Show, Don't Just Tell: A Comprehensive Survey of Visual Prompting for Multimodal Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

Multimodal large language models (MLLMs) equip pre-trained large-language models (LLMs) with visual capabilities. While textual prompting in LLMs has been widely studied, visual prompting has emerged for more fine-grained and free-form visual instructions. This paper presents the first comprehensive survey on visual prompting methods in MLLMs, focusing on visual prompting, prompt generation, compositional reasoning, and prompt learning. We categorize existing visual prompts and discuss generative methods for automatic prompt annotations on the images. We also examine visual prompts that enable better alignment between visual encoders and backbone LLMs, concerning MLLM's visual grounding, referring, and compositional reasoning abilities. This paper examines visual prompting methods developed in MLLMs and provides a vision of the future of these methods.

1 Introduction

Multimodal large language models (MLLMs)[1] (Li et al., 2023b; Liu et al., 2024b) augment pre-trained large language models (LLMs) with visual capabilities, enabling visual understanding and reasoning on complex multimodal tasks (Zhou et al., 2024b; Jia et al., 2024). However, conventional *textual* prompting describes and references visual elements only indirectly (Lin et al., 2025; Wu et al., 2024d), falling short of accurate visual grounding and detailed referring; this gap is a documented source of visual hallucinations (Bai et al., 2024; Huang et al., 2024b) and language-prior bias (Wu et al., 2024b; Qu et al., 2024).

Visual prompting – structured visual instructions overlaid on, segmenting, or perturbing the image – has emerged as a complementary paradigm (Zhang et al., 2024a; Wu et al., 2024g) that operates directly in pixel or embedding space and supports instance-level, pixel-accurate references. Unlike textual prompt templates, visual prompts are inherently heterogeneous (boxes, markers, masks, soft tokens) and typically must be *generated* per image rather than templated, raising new questions on how prompts should be represented (Section 4), generated (Section 5), aligned with the underlying MLLM (Sections 8 and 9), and used to drive learning (Sections 6 and 7). Pre-trained MLLMs may also be *misaligned* with visual prompts when prompt-style data is under-represented during training (Yan et al., 2024a; Lin et al., 2024), motivating dedicated pre-training, supervised fine-tuning, in-context learning, and preference-optimization strategies.

This survey. We present, to the best of our knowledge, the first comprehensive survey of visual prompting in MLLMs. We organize the field along the four-axis taxonomy of Figure 3 – (i) representation granularity, (ii) generation mechanism, (iii) adaptation paradigm, (iv) inference and reasoning – which we formalize via a unified visual-prompt operator family π (Definition 2.2) and an adaptation-paradigm signature (Section 3.5). Figure 1 situates these axes in a single end-to-end pipeline.

Summary of Main Contributions.

- We formalize visual prompting via a unified *visual-prompt operator family* $\pi : \mathcal{I} \times \Psi \to \mathcal{I}'$ (Definition 2.2) and an *alignment functional family* (Definition 3.2) that together subsume bounding-box, marker, pixel/mask, and soft prompts and the alignment metrics used across the literature.

[1] We use MLLM synonymously with Large Multimodal Model (LMM) and Large Vision-Language Model (LVLM).

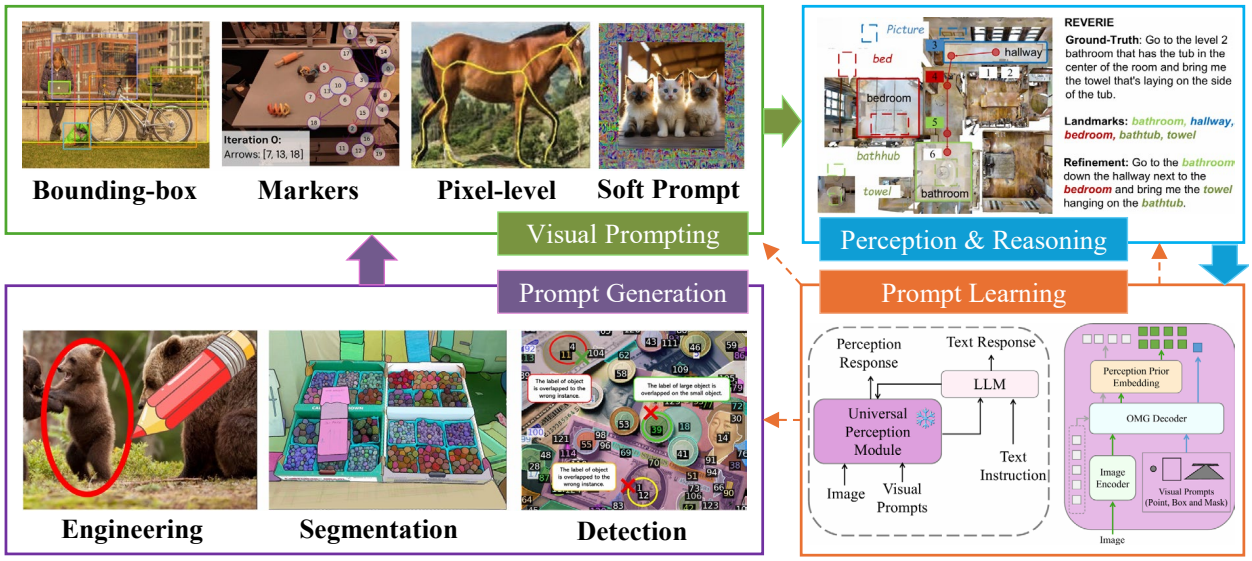

Figure 1: Taxonomy flow chart of visual prompting techniques. We illustrate in order of four stages of visual prompting including prompt generation, visual prompting, perception and reasoning, and prompt learning, where the solid arrows show the direction of each component's information flow. We explain in detail various visual prompt generation techniques (Section 5), and how these generated prompts are used to prompt MLLMs (Section 4). Then we discuss the advanced perception and reasoning abilities achieved through visual prompting (Section 8 and Section 9). Finally, model pre-training, fine-tuning, instruction tuning, and in-context learning further update previous model components, which are illustrated by the dashed arrows (Section 7 and Section 6).

- We organize the field along an explicit *four-axis taxonomy* (Figure 3) (*i.e.*, representation, generation, adaptation paradigm, capability) and show that categorizations are projections onto these axes.
- We unify training-free, in-context, supervised, and reinforcement-learning approaches via an *adaptation-paradigm signature* ($\Theta_{\text{trainable}}$, $\mathcal{D}_{\text{adapt}}$, Π_{search}) (Section 3.5).
- For each axis we surface the **trade-offs, limitations, and open challenges** of existing methods, accompanied by representative analyses and pointers to evaluation evidence.
- We provide a comprehensive **empirical comparison** of cross-task root-cause patterns (Table 3) and a curated benchmark/dataset overview (Table 2).
- We discuss **frontier applications** – jailbreaking and safety, hallucination, debiasing, controllable generation, and 3D understanding (Section 11) – and map each onto the four axes.

Scope. We focus on works that use *visual prompts* in MLLMs as explicit, visually grounded instructions complementing textual inputs. We decompose the surveyed works along the four axes above; works are sorted by their dominant axis combination. We exclude vision-only prompting (without a language backbone), text-only LLM prompting, and controllable image generation that does not involve an MLLM (Cao et al., 2024).

