, it can only operate on the information that survives in $Z$. If object identities, attribute bindings, spatial relations, or evidence from distant image regions are lost or entangled in $Z$, no amount of downstream reasoning can recover them. Much of the literature reviewed in Section 4 can be read as competing proposals for what $Z$ should contain and how it should be constructed, supervised, or verified.

**Visual reasoning.** We use *visual reasoning* as an umbrella term for tasks requiring inference beyond direct recognition. Important sub-types are:

- *Spatial reasoning*: judging relative position, orientation, distance, size, 3D layout, perspective, and motion.
- *Compositional reasoning*: correctly binding objects, attributes, and relations, such as "the red cube left of the blue sphere".
- *Diagrammatic and visual-language reasoning*: interpreting symbolic diagrams where meaning is encoded by shapes, arrows, labels, and spatial layout rather than natural image texture.
- *Deductive / abstract reasoning*: inferring hidden visual rules from examples, as in Raven-style progressive matrices.
- *Interactive reasoning*: planning and grounding actions in visual environments, including puzzles, robotics-style scenes, and GUI agents.
- *Safety reasoning*: deliberatively classifying multimodal content or model responses as safe or harmful.

**Reasoning elicitation and post-training.** Earlier VLM work relied mainly on supervised fine-tuning (SFT) and prompting. More recent work adds test-time structure, search, and reinforcement learning (RL). Inference-time methods such as SpatialPIN use external 3D experts without changing the VLM weights. Trace-structured methods such as LLaVA-CoT (Xu et al., 2024) supervise an explicit reasoning format, and CogCoM extends the trace itself with evidential image manipulations such as grounding, zooming, and counting. Grounded-model methods such as SpatialRGPT modify the representation by adding region and depth cues. Data-centric methods such as SoTA with Less select training examples based on MCTS difficulty,

and Chart-based Reasoning transfers reasoning ability from LLMs to VLMs through synthesized rationales. The most recent wave–VL-Rethinker, Reason-RFT, Perception-R1, Vision-SR1, GuardReasoner-VL, DianJin-OCR-R1, and BTL-UI–uses GRPO-style or online RL objectives to reward correct answers, valid formats, self-reflection, visual perception, safety moderation, tool-verified OCR, or GUI action grounding.

**GRPO objective.** Because most of the reviewed RL methods build on Group Relative Policy Optimization (GRPO), we state it once here. For a query $q$, the old policy $\pi_{\theta_{\text{old}}}$ samples a group of $G$ responses $\{o_1, \ldots, o_G\}$, each scored by a scalar reward $R_i$ (answer correctness, format validity, perception consistency, safety, or action grounding, depending on the paper). GRPO maximizes

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\, \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \Bigg[ \frac{1}{G}\sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min\Big( r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\big(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon\big)\hat{A}_{i,t} \Big) \Bigg] \\ - \beta\, \mathbb{D}_{\text{KL}}\big[\pi_\theta \,\|\, \pi_{\text{ref}}\big], \qquad r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})},\end{aligned} \tag{3}$$

where $\epsilon$ is the clipping range, $\beta$ weights the KL penalty against a reference policy $\pi_{\text{ref}}$, and $r_{i,t}(\theta)$ is the token-level importance ratio. Instead of training a value network as in PPO, GRPO computes the advantage by normalizing each reward within its group:

$$\hat{A}_{i,t} = \frac{R_i - \text{mean}\left(\{R_j\}_{j=1}^{G}\right)}{\text{std}\left(\{R_j\}_{j=1}^{G}\right)}, \tag{4}$$

so every token of response $o_i$ shares the group-relative advantage of that response. This formulation explains two recurring design issues in the surveyed papers: when all $G$ rewards in a group are identical, the advantage in Eq. (4) vanishes and the sample provides no gradient (the problem VL-Rethinker's Selective Sample Replay targets), and because $R_i$ is usually computed only from the final answer, the objective by itself does not supervise intermediate visual perception (the gap Perception-R1 and Vision-SR1 address with perception-level rewards).

A key conceptual shift follows: visual reasoning is no longer treated only as a property that should emerge from larger multimodal pretraining. It is treated as a behavior that must be decomposed, verified, rewarded, and stress-tested. With this vocabulary–architectures, reasoning sub-types, and the GRPO objective of Eqs. (3)–(4)–in place, we can now state how the literature is selected and organized.

## 3 Methodology / Taxonomy

**Selection criteria.** We consider recent work on visual reasoning in vision-language models published or publicly released at major conferences, workshops, and preprint venues, including NeurIPS, ICML, CVPR, ACL/NAACL, and arXiv. Our selection is selective rather than exhaustive and is intended to cover the major methodological directions relevant to the framework developed in this survey. We prioritize works that (i) introduce influential architectures, training recipes, or reasoning mechanisms, (ii) provide distinct interventions for improving visual reasoning, or (iii) introduce benchmarks that expose or measure limitations in spatial, compositional, deductive, diagrammatic, chart-based, safety, or interactive visual reasoning. We also include representative works on foundational instruction tuning, visual perception, structured and tool-augmented reasoning, reinforcement learning, OCR and GUI reasoning, and evaluation. Our thematic organization draws on the broader literature on multimodal reasoning (Li et al., 2025c; Wang et al., 2025c), but reorganizes these directions according to the visual reasoning pipeline in Eq. (1).

**Conceptual framework.** Rather than treating spatial reasoning, compositional reasoning, chain-of-thought, reinforcement learning, and evaluation as independent research areas, we analyze them as interventions that operate at different stages of the visual reasoning pipeline (Figure 2). Architecture and visual-prior studies primarily concern *perception*; spatial and binding methods concern *grounding*; chart, caption, latent, and imagination methods concern *intermediate representations*; chain-of-thought methods primarily address *reasoning*; reinforcement learning provides *training* signals that can affect multiple stages; and tool use, self-checks, and guard models introduce forms of *verification*. Benchmarks probe these stages from different

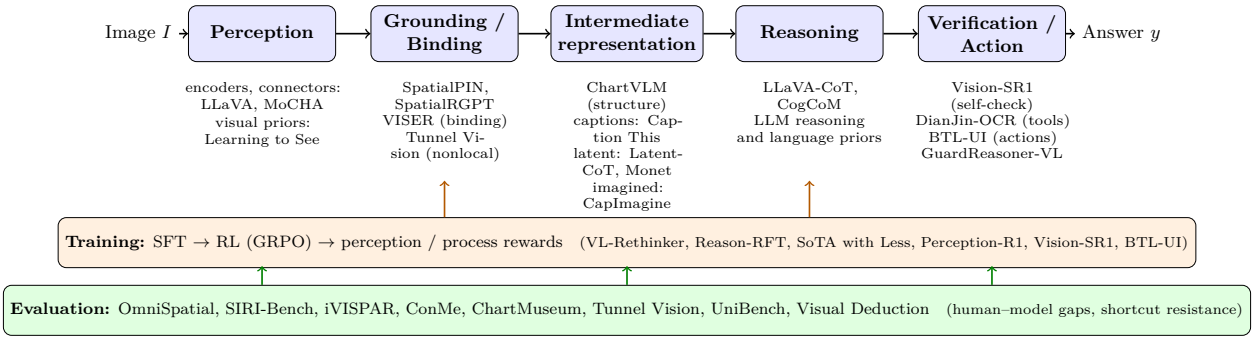


Figure 2: The visual reasoning pipeline used as this survey's organizing framework. An image is processed left to right: *perception* encodes it, *grounding/binding* associates features with the right entities and relations, an *intermediate representation* makes the evidence reasoning-ready, *reasoning* operates on that representation, and *verification/action* checks or executes the result. Below each stage we list the surveyed works whose primary contribution intervenes at that stage (a work can touch several stages; placement is by main contribution). The orange band shows that training–supervised fine-tuning followed by reinforcement learning with perception- or process-level rewards–shapes the stages from below, and the green band lists the benchmarks that probe them.

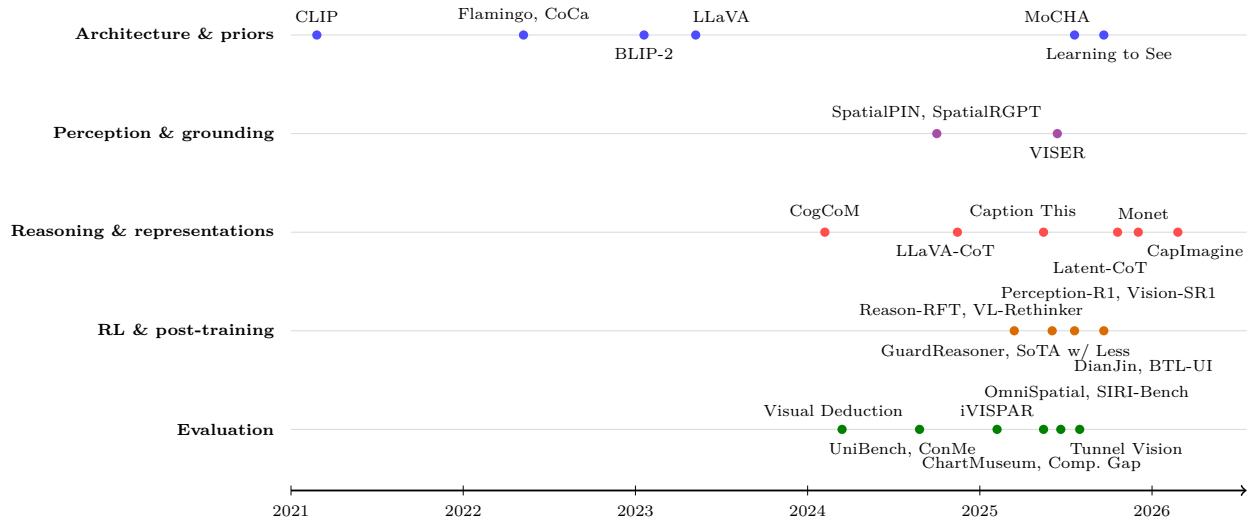


Figure 3: Conceptual evolution of visual reasoning research across five lanes. The field moved from alignment and instruction tuning (2021–2023), through structured reasoning and spatial grounding (2024), to perception-aware RL, binding, latent reasoning, visual priors, and process-level evaluation (2025–2026). Dots mark first public release; representative works only.

perspectives. Section 4 follows this pipeline, while Figure 3 situates the reviewed directions chronologically, from early vision-language alignment and instruction tuning to structured reasoning, grounding, perception-aware reinforcement learning, latent reasoning, and process-level evaluation. Table 2 maps the reviewed papers to these categories.