Relation to prior surveys. Existing surveys on LLM prompting cover textual prompt design (Gu et al., 2023; Schulhoff et al., 2024) and in-context demonstrations (Xu et al., 2024b; Li, 2023) but omit pixel-level instructions and multimodal interactions. Visual prompting has also been surveyed in pure computer vision (Lei et al., 2024b; Zhang et al., 2024c), but those treatments are limited to vision-only backbones and do not cover MLLM-specific issues such as cross-modal alignment, instruction tuning with visual prompts, or preference optimization. A SAM-specific survey (Zhang et al., 2023a) discusses SAM's role in MLLMs but does not span the broader visual-prompt design space we address here.

2 Problem Formulation

This section formalizes visual prompting in multimodal large language models (MLLMs). We first fix the input and parameter spaces (Section 2.1); next we introduce a unified *visual-prompt operator family* that subsumes the four common prompt types (Section 2.2); finally, we cast visual prompting as a constrained optimization problem and state its combined objective (Section 2.3). The survey-level abstractions that organize the rest of the paper – taxonomy, generator family, alignment functional, reasoning recurrence, and adaptation-paradigm signature – are gathered into a single companion framework in Section 3, and the notation introduced here is summarized in Table 1.

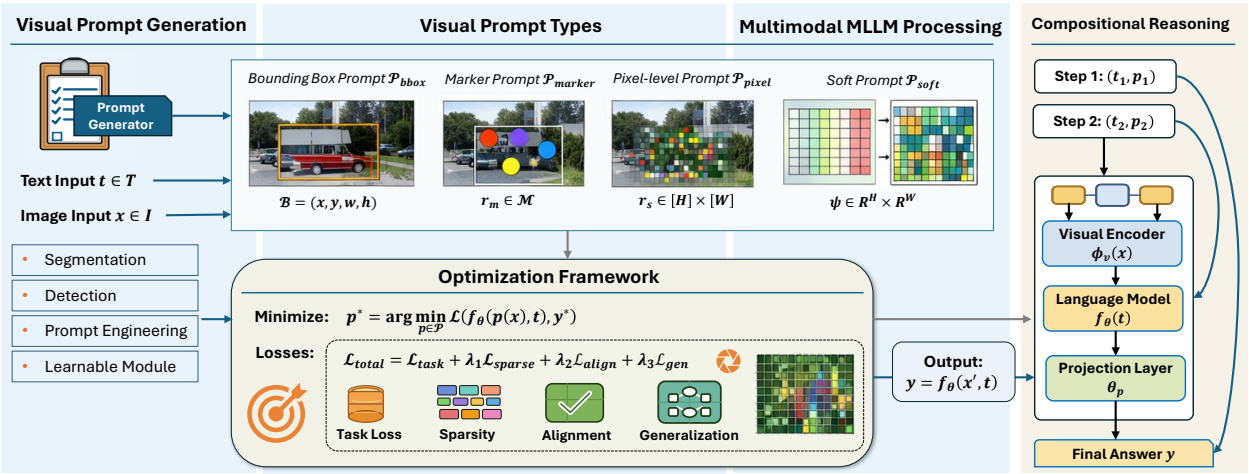

Figure 2: Overview of the visual prompting pipeline. A textual instruction $\boldsymbol{t}$ and image $\boldsymbol{x}$ feed a generator g that produces prompt parameters $\theta_\pi \in \Psi$; these parameters instantiate the visual-prompt operator $\pi(\,\cdot\,;\theta_\pi)$, yielding the augmented input $\boldsymbol{x}' = \pi(\boldsymbol{x};\theta_\pi)$ consumed by an MLLM $f_{\boldsymbol{\theta}}$ to produce $\boldsymbol{y}$. Solid arrows denote the forward inference path; dashed arrows denote training-time updates under one of four adaptation paradigms (training-free, ICL, supervised fine-tuning, reinforcement learning), discussed in Sections 6 and 7.

2.1 Preliminaries

Let $\mathcal{I} \subseteq \mathbb{R}^{H\times W\times C}$ denote the space of visual inputs (images or video frames), where H, W, and C are the height, width, and channel dimensions. Let $\mathcal{T}$ denote the space of textual inputs, i.e., token sequences over a vocabulary $\mathcal{V}$. A multimodal input is a pair $(\boldsymbol{x}, \boldsymbol{t}) \in \mathcal{I} \times \mathcal{T}$, where $\boldsymbol{x}$ is a visual input and $\boldsymbol{t}$ is a textual instruction or query. An MLLM is parameterized by $\boldsymbol{\theta} = (\boldsymbol{\theta}_v, \boldsymbol{\theta}_l, \boldsymbol{\theta}_p)$, where $\boldsymbol{\theta}_v$ are the parameters of the visual encoder $\phi_v(\,\cdot\,;\boldsymbol{\theta}_v) : \mathcal{I} \to \mathbb{R}^{N_v\times d_v}$, $\boldsymbol{\theta}_l$ are the parameters of the language-model backbone ϕ_l, and $\boldsymbol{\theta}_p$ are the parameters of the projection layer $\rho_{\boldsymbol{\theta}_p}$ that aligns visual and textual representations. The composite MLLM function is

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}', \boldsymbol{t}) \;=\; \phi_l\big(\rho_{\boldsymbol{\theta}_p}(\phi_v(\boldsymbol{x}';\boldsymbol{\theta}_v)),\, \boldsymbol{t};\, \boldsymbol{\theta}_l\big), \tag{1}$$

where the semicolon separates function inputs from the parameter block being applied. Equation (1) is the image-space case; when $\boldsymbol{x}'$ is already a visual-token sequence, as in soft prompting, the outer ϕ_v call is omitted and $\rho_{\boldsymbol{\theta}_p}$ is applied directly.

2.2 Visual Prompts and the Operator Family

Definition 2.1 (Visual Prompt) *A* visual prompt $\boldsymbol{p}$ *is a structured visual instruction that augments a base image* $\boldsymbol{x}$ *to guide an MLLM's attention, perception, or reasoning. Formally,* $\boldsymbol{p}$ *is a function* $\boldsymbol{p} : \mathcal{I} \to \mathcal{I}'$ *that transforms an input image* $\boldsymbol{x}$ *into an augmented visual input or representation* $\boldsymbol{x}' = \boldsymbol{p}(\boldsymbol{x})$*, where* $\mathcal{I}'$ *is the space of prompt-augmented visual inputs. The set of such prompt functions is denoted by* $\mathcal{P}$*.*

Visual prompts decompose into four fundamental types – bounding-box, marker, pixel-level, and soft – distinguished by the subspace of the prompt-parameter space in which their parameters live. Each type is detailed in Section 4; we collect them under a single typed operator family.