## 4 Review of Key Papers

The review follows the pipeline of Figure 2 from left to right. We begin with perception and the origins of visual knowledge (Section 4.1), move to spatial grounding and the integration of nonlocal evidence (Section 4.2), and then to the binding problem that links grounding to composition (Section 4.3). We next examine

| Category | Focus | Papers |
|---|---|---|
| Perception, architecture, and visual priors | Encoders, connectors, and where visual knowledge originates | LLaVA (Liu et al., 2023); MoCHA (Pang et al., 2025); Learning to See Before Seeing (Han et al., 2025) |
| Spatial grounding and nonlocal evidence | 3D priors, region/depth grounding, nonlocal integration, spatial benchmarks | SpatialPIN (Ma et al., 2024); SpatialRGPT (Cheng et al., 2024); Tunnel Vision (Berman and Deng, 2025); OmniSpatial (Jia et al., 2025); SIRI-Bench (Zhang et al., 2025a); iVISPAR (Mayer et al., 2025) |
| Binding and compositional reasoning | Attribute–object binding, adversarial and skill composition | VISER (Izadi et al., 2025); ConMe (Huang et al., 2024); Compositional Ability Gap (Li et al., 2025b) |
| Structured intermediate representations | Chart/diagram structure, reasoning transfer, tool-verified OCR | ChartX/ChartVLM (Xia et al., 2024); Chart-based Reasoning (Carbune et al., 2024); ChartMuseum (Tang et al., 2025); Do VLMs Understand Visual Language? (Hou et al., 2024); DianJin-OCR-R1 (Chen et al., 2025) |
| Chain-of-thought: explicit, latent, imagined | Staged traces, image manipulations, captioning bridges, latent thoughts, imagination | LLaVA-CoT (Xu et al., 2024); CogCoM (Qi et al., 2024); Caption This, Reason That (Weng et al., 2025); Latent-CoT (Sun et al., 2025); Monet (Wang et al., 2025d); CapImagine (Li et al., 2026) |
| Reinforcement learning and reward design | GRPO variants, perception/self rewards, data selection, GUI agents | VL-Rethinker (Wang et al., 2025b); Reason-RFT (Tan et al., 2025); SoTA with Less (Wang et al., 2025a); Perception-R1 (Xiao et al., 2025); Vision-SR1 (Li et al., 2025a); BTL-UI (Zhang et al., 2025b) |
| Evaluation and applied deliberation | Unified evaluation, abstract deduction, safety reasoning | UniBench (Al-Tahan et al., 2024); Visual Deductive Reasoning (Zhang et al., 2024); GuardReasoner-VL (Liu et al., 2025) |

Table 2: Taxonomy of the reviewed works. Each row is one of the seven categories of Section 4, ordered along the visual reasoning pipeline of Figure 2; the *Focus* column states the questions and mechanisms the category covers, and the *Papers* column lists the works reviewed under it. Some works span multiple categories; we place each according to its primary contribution.

explicit structured intermediate representations (Section 4.4) and the spectrum of chain-of-thought reasoning from explicit to latent to imagined (Section 4.5), before turning to how reasoning behavior is trained with reinforcement learning (Section 4.6) and how the resulting systems are evaluated and applied (Section 4.7). We discuss benchmark papers in the context of the capabilities they probe and note the main limitation of each benchmark.

### 4.1 Perception, Architecture, and Visual Priors

The modern era of open VLMs begins with LLaVA (Liu et al., 2023), which showed that a frozen CLIP encoder, a trainable linear projection, and an open LLM, tuned in two stages on GPT-generated visual conversations, suffice to produce a fluent multimodal assistant. The recipe's success is precisely what makes it the right starting point for a survey of reasoning: it demonstrated that fluent image dialogue is cheap to obtain, and thereby created the misleading impression that visual reasoning had largely been solved. The benchmarks reviewed later in this section show the opposite–the same recipe that recognizes objects reliably often fails to represent the relations among them.

If instruction tuning under-delivers on reasoning, a natural first hypothesis is that the bottleneck is architectural: a single frozen encoder and a shallow connector may simply discard the visual detail that reasoning needs, before the language model ever sees it. MoCHA (Pang et al., 2025) tests how far this hypothesis can

be pushed. It fuses four complementary vision backbones (CLIP, SigLIP, DINOv2, ConvNeXt) through a sparse Mixture-of-Experts Connector and Hierarchical Group Attention with adaptive gating, on Phi-2 and Vicuna language models, and outperforms comparable open-weight VLMs on understanding and reasoning benchmarks. In the terms of Eq. (2), MoCHA enriches $f_\phi$ and $g_\psi$ so that more visual information survives in $Z$. Yet the same benchmarks that motivated it show that feature-rich models still fail on relational and multi-step tasks: richer $Z$ is necessary but evidently not sufficient.

These observations raise a more fundamental question: where does the reasoning ability of a VLM come from in the first place? Learning to See Before Seeing (Han et al., 2025) gives a striking answer. Through more than 100 controlled experiments (500,000 GPU-hours) across five model scales, the authors show that text-only LLMs already possess substantial *visual priors*–implicit knowledge about the visual world acquired purely from language pre-training–and that these priors decompose into separable perception and reasoning components with different scaling behavior. The reasoning prior originates mainly from reasoning-centric corpora such as code, mathematics, and academic text, and multimodal training largely *unlocks* rather than creates it. The implication for this survey is double-edged. On one hand, it explains why caption-mediated strategies work (Section 4.5): once evidence is textualized, a strong inherited reasoner takes over. On the other hand, it warns that benchmark success may reflect language priors rather than visual grounding–a model can answer a diagram question from what it read about diagrams, not from the diagram in front of it, a failure mode that the evaluation literature repeatedly confirms (Sections 4.4 and 4.7).

Taken together, this first group of works establishes the frame for everything that follows. Instruction tuning provides the interface, architecture determines how much visual information reaches the language model, and language pre-training supplies much of the reasoning machinery– but none of the three guarantees that the machinery operates on the right visual evidence. The critical open question is therefore not only how to see better, but how what is seen is grounded and handed to the reasoner. That hand-off is where we turn next.

### 4.2 Spatial Grounding and Nonlocal Visual Evidence

Spatial reasoning is where the grounding gap shows first and most consistently, because spatial questions cannot be answered from object identity alone. OmniSpatial (Jia et al., 2025) documents the gap at scale: over 8.4K question–answer pairs (1.5K held out for testing) spanning dynamic reasoning, complex spatial logic, spatial interaction, and perspective taking (Figure 4). Even the strongest evaluated model, OpenAI's o3, reaches 56.3% overall accuracy against 92.6% for humans. As with most curated static benchmarks, its multiple-choice single-image format cannot verify embodied behavior, and annotation effort bounds its scale–but a 36-point human–model gap on held-out spatial questions is difficult to explain away.

Two further benchmarks sharpen the diagnosis by controlling *where* the difficulty enters. SIRI-Bench (Zhang et al., 2025a) embeds geometry problems into 9,000 synthesized 3D video scenes, so the mathematical content is held fixed while the evidence becomes visual; models that solve the same mathematics in text degrade sharply, isolating the perception-to-reasoning hand-off as the failure point (its synthetic scene style, however, limits transfer to real footage). iVISPAR (Mayer et al., 2025) performs the complementary manipulation for interaction: a sliding-tile puzzle presented to a VLM agent in 3D, 2D, or text over 300 board configurations. The best VLM completes only 28.7% of 3D episodes versus 89.7% in the simplified 2D view and 45.3% in text, while humans are near-perfect with close-to-optimal paths–planning ability exists, but collapses as the visual grounding burden grows (with the caveat that a single fully observable puzzle family with discrete actions cannot certify real-world competence).

These failures motivated two remedies that intervene at different depths. SpatialPIN (Ma et al., 2024) is training-free: it prompts and interacts with 3D priors produced by off-the-shelf experts–segmentation, depth, 3D reconstruction–and feeds the resulting spatial context back to the VLM, improving spatial VQA and robotics-style task planning at the price of a heavy, expert-dependent pipeline. SpatialRGPT (Cheng et al., 2024) instead builds region and depth grounding into the model and its data pipeline through a 3D-scene-graph data engine, enabling metric, region-specific relational reasoning on SpatialRGPT-Bench that generic VLMs cannot match. The trade-off is the recurring one between modular and learned systems: external experts are flexible but brittle; integrated grounding is robust but requires specialized data and retraining.

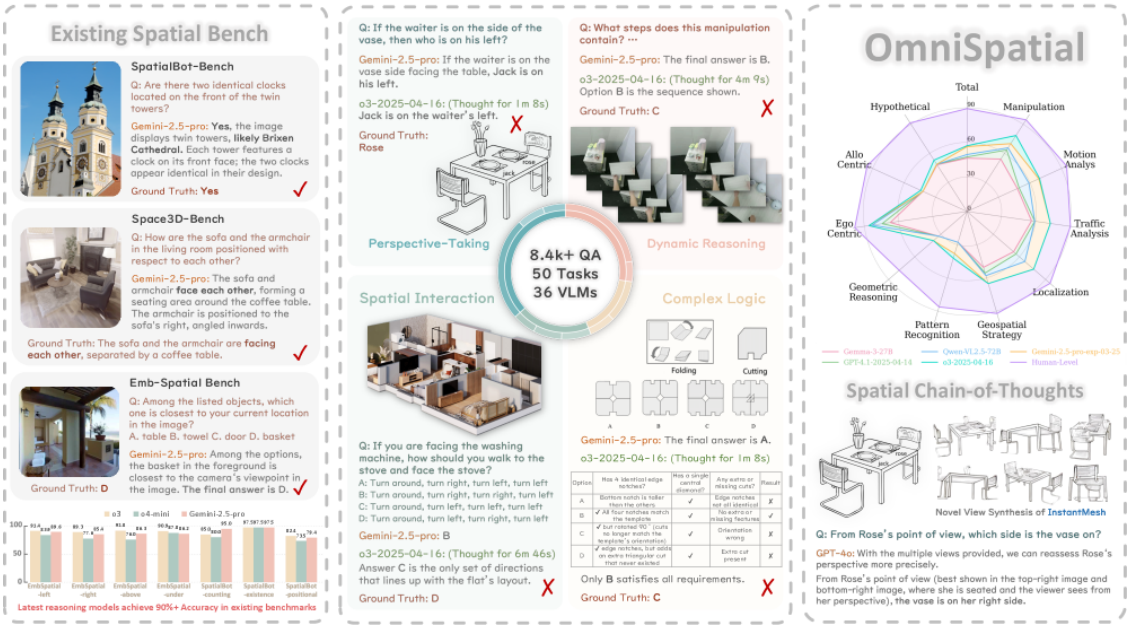


Figure 4: Overview of the OmniSpatial spatial-reasoning benchmark (8.4K+ question–answer pairs, 50 tasks, 36 evaluated VLMs). Left: examples from prior spatial benchmarks, on which recent reasoning models already score above 90%, motivating a harder test. Middle: example questions from OmniSpatial's four task categories–perspective taking, dynamic reasoning, spatial interaction, and complex logic–with answers from frontier models (Gemini-2.5-pro, o3); crosses mark errors, which occur even after long reasoning chains. Right: a radar plot of accuracy per sub-task, showing that all evaluated models (colored lines) fall well inside the human-level polygon, and an illustration of spatial chain-of-thought prompting with novel view synthesis. Figure reproduced from Jia et al. (2025).