Definition 2.2 (Visual-Prompt Operator Family) *A* visual-prompt operator family *is a typed map* $\pi : \mathcal{I} \times \Psi \to \mathcal{I}'$*, where the prompt-parameter space partitions into representation-specific subspaces* $\Psi = \Psi_{\mathrm{B}} \sqcup \Psi_{\mathrm{M}} \sqcup \Psi_{\mathrm{S}} \sqcup \Psi_{\mathrm{soft}}$*. Each prompt type instantiates* π *by restricting its second argument:*

$$\begin{aligned}
\pi(\boldsymbol{x};\, \theta_\pi = \mathcal{B} \in \Psi_{\mathrm{B}}) &= \boldsymbol{p}_{\mathrm{bbox}}(\boldsymbol{x};\, \mathcal{B}),\\
\pi(\boldsymbol{x};\, \theta_\pi = \mathcal{M} \in \Psi_{\mathrm{M}}) &= \boldsymbol{p}_{\mathrm{marker}}(\boldsymbol{x};\, \mathcal{M}),\\
\pi(\boldsymbol{x};\, \theta_\pi = (\mathcal{S}, \boldsymbol{f}) \in \Psi_{\mathrm{S}}) &= \boldsymbol{p}_{\mathrm{pixel}}(\boldsymbol{x};\, \mathcal{S},\, \boldsymbol{f}),\\
\pi(\boldsymbol{x};\, \theta_\pi = \boldsymbol{\psi} \in \Psi_{\mathrm{soft}}) &= \boldsymbol{p}_{\mathrm{soft}}(\boldsymbol{x};\, \boldsymbol{\psi}).
\end{aligned}$$

For soft prompts injected at intermediate layers – for example the FFN weights as in MemVP (Jie et al., 2024) – π *is replaced by the embedding-injection variant* $\pi_{\mathrm{inject}} : \mathbb{R}^{N\times d} \times \Psi_{\mathrm{soft}} \to \mathbb{R}^{N\times d}$*, which acts on a hidden state rather than the pixel-space image. The semicolon* ; *in* $\pi(\boldsymbol{x};\theta_\pi)$ *separates the operand* $\boldsymbol{x}$ *from the parameter* θ_π *on which* π *is conditioned.*

2.3 The Visual Prompting Optimization Problem

With π and $f_{\boldsymbol{\theta}}$ in hand, the visual-prompting problem is the search for prompt parameters that drive the MLLM toward a desired output.

Problem 2.1 (Visual Prompting in MLLMs) *Given a multimodal input* $(\boldsymbol{x}, \boldsymbol{t}) \in \mathcal{I} \times \mathcal{T}$ *and a task specification with desired output* $\boldsymbol{y}^*$*, the visual-prompting problem seeks to:*

1. ***Generate*** *prompt parameters* $\theta_\pi^* \in \Psi$ *– equivalently, a prompt* $\boldsymbol{p}^* = \pi(\,\cdot\,;\theta_\pi^*)$ *– that enhance the MLLM's ability to understand the visual input;*
2. ***Integrate*** *the prompt with the base image to produce* $\boldsymbol{x}' = \pi(\boldsymbol{x};\theta_\pi^*) \in \mathcal{I}'$*; and*
3. ***Process*** *the augmented input through* $f_{\boldsymbol{\theta}}$ *to produce an output* $\boldsymbol{y} = f_{\boldsymbol{\theta}}(\boldsymbol{x}', \boldsymbol{t})$ *satisfying the task requirements.*

Casting Problem 2.1 as an optimization over the prompt-parameter space yields

$$\theta_\pi^* \;=\; \underset{\theta_\pi \in \Psi}{\arg\min}\; \mathcal{L}\big(f_{\boldsymbol{\theta}}(\pi(\boldsymbol{x};\theta_\pi),\, \boldsymbol{t}),\; \boldsymbol{y}^*\big), \qquad \boldsymbol{p}^* \;=\; \pi(\,\cdot\,;\theta_\pi^*), \tag{2}$$

where $\mathcal{L}$ is a task-specific loss and Ψ is the prompt-parameter space of Definition 2.2.

In practice, the objective in (2) is regularized so that it balances task performance, prompt sparsity, visual-text alignment, and generalization:

$$\mathcal{L}_{\mathrm{total}}(\theta_\pi) \;=\; \mathcal{L}_{\mathrm{task}}(\theta_\pi, \boldsymbol{x}, \boldsymbol{t}, \boldsymbol{y}^*) \;+\; \lambda_1\, \mathcal{L}_{\mathrm{sparse}}(\theta_\pi) \;+\; \lambda_2\, \mathcal{L}_{\mathrm{align}}(\pi(\boldsymbol{x};\theta_\pi),\, \boldsymbol{t}) \;+\; \lambda_3\, \mathcal{L}_{\mathrm{gen}}(\theta_\pi), \tag{3}$$

with trade-off coefficients $\lambda_1, \lambda_2, \lambda_3 \geq 0$ and $\mathcal{L}_{\mathrm{align}}$ instantiated from the alignment functional of Definition 3.2. Detailed instantiations of each loss component appear in the respective downstream sections.

3 A Unifying Framework for Visual Prompting

Section 2 fixes the input/parameter spaces, defines the visual-prompt operator family π, and casts visual prompting as a constrained optimization problem. The remaining survey-level abstractions used as reading guides for the rest of the paper are gathered here. Each abstraction is introduced once – the taxonomy that organizes the surveyed works (Section 3.1), the generator family that produces prompt parameters (Section 3.2), the alignment functional that grades visual-text faithfulness (Section 3.3), the recurrence that lifts single-step prompting to compositional reasoning (Section 3.4), and the paradigm signature that unifies all four adaptation procedures (Section 3.5) – and reused thereafter by reference. Table 1 summarizes the mathematical notation introduced across Sections 2 and 3.