However, spatial grounding in the geometric sense is not the whole story. VLMs have Tunnel Vision (Berman and Deng, 2025) shows that models can succeed at local recognition yet fail to *integrate* evidence distributed across an image. Its three task families–comparative perception (holding two images in working memory), saccadic search (evidence-driven jumps between regions), and smooth visual search (following a continuous contour)–are trivial for humans, yet flagship models (Gemini 2.5 Pro, Claude Vision 3.7, GPT o4-mini) barely exceed chance on several variants even though they pass earlier primitive-vision tests. The benchmark is synthetic and narrow by design, which is both its power as a diagnostic and its limitation as a predictor of naturalistic performance. Its message extends the spatial findings: the bottleneck is not only estimating depth or relative position, but actively retrieving and chaining visual evidence across the image–a capability closer to visual attention and working memory than to geometry.

The above results raise the question of what, exactly, the language model would need in order to reason spatially. It is useful to make the target representation explicit. An idealized grounded representation of a scene is a spatial scene graph

$$G = (\mathcal{O}, \mathcal{E}), \qquad o_i = (c_i,\, a_i,\, \mathbf{p}_i) \in \mathcal{O}, \qquad (o_i,\, r,\, o_j) \in \mathcal{E}, \tag{5}$$

where each object $o_i$ carries a category $c_i$, attributes $a_i$, and a position/depth $\mathbf{p}_i$, and edges encode relations $r$ such as *left-of*, *above*, *inside*, or metric distance. Most spatial questions are then simple queries over $\mathcal{E}$–and this is precisely the point. A modern LLM can execute such symbolic queries, along with far harder mathematical and relational inference, as its text-only performance on SIRI-Bench-style problems and its inherited reasoning priors (Section 4.1) demonstrate. What it cannot do is conjure $\mathcal{E}$ from a representation $Z$ in which relations were never made explicit. This is the distinction between improving the *reasoning engine* $\pi_\theta$ and improving the *evidence* $Z$ supplied to it (Eq. (2)): SpatialPIN and SpatialRGPT enrich $Z$ with 3D structure, ChartVLM (Xia et al., 2024) will do so with symbolic chart structure (Section 4.4), and

caption-mediated reasoning will do so with text (Section 4.5). Scaling or improving the GPT-like reasoner alone cannot close the gap if the model has misidentified the objects, mis-bound their attributes, or never retrieved the distant evidence the question requires.

Spatial reasoning research thus establishes three things: there is a genuine perception/grounding bottleneck, quantified by large human–model gaps; explicit depth, region, and 3D representations demonstrably help; and the bottleneck extends beyond geometry to nonlocal evidence integration. What it does not resolve is how a model should keep track of *many* objects and their attributes simultaneously–grounding one relation correctly is not the same as maintaining a consistent assignment of properties to entities across a scene. That failure mode has a classical name, and it is the subject of the next subsection.

### 4.3 Binding and Compositional Reasoning

The binding problem–associating the right attributes and relations with the right entities–sits between perception and reasoning. It helps to distinguish a hierarchy of visual capabilities: (1) recognizing that an object is present; (2) locating it; (3) binding attributes to the correct object; (4) relating multiple objects; (5) retrieving evidence from distant regions; and (6) integrating that evidence into a multi-step inference. Current VLMs are strong at (1)–(2) and degrade progressively through (3)–(6); the spatial and nonlocal results of Section 4.2 concern (4)–(6), while compositional evaluation targets (3)–(4). The failure can be stated formally: a scene containing objects $\mathcal{O} = \{o_1, \ldots, o_n\}$ with attribute sets $\mathcal{A} = \{a_1, \ldots, a_n\}$ is only represented correctly if the model encodes the true assignment $\sigma$ with $a_{\sigma(i)}$ bound to $o_i$. A representation that stores features as an unordered bag,

$$Z \approx \{c_1, \ldots, c_n\} \cup \{a_1, \ldots, a_n\} \quad \text{rather than} \quad Z \approx \{(o_i,\, a_{\sigma(i)})\}_{i=1}^{n}, \tag{6}$$

supports recognition perfectly while making "the red cube left of the blue sphere" indistinguishable from "the blue cube left of the red sphere".

Compositional benchmarks measure exactly this. ConMe (Huang et al., 2024) showed that older compositional suites had become gameable–modern VLMs exploit their artifacts–and rebuilt the test with hard negatives generated by VLMs conversing with an LLM. Performance drops by up to 33% relative to preceding compositional benchmarks once the shortcuts are removed, though the generated negatives can inherit the generator models' own blind spots and the format remains image–text matching rather than free-form reasoning.

If the diagnosis in Eq. (6) is right, then helping the model maintain the assignment $\sigma$ should repair performance without any retraining–and this is what Visual Structures Help Visual Reasoning (VISER) (Izadi et al., 2025) finds. By adding low-level spatial structure to the image itself (for example, horizontal scan lines) and pairing it with a prompt that encourages sequential, spatially aware parsing (Figure 5), VISER improves GPT-4o by 25.0 points on visual search, 26.8 on counting, and 9.5 on spatial relations, and reduces scene-description edit-distance error by 0.32 on 2D datasets. The intervention operates purely on the input $I$, and therefore on $Z$, leaving $\pi_\theta$ untouched–strong evidence that these compositional failures are binding failures rather than knowledge or reasoning failures.

Binding failures also appear at a coarser granularity: entire *skills* fail to compose. The Compositional Ability Gap study (Li et al., 2025b) trains atomic capabilities in isolation–a perception skill, a reasoning skill–and tests their combinations in controlled cross-modal and cross-task splits. Models handle the atomic skills but degrade sharply on compositions, and RL-trained models compose better than SFT-trained ones, which tend to memorize surface formats. The study's deliberately controlled scope (few synthetic skill pairs, largely one model family) limits generalization, but it reframes ConMe's warning at the training level: composition is a property of how capabilities are learned, not only of how inputs are constructed.

This subfield thus converts a vague complaint–"VLMs don't really understand scenes"–into a specific mechanism with specific interventions: binding can be probed adversarially (ConMe), partially repaired at the input (VISER), and influenced by the training paradigm (RL versus SFT). Its central limitation is that none of these gives the model a persistent workspace in which structured evidence–the assignment $\sigma$, the scene graph $G$–is explicitly represented, inspected, and reused across reasoning steps. Constructing such

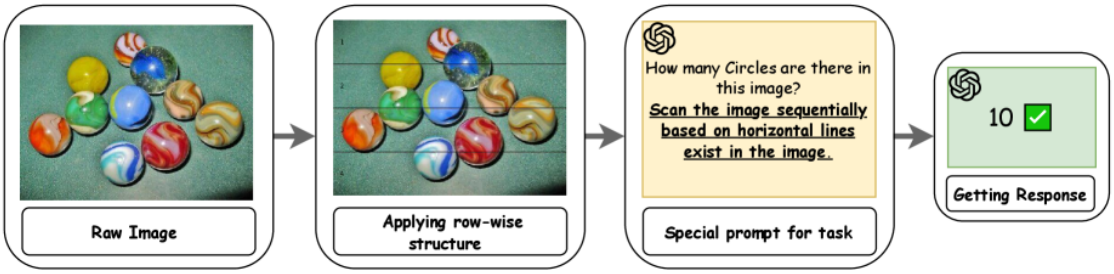


Figure 5: The VISER method for mitigating the binding problem, shown on a counting task. From left to right: the raw input image; the same image augmented with low-level visual structure (horizontal scan lines); a task prompt instructing the model to scan the image sequentially, row by row, along those lines; and the model's now-correct response. The intervention is training-free. Figure reproduced from Izadi et al. (2025).

intermediate representations deliberately, rather than hoping they emerge inside $Z$, is the strategy of the next subsection.

### 4.4 Structured Intermediate Representations: Charts, Diagrams, and OCR

Charts are the cleanest domain in which to study explicit intermediate representations, because the underlying content *is* symbolic: a chart is a rendering of a table. ChartX/ChartVLM (Xia et al., 2024) made this observation architectural. Alongside ChartX, a multi-task benchmark covering 18 chart types and 22 topics, ChartVLM decouples perception from cognition: a first stage extracts the chart's structure, and reasoning operates on that intermediate representation rather than on raw visual tokens, outperforming generic VLMs across ChartX tasks. Chart-based Reasoning (Carbune et al., 2024) strengthens both ends of the same pipeline: continued pretraining of PaLI-3 on an improved chart-to-table translation task (better perception into the symbolic intermediate), followed by fine-tuning on a 20× larger set of LLM-synthesized rationales (better reasoning over it). The resulting 5B model attains state-of-the-art ChartQA performance, rivaling systems ten times larger–a concrete demonstration that reasoning ability can be *transferred* from a language model once the evidence is symbolic.

However, this representation-level solution has a ceiling, and ChartMuseum (Tang et al., 2025) measures it precisely. Its 1,162 expert-annotated questions over real charts from 184 sources explicitly separate questions answerable by *textual* reasoning (values that could be read from a table) from questions requiring genuinely *visual* reasoning–comparing areas, slopes, or unlabeled marks (Figure 6). Humans reach about 93% while the best model at publication time reaches about 63%, and models lose 35–55 points when moving from the textual to the visual subset. The implication is sharp: chart-to-table pipelines solve exactly the subset that was already text-solvable, and the residual questions are those whose evidence no table captures. (The benchmark's modest size, string-matched answers, and web-sourced charts–which future models may see in pretraining–qualify but do not overturn this conclusion.)