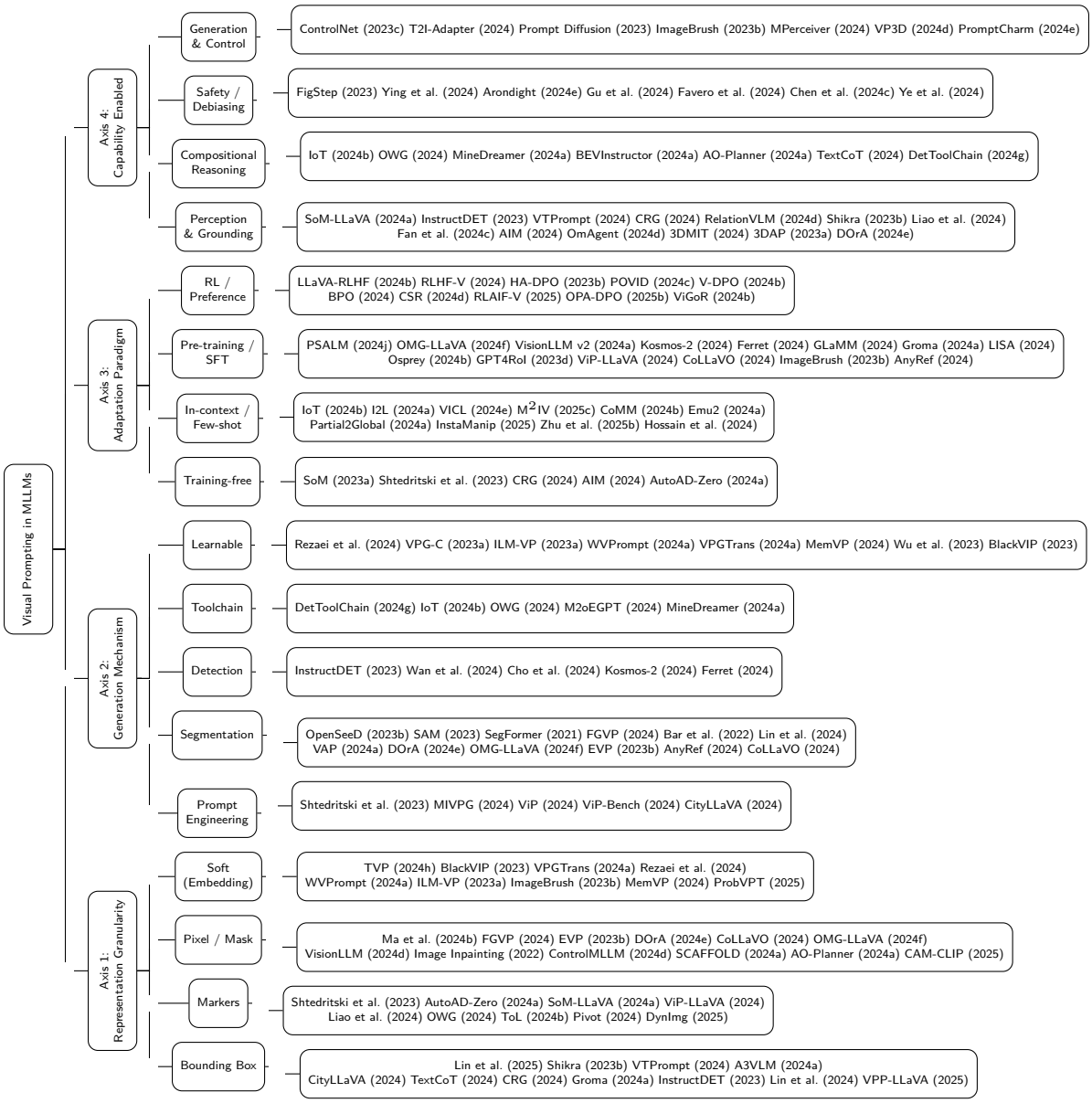

Figure 3: Multi-axis taxonomy of visual prompting in MLLMs. The taxonomy spans four orthogonal axes – representation granularity, generation mechanism, adaptation paradigm, and capability enabled. A given work typically populates more than one axis (e.g., a soft prompt produced by a learnable generator and adapted via SFT to enable grounding), so the figure is best read as a multi-axis classification rather than a strict partition; cross-axis edges are omitted for clarity.

3.1 Taxonomy Overview

The remainder of the survey is organized along four orthogonal axes (Figure 3); we use this overview both to motivate the section order and to read Table 1 top-to-bottom.

- **Axis 1 (Representation Granularity, Section 4)** – the subspace of Ψ in which a prompt parameter lives: region (bbox or marker), pixel/mask, or embedding (soft).
- **Axis 2 (Generation Mechanism, Section 5)** – the kind of generator g used to produce θ_π: manual, model-based, learned, or toolchain composition.

- **Axis 3 (Adaptation Paradigm, Sections 6 and 7)** – the training procedure that obtains θ_π and possibly $\boldsymbol{\theta}$: training-free, in-context, supervised, or reinforcement learning. Formally captured by the paradigm signature in Section 3.5.
- **Axis 4 (Inference and Reasoning, Sections 8, 9 and 11)** – the downstream capability the prompted model exhibits: perception and grounding, compositional reasoning, safety and debiasing, or controllable generation.

A given work typically populates a combination of axes rather than a single leaf, so the taxonomy is best read as a multi-axis classification, not a strict partition.

3.2 Visual Prompt Generation

Visual-prompt generation addresses how the parameters $\theta_\pi \in \Psi$ are obtained. We posit a generator family that the surveyed works specialize in different ways.

Definition 3.1 (Visual-Prompt Generator) *A* visual-prompt generator *is a map $g : \mathcal{T} \times \mathcal{I} \to \Psi$ that sends a textual instruction $\boldsymbol{t}$ and base image $\boldsymbol{x}$ to prompt parameters $\theta_\pi = g(\boldsymbol{t}, \boldsymbol{x})$. We distinguish two modes:*

- *a* fixed *generator g_{fix} – rule- or tool-based, with no learnable parameters (e.g., a frozen segmenter s, detector d, or handcrafted heuristic h_{rule}); and*
- *a* learned *generator $g_{\text{learn}}(\,\cdot\,;\boldsymbol{\varphi})$ parameterized by $\boldsymbol{\varphi}$, trained jointly with or separately from $\boldsymbol{\theta}$.*

The five generation mechanisms surveyed in Section 5 – prompt engineering, segmentation-based, detection-based, toolchain composition, and learnable prompts – are instances of either g_{fix} or g_{learn}.

3.3 Visual-Text Alignment

A fundamental requirement of visual prompting is that the prompt and the textual instruction reference the same semantic content. We capture this requirement as a granularity-parameterized functional.

Definition 3.2 (Alignment Functional Family) *Let $\gamma \in \{\text{pixel}, \text{region}, \text{token}, \text{embedding}, \text{temporal}\}$ index a granularity level. The* alignment functional family $\text{Align}_\gamma(\pi(\boldsymbol{x};\theta_\pi),\, \boldsymbol{t})$ *assigns a real-valued score measuring how well the prompt-augmented visual input localizes the content referenced by $\boldsymbol{t}$ at granularity γ. Two common specializations are*

$$\text{Align}_{\text{region}}(\boldsymbol{x}',\, \boldsymbol{t}) \;=\; \text{IoU}\big(\mathcal{R}(\boldsymbol{x}'),\, \mathcal{R}_{\boldsymbol{x}}(\boldsymbol{t})\big), \qquad \text{Align}_{\text{embedding}}(\boldsymbol{x}',\, \boldsymbol{t}) \;=\; \cos\big(\phi_v(\boldsymbol{x}'),\, \phi_l^{\text{text}}(\boldsymbol{t})\big), \tag{4}$$

where $\mathcal{R}(\boldsymbol{x}')$ extracts the prompted or predicted image regions, $\mathcal{R}_{\boldsymbol{x}}(\boldsymbol{t})$ denotes the regions in image $\boldsymbol{x}$ semantically referenced by $\boldsymbol{t}$, and ϕ_l^{text} denotes the text-side embedding from the language backbone. Pixel-, token-, and temporal-level instances follow analogously. Concrete realizations and per-task choices of γ are surveyed in Section 8.

3.4 Compositional Reasoning

Visual prompts enable compositional reasoning by providing step-wise visual grounding. Given a sequence of reasoning steps $\{(\boldsymbol{t}_t, \theta_{\pi,t})\}_{t=1}^{T}$ with running history $\boldsymbol{h}_t$, the model produces

$$\boldsymbol{y}_t \;=\; f_{\boldsymbol{\theta}}\big(\pi(\boldsymbol{x}_t;\theta_{\pi,t}),\, \boldsymbol{t}_t,\, \boldsymbol{h}_{t-1}\big), \qquad \boldsymbol{h}_t \;=\; [\boldsymbol{h}_{t-1};\, (\pi(\boldsymbol{x}_t;\theta_{\pi,t}),\, \boldsymbol{t}_t,\, \boldsymbol{y}_t)]. \tag{5}$$

Here $f_{\boldsymbol{\theta}}$ is overloaded to accept a running multimodal history. This recurrence subsumes single-step prompting (set $T = 1$) and visual chain-of-thought (Section 9); ICL with K demonstrations is the special case $T = K + 1$ in which the first K steps carry $(\boldsymbol{p}_i(\boldsymbol{x}_i),\, \boldsymbol{t}_i,\, \boldsymbol{y}_i)$ as fixed context (Section 6).