Diagrams push the same concern from values to relations. The study of Hou et al. (2024) tests entities and relationships separately in a synthetic-plus-real diagram suite and finds that LVLMs identify entities well but understand relationships far less, with background knowledge frequently masking the failure–a diagram question answered from priors looks identical, in aggregate metrics, to one answered from the diagram. This is the visual-prior warning of Section 4.1 made concrete, though the largely synthetic suite may underestimate models tuned on natural diagrams.

Where the symbolic content is text itself, the intermediate representation can be *verified* rather than merely extracted. DianJin-OCR-R1 (Chen et al., 2025) treats OCR–where hallucination is maximally costly because outputs must transcribe exactly what is printed–as a reasoning-and-tool interleaving problem: the model first reads the image with its own perception, then calls specialized expert OCR tools, and finally re-thinks by comparing the two before answering, trained with SFT followed by GRPO-based reinforcement fine-tuning. The 3B and 7B models (built on Qwen2.5-VL) outperform their base models and larger proprietary VLMs

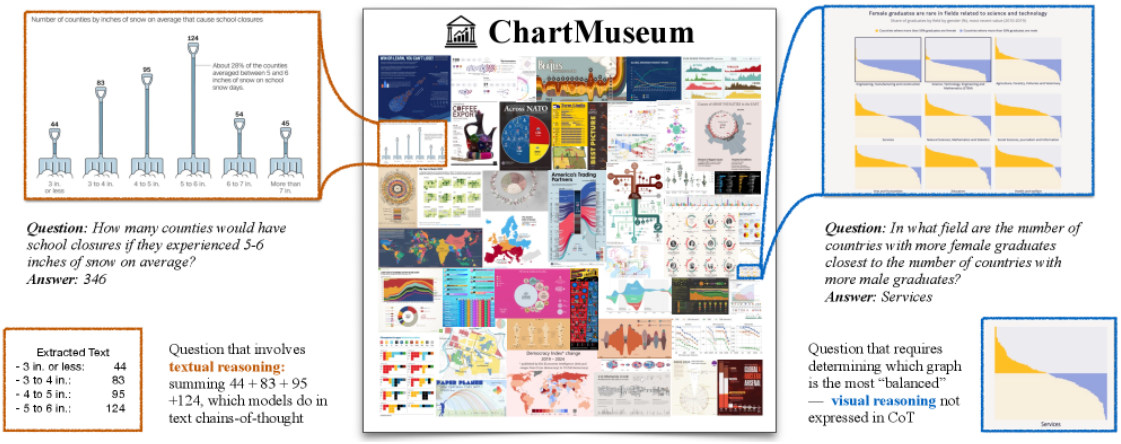


Figure 6: The ChartMuseum benchmark (1,162 expert-annotated questions over real-world charts from 184 web sources). Center: a collage of the benchmark's charts, illustrating their visual diversity. Left: an example question solvable by *textual* reasoning–the relevant values can be extracted as text and summed in an ordinary text chain-of-thought. Right: an example question requiring *visual* reasoning–judging which small-multiple panel is most "balanced" cannot be expressed as operations over extracted text. Figure reproduced from Tang et al. (2025).

on seal and formula recognition. The approach is bounded by tool coverage and has been demonstrated on narrow subtasks, but it introduces an idea that recurs in Section 4.6: an intermediate representation is most valuable when something external–a tool, a second pass, a reward–can check it.

Collectively, these works establish that explicit symbolic intermediates are highly effective exactly where the content is symbolic: tables, axis values, transcriptions. They fail, by construction, where the evidence is irreducibly visual (ChartMuseum's visual subset) or where the symbols' meaning lies in spatial arrangement (diagrams). And they leave open what the intermediate representation should be for general natural scenes, where no canonical symbolic form exists. The field's dominant answer–let the model generate its own intermediate representation as a chain of thought–comes in several competing forms, examined next.

### 4.5 Chain-of-Thought: Explicit, Latent, and Imagined Reasoning

Chain-of-thought methods differ, at bottom, in the *type* of the intermediate variable placed between evidence and answer. Writing $c$ for that variable, all variants share the factorization

$$p(y \mid Z, q) \;=\; \sum_{c} p(c \mid Z, q)\, p(y \mid c, Z, q), \tag{7}$$

but instantiate $c$ differently (a survey-level schema): *explicit textual CoT* takes $c$ to be a natural-language trace; *visual or tool-interleaved CoT* lets $c$ contain image operations, so the evidence itself is updated between steps ($Z_{t+1} = g_\psi(f_\phi(m_t(I_t)))$ for a manipulation $m_t$); *latent CoT* replaces $c$ by continuous hidden states $z_{1:T}$, making the sum an intractable integral that must be trained variationally; and *imagined CoT* takes $c$ to be a generated description (or depiction) of a visual state that does not exist in the input. The works below each occupy one of these positions, and their disagreements are informative.

LLaVA-CoT (Xu et al., 2024) establishes the explicit-textual baseline: a four-stage format (summary, caption, reasoning, conclusion) trained from 100K stage-annotated traces with stage-level beam search, improving clearly over its Llama-3.2-Vision base on reasoning-intensive benchmarks. Its weakness defines the agenda for everything after it: a well-formatted trace is not necessarily a faithful one, since every stage can rest on a false visual claim that nothing in training penalizes.

CogCoM (Qi et al., 2024) answers by making the trace *evidential*. Its 17B model reasons through a Chain of Manipulations–grounding, cropping and zooming, drawing auxiliary marks, counting, calculating–where

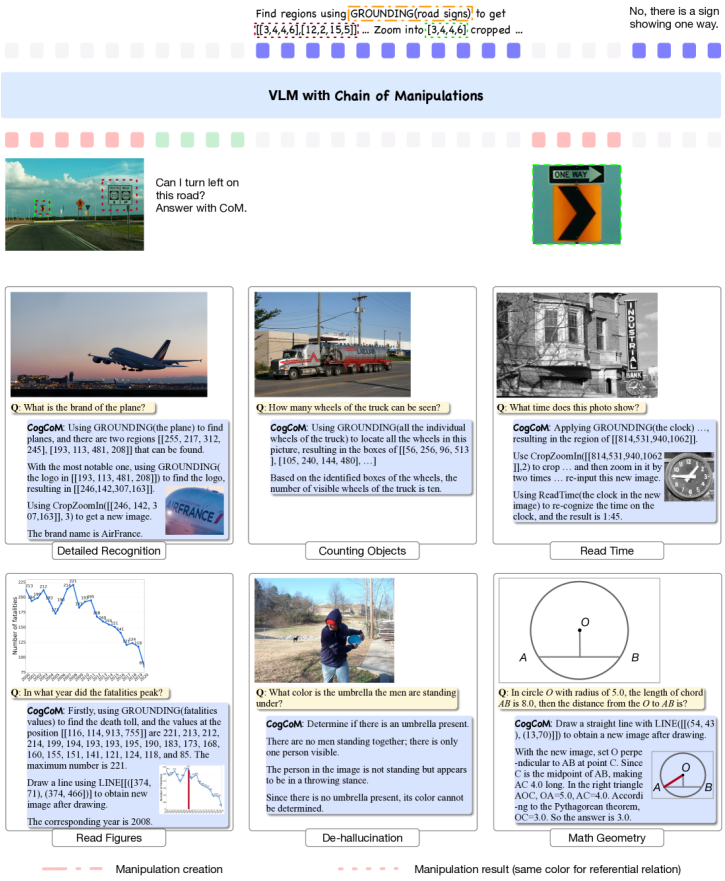


Figure 7: CogCoM's Chain of Manipulations, in which reasoning steps are concrete image operations rather than only text. Top: schematic of one reasoning turn–the model grounds a referred region (road signs), receives the cropped/zoomed evidence back as input, and answers on that basis; colors link each manipulation request to its returned result. Bottom: six example tasks (detailed recognition, counting objects, reading time, reading figures, de-hallucination, math geometry), each solved by a chain in which operations such as **GROUNDING**, **CropZoomIn**, and auxiliary line drawing produce intermediate visual evidence that feeds the next step, making every intermediate claim externally checkable. Figure reproduced from Qi et al. (2024).

each step operates on the image and its output feeds the next (Figure 7); training chains are synthesized by an LLM planner executing visual tools, and the model reaches state-of-the-art results at its scale on VQA, grounding, and detailed-reasoning benchmarks. The cost is a fixed manipulation vocabulary and the propagation of tool errors, but the conceptual gain is real: intermediate claims are tied to operations whose results anyone–including a reward function–can inspect.

A complementary line of work asks not how to structure the trace but *where reasoning actually happens*. Caption This, Reason That (Weng et al., 2025) profiles VLMs with a cognitive-science-style battery along perception, attention, and memory, finding near-ceiling category recognition alongside persistent deficits in spatial understanding and selective attention–and, crucially, that having the model first caption the image and then reason over its own caption measurably improves accuracy. Read together with the visual priors of Section 4.1, this suggests that current VLMs reason best in their language space: the caption is a lossy but reasoning-ready intermediate representation, and the bottleneck is getting the right information into it.

If reasoning in language space is expensive and verbose, one might hope to move it into the model's hidden states instead. Latent Chain-of-Thought for Visual Reasoning (Sun et al., 2025) develops the principled version of this idea: treating the reasoning chain as a latent variable, it trains with amortized variational inference (optimizing an evidence lower bound rather than the intractable marginal in Eq. (7)), adds a sparse

| Intermediate repr. | Representative works | Strengths | Main concern |
|---|---|---|---|
| Explicit textual trace | LLaVA-CoT; Caption This, Reason That | Interpretable; exploits inherited language reasoning | May be ungrounded; verbose; lossy for visual detail |
| Visual / tool trace | CogCoM; DianJin-OCR-R1 | Evidence-grounded; steps externally checkable | Fixed operation vocabulary; tool errors; expensive |
| Symbolic structure | ChartVLM; Chart-based Reasoning | Reliable where content is symbolic; enables transfer from LLMs | Ceiling on genuinely visual questions (ChartMuseum) |
| Latent states | Latent-CoT; Monet | Compact; potentially scalable; no verbalization cost | Hard to verify; can be causally disconnected (CapImagine) |
| Textual imagination | CapImagine | Easy to reason over; empirically strong | May lose visual information in verbalization |

Table 3: The design space of intermediate representations for visual chain-of-thought. Each row is one way of instantiating the intermediate variable $c$ that stands between the visual evidence and the answer in Eq. (7): an explicit natural-language trace, a trace of image operations or tool calls, an extracted symbolic structure, continuous latent states, or an explicitly generated textual "imagination" of visual content. Columns list representative surveyed works, the main strengths, and the main concern for each choice.

diversity-seeking reward to obtain token-level signals without a learned reward model, and replaces best-of-$N$ selection with Bayesian inference-time scaling, reporting consistent gains in accuracy, generalization, and interpretability across seven multimodal reasoning benchmarks. Monet (Wang et al., 2025d) pushes the latent program further and makes it *visual*: the model generates continuous latent embeddings as intermediate visual thoughts–standing in for operations such as cropping key regions, generating new visual states, or drawing auxiliary lines–rather than emitting actual images. Addressing the two obstacles that had made this impractical, the high cost of latent–vision alignment and the lack of supervision over latent embeddings, it combines SFT on Monet-SFT-125K (125K text–image interleaved CoTs spanning real-world, chart, OCR, and geometry data) with Visual-latent Policy Optimization (VLPO), an RL method that incorporates the latent embeddings directly into policy-gradient updates. Monet-7B (built on Qwen2.5-VL-7B) improves consistently across perception and reasoning benchmarks and generalizes out-of-distribution to abstract visual reasoning–in effect, a latent-space successor to CogCoM's explicit manipulations.