Table 1: **Mathematical notation.** Symbols are grouped by spaces and sets, vectors and scalars, functions and operators, sizes and coordinates, and losses and hyperparameters. Italic φ denotes a function argument; bold $\boldsymbol{\varphi}$ denotes learned parameters. Prompt-parameter subspaces $\Psi_{\mathrm{B}}, \Psi_{\mathrm{M}}, \Psi_{\mathrm{S}}, \Psi_{\mathrm{soft}}$ partition the unified parameter space Ψ along Axis 1 of the taxonomy (Figure 3).

Symbol	Description	Symbol	Description
Spaces, sets, and datasets			
$\mathcal{I}$	Space of visual inputs (images/videos)	$\mathcal{T}$	Space of textual inputs
$\mathcal{V}$	Vocabulary of tokens	$\mathcal{P}$	Space of visual prompts
$\mathcal{I}'$	Space of prompt-augmented visual inputs	$\mathcal{B}$	Set of bounding-box regions
$\mathcal{M}$	Set of visual markers	$\mathcal{S}$	Set of pixel coordinates
$\mathcal{R}(\boldsymbol{x}')$, $\mathcal{R}_{\boldsymbol{x}}(\boldsymbol{t})$	Prompted / text-referred image regions	Ψ	Prompt-parameter space (union over types)
Ψ_{B}	Bbox prompt parameters ($\subseteq \mathbb{R}^{4K}$)	Ψ_{M}	Marker prompt parameters
Ψ_{S}	Pixel/mask prompt parameters	Ψ_{soft}	Soft (embedding-space) prompt parameters
$\mathcal{D}_{\mathrm{train}}$, $\mathcal{D}_{\mathrm{test}}$	Train / test data distributions	$\mathcal{D}_{\mathrm{demo}}$, $\mathcal{D}_{\mathrm{support}}$	ICL demonstrations / few-shot support set
$\mathcal{D}_{\mathrm{adapt}}$	Adaptation data in paradigm signature	$\mathcal{D}_{\mathrm{train}}^{\mathrm{ground}}$, $\mathcal{D}_{\mathrm{train}}^{\mathrm{instr}}$, $\mathcal{D}_{\mathrm{train}}^{\mathrm{pref}}$	Pre-training / SFT / preference data
$\Theta_{\mathrm{trainable}}$	Trainable parameter blocks	Π_{search}	Adaptation/search procedure
Vectors and scalars			
$\boldsymbol{x} \in \mathcal{I}$	Visual input (image/video frame)	$\boldsymbol{t} \in \mathcal{T}$	Textual instruction/query
$\boldsymbol{x}' \in \mathcal{I}'$	Prompt-augmented visual input	$\boldsymbol{x}_t$	Visual state at reasoning step t
$\boldsymbol{y}, \boldsymbol{y}^*$	Model output, ground-truth output	$\boldsymbol{p} \in \mathcal{P}$	Visual prompt (instance)
$\boldsymbol{p}^*$	Optimal visual prompt	$\boldsymbol{p}_t$	Visual prompt at reasoning step t
$\theta_\pi \in \Psi$	Prompt parameters	$\theta_{\pi,t}$	Prompt parameters at step t
$\boldsymbol{\theta} = (\boldsymbol{\theta}_v, \boldsymbol{\theta}_l, \boldsymbol{\theta}_p)$	MLLM parameters	$\boldsymbol{\psi}$	Soft / learnable prompt parameters
$\boldsymbol{\varphi}$	Prompt-generator parameters	$\boldsymbol{h}_t$	Reasoning history at step t
Functions and operators			
$\phi_v(\cdot;\boldsymbol{\theta}_v)$	Visual encoder	$\phi_l(\cdot;\boldsymbol{\theta}_l)$	Language-model backbone
$\rho_{\boldsymbol{\theta}_p}$	Vision-language projection	$f_{\boldsymbol{\theta}}$	MLLM function
$g : \mathcal{T} \times \mathcal{I} \to \Psi$	Visual-prompt generator	g_{fix}	Fixed (rule/tool-based) generator
$g_{\mathrm{learn}}(\,\cdot\,;\boldsymbol{\varphi})$	Learnable generator	$\pi(\,\cdot\,;\theta_\pi)$	Unified visual-prompt operator family
π_{inject}	Embedding-injection variant	$\boldsymbol{p}_{\mathrm{bbox}}$	Bounding-box prompt
$\boldsymbol{p}_{\mathrm{marker}}$	Marker prompt	$\boldsymbol{p}_{\mathrm{pixel}}$	Pixel-level prompt
$\boldsymbol{p}_{\mathrm{soft}}$	Soft visual prompt	$\mathrm{Align}_\gamma\left(\boldsymbol{x}',\, \boldsymbol{t}\right)$	Visual-text alignment functional
$s,\ d,\ h_{\mathrm{rule}}$	Segmenter, detector, rule heuristic	ϕ_l^{text}	Text-side embedding map
Sizes, coordinates, and pixel-space objects			
$H,\ W,\ C$	Image height, width, channels	$[H] \times [W]$	Pixel coordinate grid
$N_v,\ d_v$	Number and dimension of visual tokens	$N_p,\ d_p$	Prompt-token count and native dimension
$K,\ M,\ T$	num. boxes, markers, reasoning steps	$b_i = (x_i, y_i, w_i, h_i)$	Bounding-box coordinates
$\boldsymbol{r}_m \subseteq [H] \times [W]$	Marker region	$\boldsymbol{m}_{\mathcal{B}}$	Binary mask for bounding boxes
$\boldsymbol{P} \in \mathbb{R}^{N_v \times N_p}$	Soft-prompt placement matrix	$\boldsymbol{f} : \mathbb{R}^C \to \mathbb{R}^C$	Pixel transformation function
$\boldsymbol{g}_m$	Graphical primitive for marker m	$\boldsymbol{x}_{\mathrm{overlay}}$	Visual annotation overlay
$2^{[H]\times[W]}$	Power set of pixel coordinates	$2^{\mathbb{R}^4}$	Power set of bbox coordinates
$\odot$	Hadamard (element-wise) product	$[\,\cdot\,;\,\cdot\,]$	Multimodal sequence concatenation
Losses and hyperparameters			
$\mathcal{L}$	Loss function	$\mathcal{L}_{\mathrm{task}}$	Task-specific loss
$\mathcal{L}_{\mathrm{sparse}}$	Sparsity regularizer	$\mathcal{L}_{\mathrm{align}}$	Alignment loss
$\mathcal{L}_{\mathrm{gen}}$	Generalization regularizer	$\lambda_1, \lambda_2, \lambda_3$	Loss-trade-off coefficients
$\mathbb{E}$	Expectation	γ	Alignment granularity index
$\delta,\ \epsilon$	Performance improvement, error bound	β	RL KL coefficient

3.5 Adaptation-Paradigm Signature

To unify the training procedures surveyed in Section 6 (ICL and few-shot) and Section 7 (pre-training, supervised fine-tuning, reinforcement learning), we associate each method with a typed *paradigm signature*

$$(\Theta_{\mathrm{trainable}} \subseteq \{\boldsymbol{\theta},\, \boldsymbol{\varphi},\, \theta_\pi\},\, \mathcal{D}_{\mathrm{adapt}},\, \Pi_{\mathrm{search}}), \tag{6}$$

where $\Theta_{\mathrm{trainable}}$ is the subset of parameter blocks updated during adaptation, $\mathcal{D}_{\mathrm{adapt}}$ is the data used (e.g., $\mathcal{D}_{\mathrm{demo}}$ for ICL, $\mathcal{D}_{\mathrm{support}}$ for few-shot, $\mathcal{D}_{\mathrm{train}}$ for pre-training and SFT, an online policy distribution for RL), and Π_{search} is the search procedure (analytical, gradient-based, preference optimization, and so on). Training-free methods correspond to $\Theta_{\mathrm{trainable}} = \emptyset$; in-context learning leaves $\boldsymbol{\theta}$ frozen and absorbs $\mathcal{D}_{\mathrm{demo}}$ into the prompt; SFT updates $\boldsymbol{\theta}$ (or a subset thereof) on $\mathcal{D}_{\mathrm{train}}$; RL replaces direct loss minimization with a reward-shaped objective. Sections 6 and 7 populate this signature for representative works.

4 Types of Visual Prompt

This section instantiates the visual-prompt operator family π (Definition 2.2, Axis 1 of the taxonomy) along the four fundamental *types of visual prompt*, bounding box (Section 4.1), marker (Section 4.2), pixel/mask (Section 4.3), and soft (Section 4.4), distinguished by the subspace of Ψ that their parameters occupy. Each subsection gives a formal definition, surveys representative methods organized by their key challenges, and ends with a limitation analysis. Visual prompts of every type were originally introduced as auxiliary cues (Wu et al., 2024g) that complement textual prompts at finer spatial granularity than language alone can express.

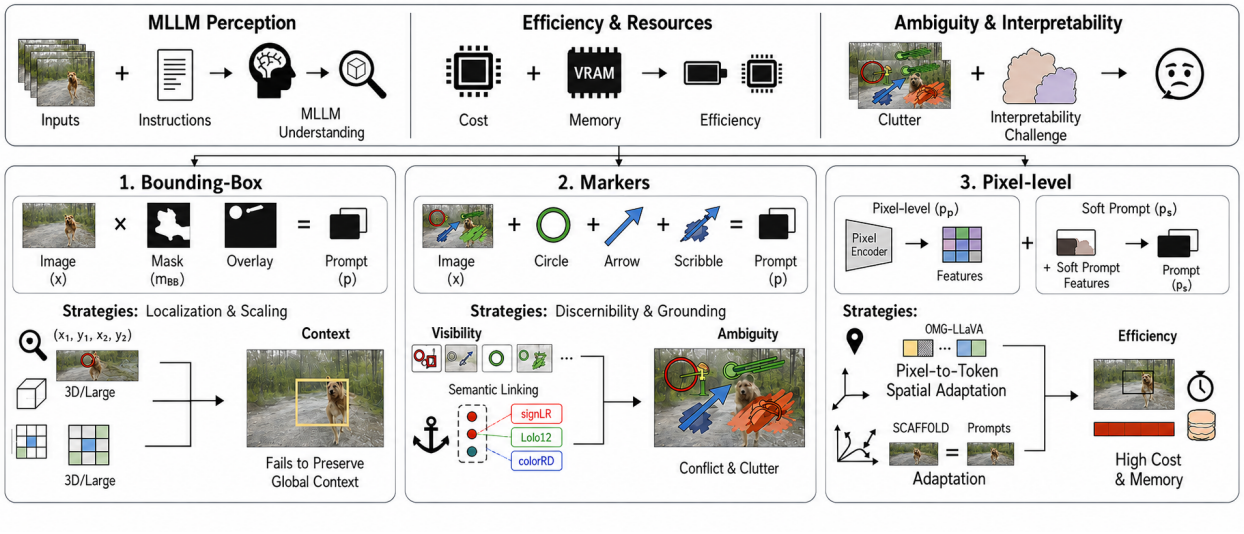

Figure 4: The four fundamental types of visual prompt surveyed in Section 4: bounding box (rectangular region highlights), marker (discrete overlays such as numbers, arrows, or scribbles), pixel/mask (fine-grained region edits), and soft (learned embedding-space perturbations). Each panel corresponds to a subspace of the prompt-parameter space Ψ (Definition 2.2).

4.1 Bounding-Box

Definition 4.1 (Bounding-Box Prompt) *A* bounding-box prompt $\boldsymbol{p}_{\text{bbox}}$ *is parameterized by a set of rectangular regions* $\mathcal{B} = \{b_i = (x_i, y_i, w_i, h_i)\}_{i=1}^{K} \subseteq \mathbb{R}^4$, *where* (x_i, y_i) *denotes the top-left corner and* (w_i, h_i) *denotes width and height. The prompt acts as*

$$\boldsymbol{p}_{\text{bbox}}(\boldsymbol{x}; \mathcal{B}) \;=\; \boldsymbol{x} \odot (\mathbf{1} - \boldsymbol{m}_{\mathcal{B}}) \;+\; \boldsymbol{x}_{\text{overlay}} \odot \boldsymbol{m}_{\mathcal{B}}, \tag{7}$$

where $\boldsymbol{m}_{\mathcal{B}} \in \{0,1\}^{H \times W \times C}$ *is a broadcast binary mask indicating pixels within bounding boxes and* $\boldsymbol{x}_{\text{overlay}}$ *contains the visual annotations (e.g., colored boxes, highlights). The parameter subspace is* $\Psi_{\text{B}} \subseteq \mathbb{R}^{4K}$.

Bounding boxes are used to demarcate objects or regions within an image, enabling MLLMs to extract visual features (Lin et al., 2025). These features help the model understand the image content and correlate it with the corresponding text, thereby enhancing fine-grained and grounded image understanding.

Challenge 4.1.1 (Localization) To enable the localization of user instructions, several approaches leverage bounding-box encoding for fine-grained understanding. Groma (Ma et al., 2024a) and InstructDET (Dang et al., 2023) encode user-specified regions (i.e., bounding boxes) into visual tokens, allowing MLLMs to focus on specific areas of interest. In contrast, Lin et al. (2024) integrates contextual embeddings from external knowledge directly into bounding boxes, effectively using these as visual prompts to enhance fine-grained cognitive abilities. Shikra (Chen et al., 2023b) and VTPrompt (Jiang et al., 2024) take an innovative approach by quantizing bounding boxes to numerically represent key objects, modeling both their input and output positions. VPP-LLaVA (Tang et al., 2025) addresses the inherent coordinate-alignment challenge

by introducing explicit global spatial references, implemented as a learnable global visual-position prompt alongside local position-aware queries to enhance spatial grounding and alignment.