However, the latent program has also drawn a direct empirical challenge. Imagination Helps Visual Reasoning, But Not Yet in Latent Space (Li et al., 2026) perturbs latent-reasoning systems and finds two disconnects: large changes to the input barely move the latent tokens (input–latent disconnect), and perturbing the latent tokens barely changes the answer (latent–answer disconnect)–evidence that, in the systems examined, the latents are not causally load-bearing. Their alternative, CapImagine, performs imagination *explicitly* in text and outperforms substantially more complex latent-space baselines across visual reasoning benchmarks. The critique should be read with care: it concerns the latent-reasoning systems evaluated there, and whether training regimes like Monet's VLPO–which explicitly supervises the latents–escape it is an open empirical question. But the burden of proof it establishes is exactly right: a latent trace that can be ablated without consequence is not a reasoning trace, however much it improves benchmark averages.

The picture that emerges is a genuine conceptual tension rather than a solved problem. The empirical evidence currently favors explicit, text-mediated intermediates: captioning helps, explicit imagination beats latent baselines, and explicit traces are the only ones a human or a reward can audit. Yet explicit traces are costly, lossy, and–as LLaVA-CoT shows–formattable without being faithful, while latent approaches are young and improving quickly under direct supervision (Monet). What all variants share is that their value ultimately depends on whether the intermediate state is *caused by the image and causally used for the answer*–a property that neither format guarantees by construction and that supervision must therefore enforce. How to build such supervision is the subject of the next subsection.

### 4.6 Reinforcement Learning and Reward Design

The RL wave applies the GRPO objective of Eq. (3) to visual reasoning, and its internal history is best understood as an evolution of the reward. It is useful to write the reward of a rollout $o$ as a weighted combination

$$R(o) \;=\; \lambda_{\text{ans}} R_{\text{ans}} + \lambda_{\text{fmt}} R_{\text{fmt}} + \lambda_{\text{vis}} R_{\text{vis}} + \lambda_{\text{act}} R_{\text{act}}, \tag{8}$$

where $R_{\text{ans}}$ verifies the final answer, $R_{\text{fmt}}$ the trace format, $R_{\text{vis}}$ the visual-perception stage, and $R_{\text{act}}$ grounded actions. The decomposition makes the field's central lesson visible at a glance: with $\lambda_{\text{vis}} = \lambda_{\text{act}} = 0$, nothing in Eqs. (3)–(8) penalizes a policy that reaches correct answers through language priors while ignoring the image–the "thinking over seeing" failure.

The first generation optimized answers and format. VL-Rethinker (Wang et al., 2025b) confronts two practical obstacles: vanishing advantages–when all $G$ rollouts in a group earn equal reward, Eq. (4) yields zero gradient–addressed by Selective Sample Replay, and shallow reflection, addressed by Forced Rethinking, which appends a rethinking trigger to rollouts so that self-correction is incentivized rather than merely formatted. From a Qwen2.5-VL base it reaches state-of-the-art or open-source state-of-the-art results on MathVista (∼80.4%), MathVerse (∼63.5%), MathVision, MMMU-Pro, EMMA, and MEGA-Bench. Reason-RFT (Tan et al., 2025) shows that the RL stage interacts strongly with its initialization: SFT on curated CoT activates reasoning, and GRPO then restores the generalization that pure SFT destroys, outperforming open and proprietary baselines on visual counting, structural perception, and spatial transformation, with better domain-shift robustness and few-shot regimes that beat full-dataset SFT. SoTA with Less (Wang et al., 2025a) adds the complementary data-side lever: scoring each sample's difficulty by the MCTS search effort needed to solve it, and keeping only informative examples, matches or beats full-dataset self-improvement across VLM backbones at a fraction of the data–structure enters through example selection rather than the reward.

These findings suggest that RL reliably improves *answering*; whether it improves *seeing* is a separate question, and the second generation answers it in the negative before repairing it. Perception-R1 (Xiao et al., 2025) observes that RL with verifiable answer rewards can raise final accuracy without improving multimodal perception, and therefore sets $\lambda_{\text{vis}} > 0$: visual annotations extracted from CoT trajectories serve as references, and an LLM judge scores the consistency of the policy's response with them. Trained from Qwen2.5-VL-7B on only 1,442 Geometry3K-derived samples, it outperforms baselines trained on 10–100× more data on multimodal math benchmarks–evidence that rewarding evidence is more sample-efficient than rewarding answers. Vision-SR1 (Li et al., 2025a) removes the external judge by making the perception reward *self-generated*: the model first emits a self-contained visual description, is re-prompted *without* the image to answer from that description alone, and earns the visual reward when the description suffices (Figure 8); a multi-reward objective separates the visual and language signals. Across general, science, and multimodal-math suites it reduces visual hallucination and language shortcuts without extra reward-model GPUs. In pipeline terms, Vision-SR1 operationalizes verification: the second rollout is a causal test of whether the intermediate representation carried the evidence–precisely the property that Section 4.5 identified as the crux of the explicit-versus- latent debate.

Process rewards extend naturally from perception to action. BTL-UI (Zhang et al., 2025b) decomposes GUI interaction into Blink (attend to relevant regions), Think (reason about the next step), and Link (emit executable commands), each shaped by process rewards ($\lambda_{\text{act}} > 0$). The 7B model reaches 87.2% grounding accuracy on ScreenSpot versus 84.8% for Qwen2.5-VL-7B and 82.4% for Aria-UI; the 3B model reaches 27.1% on ScreenSpot-Pro versus 17.8% for UI-R1; and on AndroidControl-Low the 7B model attains 88.0% versus 85.2% for OS-Atlas-Pro-7B and 66.5% for GUI-R1-7B–the constructive counterpart to iVISPAR: interactive visual planning is hard, but decomposed attention–reasoning–action rewards measurably help.

The RL literature thus converges on a clear principle: policy optimization amplifies whatever the reward measures, no more. Answer-only rewards produce better answerers; perception, process, and action rewards move supervision earlier in the pipeline, where the grounding failures documented in Sections 4.2–4.5 actually occur. The costs are equally clear–judge models, second rollouts, curated references, and new reward-hacking surfaces–and every reward term is only as good as the verifier behind it. This dependence on verification makes the final question unavoidable: how do we evaluate whether any of this produces reasoning that is genuinely grounded in the image?

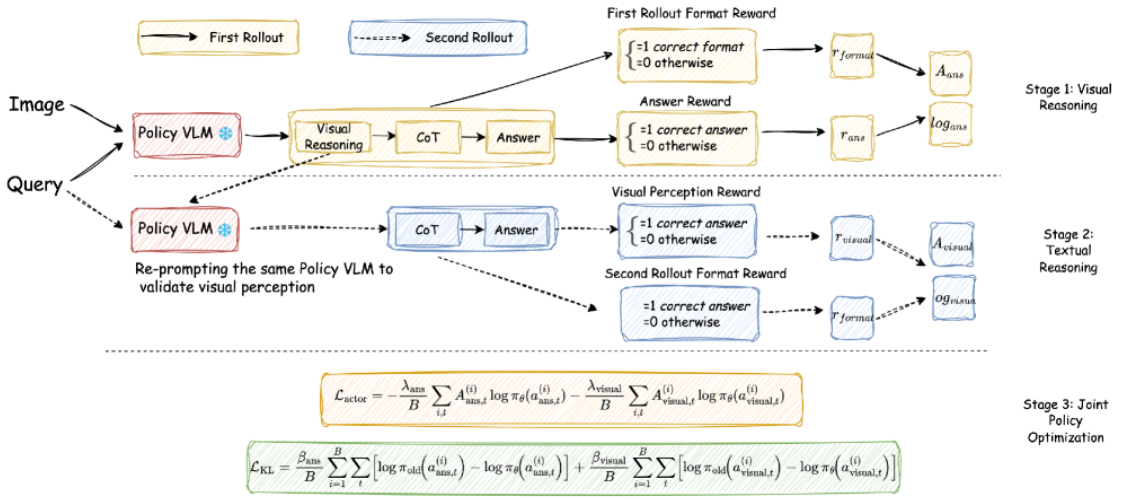


Figure 8: Vision-SR1's self-rewarding training decomposition. Stage 1 (solid arrows): given the image and query, the policy VLM produces a self-contained visual description, a chain of thought, and an answer, earning a format reward and an answer-correctness reward. Stage 2 (dashed arrows): the *same* policy is re-prompted *without* the image, using only its own visual description; if it still answers correctly, the description demonstrably carried the necessary evidence, and a separate visual-perception reward is granted. Stage 3: all reward signals are combined in a joint policy optimization objective (boxed equations) that keeps the visual and language signals separate. This second, image-free rollout is a causal test of the intermediate representation, requiring no external judge model. Figure reproduced from Li et al. (2025a).

### 4.7 Evaluation and Applied Deliberation

Training advances are only meaningful if evaluation can detect what they change, and the broadest evidence says that it often cannot. UniBench (Al-Tahan et al., 2024) re-implements more than 50 VLM benchmarks under one protocol and evaluates roughly 60 models, finding that scaling model and data size improves recognition far more than relational and reasoning abilities, and that data quality matters more than quantity. Because it aggregates existing benchmarks, it also inherits their artifacts and language shortcuts–the very confounds documented throughout this survey–and its unified multiple-choice protocol cannot probe generative, interactive, or long-horizon behavior. Aggregate leaderboards, in other words, systematically overstate reasoning progress.

At the opposite extreme of specificity, Visual Deductive Reasoning (Zhang et al., 2024) strips language priors away almost entirely: Raven-style progressive matrices from three RPM sources (Mensa, IntelligenceTest, RAVEN) require inducing abstract visual rules with no factual knowledge to fall back on. Frontier VLMs such as GPT-4V and Gemini often remain near chance, and the authors trace the failure to perception of the confounding abstract patterns rather than to rule manipulation; strategies effective for LLMs, such as few-shot prompting and self-consistency, do not transfer. The domain is narrow and synthetic, but as a purified probe of stages (1)–(6) with priors removed, it marks a lower bound on genuinely visual reasoning that has barely moved.

Deliberative reasoning also transfers to applied settings where the output is a judgment rather than an answer. GuardReasoner-VL (Liu et al., 2025) trains a guard model to reason before moderating multimodal content: SFT on GuardReasoner-VLTrain (123K samples with 631K reasoning steps across text, image, and text–image inputs), then online RL with a length-aware safety reward balancing accuracy, format, and token cost, yielding an average F1 advantage of 19.27 points over the runner-up. Beyond its practical value, it demonstrates that the reason-then-decide pattern–and its reward design–generalizes beyond VQA, while inheriting the same open question as every explicit-trace system: whether the stated reasons are the operative ones.

Evaluation, then, is not a bookkeeping appendix to this survey: it is central to determining how strongly we can interpret claims about visual reasoning. Recent work increasingly shows that aggregate scores can conceal reasoning failures, that priors-free probes remain challenging, and that deliberation can be trained in applied domains. What remains missing is a broadly accepted methodology for verifying that a model's answer causally depended on the image–visual-versus-textual splits (ChartMuseum), re-prompting without the image (Vision-SR1), input perturbations (CapImagine), and nonlocal probes (Tunnel Vision) are all steps toward such a methodology, and we return to them in the discussion.

## 5 Comparison and Discussion

Each sub-field above closed with its own synthesis; this section reads those syntheses against one another. We first compare the evaluation benchmarks across categories, quantitatively in Figure 9 and qualitatively in Table 4, and then draw the cross-cutting lessons that no single sub-field states on its own.

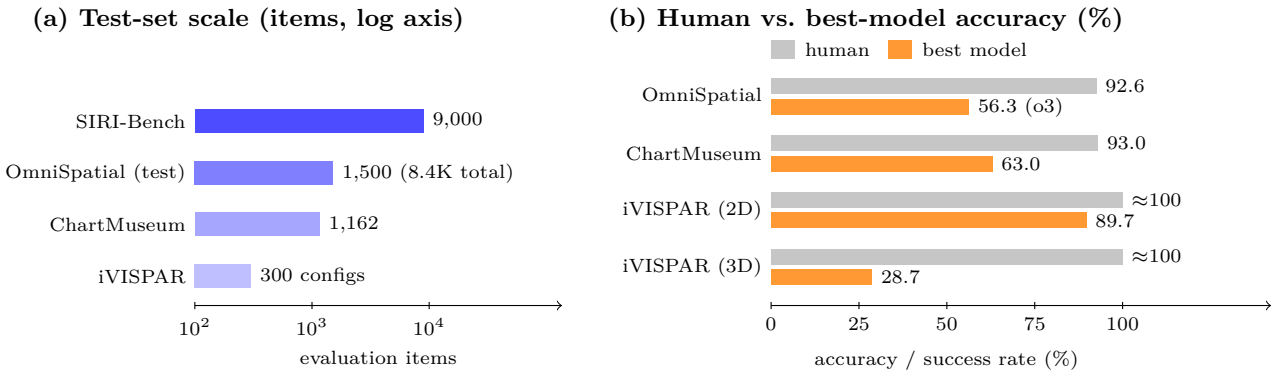


Figure 9: Quantitative comparison of selected evaluation benchmarks. (a) Test-set scale on a logarithmic axis. UniBench is omitted because it aggregates more than 50 existing benchmarks across approximately 60 VLMs, rather than defining a single test set; ConMe and the Compositional Ability Gap study are likewise omitted because they report relative performance drops rather than a single test-set size. (b) Human performance versus the best reported model performance. On OmniSpatial, o3 achieves 56.3% versus 92.6% human performance; on ChartMuseum, the best reported LVLM achieves approximately 63% versus 93% human performance; and on iVISPAR, the best VLM agent achieves 89.7% on 2D tasks but only 28.7% on 3D tasks, while human performance is near-perfect. Visual Deductive Reasoning is omitted because frontier VLMs perform near chance. The reported values are reproduced from the respective papers. They are intended to illustrate differences in benchmark scale and human–model gaps rather than provide a strictly standardized cross-benchmark ranking, since task definitions, evaluation protocols, and metrics differ across benchmarks.

We then organize the discussion around five questions that the surveyed literature, read as a whole, allows us to pose sharply. Throughout, we distinguish empirical findings reported by individual papers from interpretations that are inferences of this survey.

**Q1: Is visual reasoning primarily a perception problem or a reasoning problem?** Among the works surveyed here, the evidence more often locates the bottleneck on the evidence side of Eq. (2), though this reflects the specific set of papers we review rather than a literature-wide consensus. VISER repairs counting and search by structuring the *input* while leaving the reasoner untouched; SIRI-Bench (Section 4.2) holds the mathematics fixed and observes collapse when evidence becomes visual; ChartMuseum (Section 4.4) shows that models solve the text-solvable subset and fail the visual one; Tunnel Vision (Section 4.2) shows chance-level integration of distributed evidence despite intact local recognition; and Visual Deductive Reasoning's authors trace near-chance RPM performance to perception of the confounding patterns rather than to rule manipulation. Meanwhile, language models appear to bring substantial reasoning capability of their own, much of which may be inherited from language pre-training (Han et al., 2025)–though this same result is also a source of concern for evaluation, since it means models can sometimes produce correct-looking answers without grounding them in the image (Section 4.7). The honest qualification is that the two are not fully

| Benchmark | Capability probed | Format / scale | Key finding | Main limitation |
|---|---|---|---|---|
| UniBench | Unified recognition vs. reasoning | 50+ aggregated benchmarks; ∼60 VLMs | Scaling helps recognition far more than relations/reasoning | Inherits artifacts of aggregated benchmarks; static multiple-choice protocol |
| ConMe | Compositional reasoning | Adversarial image–text matching | Up to 33% drop once dataset shortcuts are removed | Hard negatives generated by VLM/LLM pipeline; confined to matching format |
| Compositional Gap | Cross-skill composition after post-training | Controlled cross-modal / cross-task splits | Atomic skills do not compose; RL composes better than SFT | Few synthetic skill pairs; mostly one model family |
| Visual Deduction | Abstract rule induction | Raven-style puzzles (3 RPM sources) | Frontier VLMs near chance despite easy human solutions | Small, synthetic, single puzzle domain |
| Tunnel Vision | Nonlocal visual reasoning | Comparative / saccadic / smooth search tasks | Flagship VLMs barely exceed chance on human-trivial variants | Synthetic primitives; narrow task family |
| *Diagram and chart* | | | | |
| Visual Language | Diagram entity/relation understanding | Synthetic + real diagram suite | Entities recognized, relations not; priors mask failures | Largely synthetic diagrams, cleaner than real ones |
| ChartMuseum | Visual vs. textual chart reasoning | 1,162 expert questions; 184 real sources | 35–55 point drop on visual-reasoning subset; best model ∼63% vs. 93% human | Modest size; static string-matched QA; web charts may leak into pretraining |
| *Spatial and interactive* | | | | |
| OmniSpatial | Broad spatial reasoning incl. perspective | 8.4K QA pairs (1.5K test); multiple choice | Best model (o3) 56.3% vs. 92.6% human; perspective and dynamic reasoning hardest | Static single images; no embodied verification; annotation-bound scale |
| SIRI-Bench | Spatial intelligence in 3D video | 9,000 video–QA triplets, synthesized 3D scenes | Text-strong models degrade sharply with visual 3D grounding | Synthetic scene style; domain gap to real video |
| iVISPAR | Interactive visual planning | 300 configs × 3D/2D/text | Best VLM 28.7% (3D) vs. 89.7% (2D) vs. near-perfect humans | Single fully observable puzzle family; discrete actions |

Table 4: Qualitative comparison of the evaluation benchmarks surveyed in this paper, grouped by the capability family they probe. For each benchmark we list the capability, format and scale, headline finding, and the main limitation discussed in the text; Figure 9 shows the corresponding quantitative comparison.

separable: the Compositional Ability Gap study (Section 4.3) shows that even correctly perceived skills fail to *compose* depending on how training was done, so perception-side fixes alone will not finish the job.

**Q2: What should the intermediate representation look like?** Table 3 summarizes the design space for the chain-of-thought variants of Section 4.5 (raw visual tokens, regions and depth, symbolic structures, captions, tool outputs, latent states, and imagined descriptions). Reading this alongside the pipeline interventions discussed elsewhere in the survey, the more useful choice appears to depend on the task rather than on a single universal format: 3D and depth structure support spatial reasoning (SpatialRGPT), symbolic representations support chart reasoning (ChartVLM), expert-verified transcriptions support OCR reasoning (DianJin-OCR-R1), attention–action decompositions support GUI agency (BTL-UI), and captions or explicit imagination can support reasoning about general scenes (Caption This, Reason That; CapImagine)–though only the last group of these is directly captured in Table 3. Across these domains, a recurring design consideration is *verifiability*: representations become particularly useful when their contents can be checked by a tool, a second rollout, a reward signal, or a human. This criterion also highlights a current limitation of latent representations, whose internal contents are generally harder to inspect or verify. Our reading therefore favors task-matched representations that expose or otherwise make checkable the visual evidence needed for the downstream reasoning process, rather than a single representation format for all tasks.

**Q3: Does better language reasoning automatically produce better visual reasoning?** Not automatically, and the surveyed evidence on this point is indirect rather than a direct test of language-reasoning strength alone. UniBench finds that scaling model and data size improves recognition far more than relational reasoning, which speaks to scaling broadly rather than to language reasoning specifically; MoCHA shows that aggressive encoder- and connector-side engineering does not close the relational gap, which bears more directly on the perception side of Eq. (2) than on the language-reasoning side. Taken together, these results are consistent with–but do not directly establish–the idea that a stronger language reasoner does

not by itself repair grounding failures. The visual-priors results sharpen this into a genuine double edge: inherited language reasoning is plausibly what makes caption-mediated strategies work, and simultaneously what lets models answer diagram and knowledge-aligned questions without using the image (Hou et al., 2024). Better language reasoning may therefore raise the ceiling of what grounded evidence can support, while also raising the risk that benchmarks mistake priors for perception–so the honest answer is conditional: language reasoning helps only once the relevant evidence has reached the reasoner, and none of the works reviewed here shows it compensating for evidence that never arrived.