While Groma and InstructDET excel in general bounding-box localization, frameworks like Shikra and VTPrompt emphasize structured representation, enabling more precise localization of key objects. The trend is also evolving toward unified architectures capable of processing bounding boxes alongside other prompt formats: the VP-MLLM framework (Lin et al., 2025) employs a unified visual-prompt encoder to handle diverse forms, facilitating fine-grained understanding within a single model.

Challenge 4.1.2 (Scalability) Scaling bounding-box methods to broader applications involves new rescaling techniques and dynamic adjustments. A3VLM (Huang et al., 2024a) employs 3D bounding boxes to identify actionable parts of images, extending the dimensionality of object annotations for richer understanding. CityLLaVA (Duan et al., 2024) scales bounding boxes by accommodating larger areas, and TextCoT (Luan et al., 2024) addresses dimensional limitations by extending the shorter side of each box to ensure complete region coverage. CRG (Wan et al., 2024) takes a different approach by masking specific regions with black pixels, reducing visual priors and allowing the model to adjust predictions without additional training. Compared to 2D-based methods like CityLLaVA and TextCoT, A3VLM's 3D annotations provide deeper spatial understanding, while CRG offers a flexible way to improve robustness against spurious correlations.

Limitation 4.1 (Contextual Awareness) While bounding-box techniques have advanced fine-grained visual understanding, they often struggle to capture broader contextual relationships between objects. Methods like Shikra (Chen et al., 2023b) and VTPrompt (Jiang et al., 2024) excel at localizing objects but rely heavily on predefined regions, potentially overlooking critical interactions between objects outside these boundaries. Scalability-focused methods such as A3VLM (Huang et al., 2024a) and TextCoT (Luan et al., 2024) extend bounding boxes effectively but may fail to dynamically adapt to scenes with complex, interdependent object relationships. Emerging benchmarks like VP-Bench (Xu et al., 2025b) systematically evaluate bounding boxes and other visual prompts across a wide range of attributes and downstream tasks, underscoring the community's push toward more systematic assessment of these context-awareness limitations. Developing approaches that seamlessly integrate global context and dynamic object relationships alongside localized bounding-box annotations remains an essential research challenge.

4.2 Markers

Definition 4.2 (Marker Prompt) *A* marker prompt $\boldsymbol{p}_{\text{marker}}$ *overlays discrete visual markers – circles, arrows, scribbles, or numerical tags – on specific image regions. Formally,*

$$\boldsymbol{p}_{\text{marker}}(\boldsymbol{x};\, \mathcal{M}) \;=\; \boldsymbol{x} \;+\; \sum_{m \in \mathcal{M}} \boldsymbol{g}_m(\boldsymbol{x}, \boldsymbol{r}_m), \tag{8}$$

where $\mathcal{M} = \{m_i\}_{i=1}^{M}$ *is a set of markers, each* m_i *associated with a region* $\boldsymbol{r}_m \subseteq [H] \times [W]$ *and a graphical primitive* $\boldsymbol{g}_m$ *– e.g., a circle or an arrow – that renders the marker. The parameter subspace is* Ψ_{M}.

Similar to bounding boxes, visual markers are specific elements within visual data – images or videos – used to highlight, identify, or draw attention to particular features or regions. They are often employed to indicate particular parts of an image that are relevant to the task. The concept of markers also extends from static images to dynamic video understanding, where they can serve as *temporal prompts* that guide attention across frames.

Challenge 4.2.1 (Discernibility) Visual markers improve the discernibility of prompts in MLLMs by highlighting specific regions. Shtedritski et al. (2023) demonstrate that models trained on web-scale data respond to markers like red circles without requiring cropped images. Set-of-Mark (SoM) prompting (Yang et al., 2023a) overlays markers on images to ground model outputs in specific regions. ViP-LLaVA (Cai et al., 2024) extends this by using diverse cues like scribbles and arrows for enhanced recognition. SoM-LLaVA (Yan et al., 2024a) improves SoM by associating markers with detailed descriptions of all tagged items.

Challenge 4.2.2 (Grounding) Recent work on the visual-grounding challenge introduced by visual prompting employs visual markers to better link semantic concepts within image frames. AutoAD-Zero (Xie et al.,

2024a) introduces a training-free approach using circles and color-coded markers to identify characters in frames. Liao et al. (2024) enhance SoM by integrating feedback via textual or visual marks to improve grounding. ToL (Fan et al., 2024b) and OWG (Tziafas & Kasaei, 2024) assign unique IDs to segments, while Pivot (Nasiriany et al., 2024) uses projected 3D markers for spatial references. DynImg (Bao et al., 2025) treats non-key video frames as *temporal visual prompts* to highlight spatial regions of fast-moving objects, extending the role of markers from spatial to spatio-temporal domains for enhanced video understanding.

Limitation 4.2 (Ambiguity) One major limitation of visual markers lies in their potential ambiguity when interpreted by models, especially in complex or cluttered scenes. For example, while SoM (Yang et al., 2023a) employs markers such as circles, arrows, or scribbles to guide model attention, these visual cues can sometimes overlap or conflict with each other, leading to misinterpretation. Further refinements are necessary to enhance their interpretability and precision.

4.3 Pixel-level

Definition 4.3 (Pixel-level Prompt) *A* pixel-level prompt $\boldsymbol{p}_{\text{pixel}}$ *operates at the finest spatial granularity, modifying individual pixels or pixel neighborhoods (segmentation masks, inpainting regions, coordinate-based annotations):*

$$\big[\boldsymbol{p}_{\text{pixel}}(\boldsymbol{x};\, \mathcal{S},\, \boldsymbol{f})\big]_{i,j} \;=\; \begin{cases} \boldsymbol{f}(\boldsymbol{x}_{i,j}) & \text{if } (i,j) \in \mathcal{S}, \\ \boldsymbol{x}_{i,j} & \text{otherwise,} \end{cases} \tag{9}$$

where $\mathcal{S} \subseteq [H] \times [W]$ *is a set of pixel coordinates and* $\boldsymbol{f} : \mathbb{R}^C \to \mathbb{R}^C$ *is a pixel transformation (e.g., a color overlay or a semantic mask encoding). The parameter subspace is* Ψ_{S}.

Previous approaches relied on coarse markers such as colorful boxes or circles, which introduced ambiguity in accurately highlighting objects. To address this, pixel-level prompts (Ma et al., 2024b) operate on individual pixels in images or videos, enhancing the semantic localization capability of MLLMs.

Challenge 4.3.1 (Localization and Spatial Reasoning) Pixel-level prompts enable precise object localization and spatial reasoning by embedding semantic and geometric information directly into fine-grained visual representations. FGVP (Yang et al., 2024), EVP (Liu et al., 2023b), DOrA (Wu et al., 2024e), and CoLLaVO (Lee et al., 2024) exploit pixel-level representations to localize objects accurately, while OMG-LLaVA (Zhang et al., 2024f) and VisionLLM (Wang et al., 2024d) tokenize images into pixel-centric visual tokens, aligning dense visual tasks with language instructions. Other methods further refine or structure these representations: Image Inpainting (Bar et al., 2022) decodes visual tokens back into pixels to improve fine-grained reconstruction, ControlMLLM (Wu et al., 2024d) models semantic relationships between pixels and prompts to support reasoning, and CAM-CLIP (Ji et al., 2025) extends pixel-level prompting to industrial manufacturing and safety monitoring through mask-text matching and learnable mask-prompting mechanisms. Beyond localization, pixel-level prompts also support spatial reasoning by grounding visual understanding in coordinate-aware representations. SCAFFOLD (Lei et al., 2024a) and AO-Planner (Chen et al., 2024a) convert images into coordinate systems, enabling models to better capture spatial relationships, affordances, and object interactions within a scene.