**Q4: Does RL teach visual reasoning, or only better answer generation?** The evidence suggests that RL can improve visual reasoning, but its effect depends strongly on what the reward supervises. In the decomposition of Eq. (8), answer-only rewards directly constrain the final response and therefore do not necessarily provide supervision for perception, grounding, or intermediate reasoning. Perception-R1 illustrates this distinction by showing that improvements in final-answer accuracy do not automatically translate into corresponding improvements in visual perception. This motivates reward designs that place supervision at earlier stages of the pipeline: Perception-R1 introduces perception-level rewards, Vision-SR1 uses self-generated evidence and multi-reward optimization, and BTL-UI incorporates process- and action-level supervision for GUI interaction, with the grounding-accuracy gains reported in Section 4.6. RL is not limited to perception-level supervision, however: VL-Rethinker reports gains on multimodal mathematical reasoning, and the Compositional Ability Gap study finds that RL-trained models compose previously separate skills better than SFT-trained ones. We read these two results as complementary to, rather than as direct evidence for, the perception-reward findings above, since neither isolates whether the improvement is specifically perceptual rather than a more general effect of RL on generalization. Our synthesis is therefore more nuanced than a strict distinction between "reasoning" and "answer generation": RL can improve both, but the reward determines which behaviors receive direct optimization pressure. Answer-level rewards primarily constrain the final output, whereas perception-, process-, and action-level rewards provide more explicit incentives for the intermediate visual evidence and behaviors on which reliable visual reasoning depends.

**Q5: How can we know that the model actually used the image?** The human–model gaps are now quantified (Figure 9): roughly 36 points on OmniSpatial (92.6% vs. 56.3%), 30 on ChartMuseum (93% vs. ∼63%), and more than 70 on iVISPAR's 3D setting (≈100% vs. 28.7%)–but a gap measures how much models fail, not *why*. The surveyed literature has begun to assemble a genuine verification toolkit: visual-versus-textual splits that isolate image-dependent questions (ChartMuseum); re-answering without the image, which tests whether the intermediate description carried the evidence (Vision-SR1); input and latent perturbations that test causal dependence (CapImagine); nonlocal probes that cannot be solved from local features (Tunnel Vision); adversarial hard negatives that remove shortcuts (ConMe); and priors-free abstract probes (Visual Deductive Reasoning). None of these alone is sufficient, and several are vulnerable to contamination as their materials enter pretraining corpora, but jointly they sketch what process-level evaluation of visual reasoning should look like. In our assessment, adopting such causal probes as standard practice–rather than reporting only aggregate accuracy–is the single most consequential change the evaluation community could make.

These comparisons settle what the field has established; the questions they leave open define what it should do next.

## 6 Open Problems and Future Directions

**1. Closing the 3D, egocentric, and interactive gap.** OmniSpatial, SIRI-Bench, and iVISPAR localize failures in 3D perception, perspective taking, dynamic reasoning, and multi-step visual planning. SpatialPIN and SpatialRGPT show that 3D priors and depth/region grounding help, but the field still lacks a native, scalable, general-purpose 3D-aware VLM. Future systems may need embodied video data, egocentric experience, and spatial memory rather than static-image instruction tuning alone.

**2. Faithful reasoning, not merely formatted reasoning.** LLaVA-CoT, Reason-RFT, VL-Rethinker, GuardReasoner-VL, and BTL-UI all encourage explicit reasoning. The open question is whether the generated trace is causally grounded in the image or merely a plausible explanation. Perception-R1 and Vision-SR1 are important because they reward visual evidence, but they still rely on textualized visual descriptions and

judge/self-judge mechanisms. A stronger future direction is step-level visual verification: every intermediate visual claim should be checked against regions, depth, objects, or tools. CogCoM's evidential manipulations and DianJin-OCR-R1's tool-verified perception are early steps in this direction, but both are restricted to fixed operation or tool vocabularies.

**3. Shortcut-resistant and continuously refreshed evaluation.** ConMe and Do VLMs Understand Visual Language? show that high scores can arise from artifacts, background knowledge, or language priors, ChartMuseum shows that text-solvable questions can mask purely visual failures, and the Compositional Ability Gap study shows that per-skill scores can mask an inability to compose skills. Benchmarks should therefore include counterfactual diagrams, adversarial relation swaps, fresh generated hard cases, explicit visual-vs-textual splits, compositional skill probes, and interactive tasks where memorized priors are less useful. UniBench-style aggregation remains valuable, but the underlying benchmarks must be continually updated to avoid saturation and training-data contamination.

**4. Data quality, difficulty, and reward quality over raw scale.** UniBench suggests that more data is not always better; SoTA with Less shows that MCTS-selected difficult samples can be more valuable than large unfiltered data; Perception-R1 reports strong performance from only 1,442 high-quality RL samples; Vision-SR1 constructs a broader self-reward dataset but separates visual and language rewards. The shared lesson is that data and reward quality control are central to visual reasoning.

**5. Reward hacking and the "thinking over seeing" failure.** Recent RL-based VLMs can improve benchmark accuracy, but if the reward is based only on final answer matching, the model may learn language shortcuts or produce convincing but ungrounded rationales. Perception-R1 and Vision-SR1 explicitly identify this risk. Future RL training should separate perception rewards, reasoning rewards, format rewards, and efficiency rewards, while auditing for cases where the model obtains reward without actually using the image.

**6. Calibrated confidence and high-certainty hallucination.** The paper *Trust Me, I'm Wrong* (Simhi et al., 2025) introduces CHOKE–certain hallucinations overriding known evidence–where a model can know the correct answer but produce an incorrect answer with high certainty after a small perturbation. Although this work studies LLMs rather than VLMs, it is directly relevant to multimodal reasoning: a VLM may confidently hallucinate a visual relation, chart value, safety judgment, or GUI target even when it has enough information to answer correctly. Open evaluation should therefore measure not only accuracy, but also calibration, abstention, and whether confidence tracks visual evidence.

**7. Bridging perception and symbolic reasoning.** ChartVLM, Chart-based Reasoning, CogCoM, DianJin-OCR-R1, SpatialPIN, SpatialRGPT, Perception-R1, and Vision-SR1 all point toward a neuro-symbolic middle ground: perception should produce explicit evidence that a reasoner can operate on and verify. This is especially important for visual deduction and diagrams, where the relevant information is relational and symbolic rather than texture-based, and ChartMuseum's visual-reasoning subset shows the ceiling of purely symbolic intermediates: some questions require reasoning directly over visual properties that no table captures.

Taken together, these seven directions reveal several recurring bottlenecks and a growing set of complementary remedies.

## 7 Conclusion

Rather than restating the sections, we close with the claims that the surveyed literature, read through the pipeline of Figure 2, collectively supports.

*First, reasoning capability and visual grounding are inseparable.* A language model can inherit powerful symbolic reasoning from pre-training (Han et al., 2025), yet spatial, nonlocal, and binding failures show that this capability is inert when the grounded representation $Z$ of Eq. (2) lacks the relations the question requires (Sections 4.2–4.3).

*Second, intermediate representations have become the central design object.* Chart structure, depth-grounded regions, evidential manipulations, captions, tool outputs, latent thoughts, and explicit imagination are all

answers to the same question–what should stand between the image and the reasoner–and their comparative record (Table 3) increasingly decides system performance.

*Third, language priors both enable and mislead.* The same inherited knowledge that makes caption-mediated reasoning work lets models answer visual questions without using the image, inflating benchmark scores and complicating the attribution of progress.

*Fourth, explicit reasoning traces do not guarantee faithful visual reasoning*–a formatted trace can rest on a false visual claim–and latent traces are harder still to trust, as perturbation analyses show they can be causally disconnected from both input and answer. These findings suggest that verifiability may be more important than the surface format of a reasoning trace.

*Fifth, reinforcement learning is a powerful amplifier whose reward determines whether models learn to see, to reason, or merely to answer.* In the decomposition of Eq. (8), answer-level rewards can improve final-answer performance without necessarily improving visual perception or grounding.

*Sixth, evaluation still struggles to distinguish genuine visual reasoning from language priors and benchmark shortcuts,* but the elements of a causal methodology now exist: visual-versus-textual splits, image-free re-answering, perturbation probes, nonlocal tasks, and adversarial negatives.

*Finally*, these claims point in one direction. Spatial grounding, chart-to-table extraction, chains of manipulations, perception rewards, tool-verified OCR, and input-level visual structuring were developed by largely separate communities, yet each answers the same question: how should a model produce, and then verify, the intermediate visual evidence on which its reasoning depends? Our synthesis suggests that future VLMs may benefit from treating perception, grounding, intermediate representation, reasoning, and verification as a more integrated loop–with persistent visual memory and reliable imagination as its missing components–rather than as independent modules improved in isolation. The reviewed evidence suggests that scaling alone may not be sufficient; structure, supervision, and verification at the right stages of the pipeline may be necessary complements

## Acknowledgments

This survey was prepared for the System 2 course at Sharif University of Technology. We thank the course staff for the template and guidance. No external funding was used.

## References

Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023). Visual Instruction Tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xia, R., Zhang, B., Ye, H., Yan, X., Liu, Q., Zhou, H., Chen, Z., Dou, M., Shi, B., Yan, J., and Qiao, Y. (2024). ChartX and ChartVLM: A Versatile Benchmark and Foundation Model for Complicated Chart Reasoning. *arXiv preprint arXiv:2402.12185*.

Xu, G., Jin, P., Li, H., Song, Y., Sun, L., and Yuan, L. (2024). LLaVA-CoT: Let Vision Language Models Reason Step-by-Step. *arXiv preprint arXiv:2411.10440*.

Ma, C., Lu, K., Cheng, T.-Y., Trigoni, N., and Markham, A. (2024). SpatialPIN: Enhancing Spatial Reasoning Capabilities of Vision-Language Models through Prompting and Interacting 3D Priors. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Cheng, A.-C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., and Liu, S. (2024). SpatialRGPT: Grounded Spatial Reasoning in Vision-Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, X., Yang, Z., Feng, C., Lu, H., Li, L., Lin, C.-C., Lin, K., Huang, F., and Wang, L. (2025). SoTA with Less: MCTS-Guided Sample Selection for Data-Efficient Visual Reasoning Self-Improvement. *arXiv preprint arXiv:2504.07934*.

Jia, M., Duan, Z., Wu, S., Hu, H., Geng, Y., Yi, H., et al. (2025). OmniSpatial: Towards Comprehensive Spatial Reasoning Benchmark for Vision Language Models. *arXiv preprint arXiv:2506.03135*.

Zhang, Z., Chen, X., Liu, X., et al. (2025). SIRI-Bench: Challenging VLMs' Spatial Intelligence through Complex Reasoning Tasks. *arXiv preprint arXiv:2506.14512*.

Mayer, J., Nayak, S., Ballout, M., Bruni, E., et al. (2025). iVISPAR – An Interactive Visual-Spatial Reasoning Benchmark for VLMs. *arXiv preprint arXiv:2502.03214*.

Al-Tahan, H., Garrido, Q., Balestriero, R., Bouchacourt, D., Hazirbas, C., and Ibrahim, M. (2024). UniBench: Visual Reasoning Requires Rethinking Vision-Language Beyond Scaling. *arXiv preprint arXiv:2408.04810*.

Huang, I., Lin, W., Mirza, M. J., Hansen, J. A., Doveh, S., Butoi, V. I., Herzig, R., Arbelle, A., Kuehne, H., Darrell, T., Gan, C., Oliva, A., Feris, R., and Karlinsky, L. (2024). ConMe: Rethinking Evaluation of Compositional Reasoning for Modern VLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhang, Y., Bai, H., Zhang, R., Gu, J., Zhai, S., Susskind, J., and Jaitly, N. (2024). How Far Are We from Intelligent Visual Deductive Reasoning? *arXiv preprint arXiv:2403.04732*.

Hou, Y., Gile dereli, B., Tu, Y., and Sachan, M. (2024). Do Vision-Language Models Really Understand Visual Language? *arXiv preprint arXiv:2410.00193*.

Wang, H., Qu, C., Huang, Z., Chu, W., Lin, F., and Chen, W. (2025). VL-Rethinker: Incentivizing Self-Reflection of Vision-Language Models with Reinforcement Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tan, H., Ji, Y., Hao, X., Chen, X., Wang, P., Wang, Z., and Zhang, S. (2025). Reason-RFT: Reinforcement Fine-Tuning for Visual Reasoning of Vision Language Models. *arXiv preprint arXiv:2503.20752*.

Xiao, T., Xu, X., Huang, Z., Gao, H., Liu, Q., Liu, Q., and Chen, E. (2025). Perception-R1: Advancing Multimodal Reasoning Capabilities of MLLMs via Visual Perception Reward. *arXiv preprint arXiv:2506.07218*.

Li, Z., Yu, W., Huang, C., Liang, Z., Liu, R., Liu, F., Che, J., Yu, D., Boyd-Graber, J., Mi, H., and Yu, D. (2025). Vision-SR1: Self-Rewarding Vision-Language Model via Reasoning Decomposition and Multi-Reward Policy Optimization. *arXiv preprint arXiv:2508.19652*.

Liu, Y., Zhai, S., Du, M., Chen, Y., Cao, T., Gao, H., Wang, C., Li, X., Wang, K., Fang, J., Zhang, J., and Hooi, B. (2025). GuardReasoner-VL: Safeguarding VLMs via Reinforced Reasoning. *arXiv preprint arXiv:2505.11049*.

Zhang, S., Zhang, R., Fu, P., Wang, S., Yang, J., Du, X., Cui, S., Qin, B., Huang, Y., Luo, Z., and Luan, J. (2025). BTL-UI: Blink-Think-Link Reasoning Model for GUI Agent. *arXiv preprint arXiv:2509.15566*.

Chen, Q., Zhang, X., Guo, L., Chen, F., and Zhang, C. (2025). DianJin-OCR-R1: Enhancing OCR Capabilities via a Reasoning-and-Tool Interleaved Vision-Language Model. *arXiv preprint arXiv:2508.13238*.

Pang, Y., Yang, B., Cao, Y., Fan, R., Li, X., and He, C. (2025). MoCHA: Advanced Vision-Language Reasoning with MoE Connector and Hierarchical Group Attention. *arXiv preprint arXiv:2507.22805*.

Li, T., Zhang, J., Rao, Y., and Cheng, Y. (2025). Unveiling the Compositional Ability Gap in Vision-Language Reasoning Model. *arXiv preprint arXiv:2505.19406*.

Tang, L., Kim, G., Zhao, X., Lake, T., Ding, W., Yin, F., Singhal, P., Wadhwa, M., Liu, Z. L., Sprague, Z., Namuduri, R., Hu, B., Rodriguez, J. D., Peng, P., and Durrett, G. (2025). ChartMuseum: Testing Visual Reasoning Capabilities of Large Vision-Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Carbune, V., Mansoor, H., Liu, F., Aralikatte, R., Baechler, G., Chen, J., and Sharma, A. (2024). Chart-based Reasoning: Transferring Capabilities from LLMs to VLMs. In *Findings of the Association for Computational Linguistics: NAACL 2024*.

Qi, J., Ding, M., Wang, W., Bai, Y., Lv, Q., Hong, W., Xu, B., Hou, L., Li, J., Dong, Y., and Tang, J. (2024). CogCoM: A Visual Language Model with Chain-of-Manipulations Reasoning. *arXiv preprint arXiv:2402.04236*.

Weng, Z., Gomez, L., Webb, T. W., and Bashivan, P. (2025). Caption This, Reason That: VLMs Caught in the Middle. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Sun, G., Hua, H., Wang, J., Luo, J., Dianat, S., Rabbani, M., Rao, R., and Tao, Z. (2025). Latent Chain-of-Thought for Visual Reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Berman, S. and Deng, J. (2025). VLMs have Tunnel Vision: Evaluating Nonlocal Visual Reasoning in Leading VLMs. *arXiv preprint arXiv:2507.13361*.

Han, J., Tong, S., Fan, D., Ren, Y., Sinha, K., Torr, P., and Kokkinos, F. (2025). Learning to See Before Seeing: Demystifying LLM Visual Priors from Language Pre-training. *arXiv preprint arXiv:2509.26625*.

Li, Y., Chen, C., Li, Y., Zeng, F., Huang, K., Xu, J., and Sun, M. (2026). Imagination Helps Visual Reasoning, But Not Yet in Latent Space. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*.

Izadi, A., Banayeeanzade, M. A., Askari, F., Rahimiakbar, A., Vahedi, M. M., Hasani, H., and Soleymani Baghshah, M. (2025). Visual Structures Help Visual Reasoning: Addressing the Binding Problem in LVLMs. *arXiv preprint arXiv:2506.22146*.

Wang, Q., Shi, Y., Wang, Y., Zhang, Y., Wan, P., Gai, K., Ying, X., and Wang, Y. (2025). Monet: Reasoning in Latent Visual Space Beyond Images and Language. *arXiv preprint arXiv:2511.21395*.

Li, Y., Liu, Z., Li, Z., Zhang, X., Xu, Z., Chen, X., et al. (2025). Perception, Reason, Think, and Plan: A Survey on Large Multimodal Reasoning Models. *arXiv preprint arXiv:2505.04921*.

Su, Z., Xia, P., Guo, H., et al. (2025). Thinking with Images for Multimodal Reasoning: Foundations, Methods, and Future Frontiers. *arXiv preprint arXiv:2506.23918*.

Zhou, C., Wang, M., Ma, Y., et al. (2025). From Perception to Cognition: A Survey of Vision-Language Interactive Reasoning in Multimodal Large Language Models. *arXiv preprint arXiv:2509.25373*.

Wang, Y., Wu, S., Zhang, Y., Yan, S., Liu, Z., Luo, J., and Fei, H. (2025). Multimodal Chain-of-Thought Reasoning: A Comprehensive Survey. *arXiv preprint arXiv:2503.12605*.

Simhi, A., Itzhak, I., Barez, F., Stanovsky, G., and Belinkov, Y. (2025). Trust Me, I'm Wrong: LLMs Hallucinate with Certainty Despite Knowing the Answer. *arXiv preprint arXiv:2502.12964*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., and others (2021). Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., and Gao, J. (2022a). Grounded Language-Image Pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. (2022). CoCa: Contrastive Captioners are Image-Text Foundation Models. In *Transactions on Machine Learning Research (TMLR)*.

Li, J., Li, D., Savarese, S., and Hoi, S. C. H. (2023). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., and others (2022). Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

## A Additional Details

**Scope and selection.** The survey began from a core visual reasoning set and was expanded in several rounds with recent works on reinforced reasoning, perception and self rewards, safety reasoning, GUI agents, visual language, architecture design, chart-based reasoning transfer and evaluation, evidential and tool-interleaved reasoning, compositional skill analysis, latent and explicit imagination, nonlocal visual evaluation, visual priors from language pre-training, and input-level visual structuring. The open-problems section additionally discusses Simhi et al. (2025) because high-certainty hallucination is directly relevant to faithful visual reasoning even though the paper focuses on LLMs.

**Reading guide by reasoning sub-type.** For spatial reasoning, start with OmniSpatial, SIRI-Bench, and iVISPAR, then read SpatialPIN (Ma et al., 2024) and SpatialRGPT (Cheng et al., 2024) as remedies. For multi-step reasoning, pair LLaVA-CoT (Xu et al., 2024), CogCoM (Qi et al., 2024), Caption This, Reason That (Weng et al., 2025), Latent-CoT (Sun et al., 2025), Monet (Wang et al., 2025d), CapImagine (Li et al., 2026), VL-Rethinker (Wang et al., 2025b), and Reason-RFT (Tan et al., 2025). For chart and diagram reasoning, read ChartX/ChartVLM (Xia et al., 2024), Chart-based Reasoning (Carbune et al., 2024), and ChartMuseum. For reward design, read Perception-R1 (Xiao et al., 2025) and Vision-SR1 (Li et

al., 2025a). For compositional reasoning and binding, read ConMe, the Compositional Ability Gap study, and VISER (Izadi et al., 2025). For evaluation methodology, read UniBench, Tunnel Vision, and Do VLMs Understand Visual Language?. For architecture-side improvements and foundations, read MoCHA (Pang et al., 2025) and Learning to See Before Seeing (Han et al., 2025). For applied reasoning beyond ordinary VQA, read GuardReasoner-VL (Liu et al., 2025), DianJin-OCR-R1 (Chen et al., 2025), and BTL-UI (Zhang et al., 2025b).