Limitation 4.3 (Efficiency) Pixel-level prompts improve semantic localization and spatial reasoning but at the cost of increased computational and memory demands. Methods like OMG-LLaVA (Zhang et al., 2024f) and VisionLLM (Wang et al., 2024d) tokenize images at a fine-grained pixel level, which is resource-intensive, especially for high-resolution or real-time tasks. Optimizing these approaches to maintain accuracy while improving scalability and efficiency is a crucial research direction.

4.4 Soft Visual Prompts

Definition 4.4 (Soft Visual Prompt) *A* soft visual prompt $\boldsymbol{p}_{\text{soft}}$ *is a learnable perturbation acting in the embedding space rather than the pixel space. We write* $\boldsymbol{\psi} \in \Psi_{\text{soft}} \subseteq \mathbb{R}^{N_p \times d_v}$ *for the prompt-token matrix after any required channel projection:*

$$\boldsymbol{p}_{\text{soft}}(\boldsymbol{x};\, \boldsymbol{\psi}) \;=\; \phi_v(\boldsymbol{x}) \;+\; \boldsymbol{P}\,\boldsymbol{\psi}, \tag{10}$$

where $\boldsymbol{P} \in \mathbb{R}^{N_v \times N_p}$ is a placement matrix that integrates N_p prompt tokens into the N_v-token visual sequence, and the addition occurs in the embedding space after visual encoding. The embedding-injection variant π_{inject} (Definition 2.2) generalizes this to deeper hidden states; for example, MemVP (Jie et al., 2024) injects $\boldsymbol{\psi}$ directly into FFN weights rather than the visual-token stream.

Unlike the previous three types, soft prompts trade interpretability for adaptability: they cannot be inspected as pixel-space overlays, but they can be optimized end-to-end via gradient descent on $\boldsymbol{\psi}$.

Challenge 4.4.1 (Universal and Image-Specific Perturbations) Soft prompts can act as either universal vectors or image-conditional perturbations. TVP (Zhang et al., 2024h), BlackVIP (Oh et al., 2023), and VPGTrans (Zhang et al., 2024a) apply universal or image-specific prompts to pixel or embedding spaces, improving task performance. Learned Prompt (Rezaei et al., 2024), WVPrompt (Ren et al., 2024a), and ILM-VP (Chen et al., 2023a) introduce perturbations to modify inputs, while ImageBrush (Yang et al., 2023b) extracts tokenized features to enhance semantic understanding. Beyond static or image-specific perturbations, Probabilistic Visual Prompt Tuning (Sun et al., 2025) generates prompts with uncertainty by sampling from a Gaussian distribution, more accurately capturing the data distribution of tasks.

Limitation 4.4 (Interpretability and Transferability) Soft prompts are opaque relative to their pixel-space counterparts, and their transferability across backbones and downstream tasks remains an open challenge. Because $\boldsymbol{\psi}$ lives in an embedding subspace tied to a particular visual encoder, prompts learned for one MLLM rarely transfer verbatim to another, motivating the prompt-transfer line of work surveyed in Section 5.5.

5 Visual Prompt Generation

This section surveys methods that instantiate the generator g of Definition 3.1 (Axis 2 of the taxonomy). Unlike textual prompt templates, visual prompts are position-aware and instance-specific, so they are typically *generated* per image rather than templated. We organize generation mechanisms by the source of the prompt parameters: human-engineered (Section 5.1), produced by a frozen segmenter s (Section 5.2) or detector d (Section 5.3), composed through a toolchain (Section 5.4), or learned end-to-end (Section 5.5). The first four correspond to instances of g_{fix}; the last is the canonical $g_{\text{learn}}(\,\cdot\,;\boldsymbol{\varphi})$.

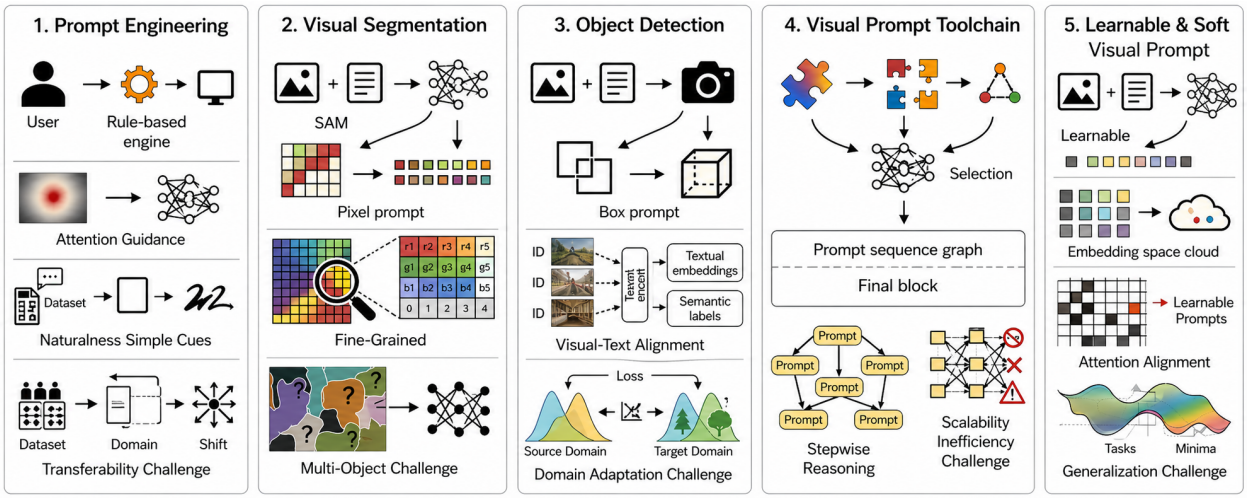

Figure 5: Pipeline for visual-prompt generation. A textual instruction $\boldsymbol{t}$ and image $\boldsymbol{x}$ feed one of five generator branches (prompt engineering, segmentation, object detection, toolchain composition, or a learnable generator) producing prompt parameters $\theta_\pi \in \Psi$ that drive the visual-prompt operator $\pi(\boldsymbol{x};\theta_\pi)$.

The five subsections below populate Definition 3.1 by specializing the generator g to a concrete construction.

5.1 Prompt Engineering

Definition 5.1 (Prompt-Engineering Generator) *A* prompt-engineering generator $g_{\text{eng}} : \mathcal{T} \times \mathcal{I} \to \Psi$ *applies a rule-based heuristic* h_{rule} *to extract spatial or semantic information from the textual instruction:*

$$g_{\text{eng}}(\boldsymbol{t},\, \boldsymbol{x}) \;=\; h_{\text{rule}}(\boldsymbol{t}, \boldsymbol{x}), \tag{11}$$

where h_{rule} *is a handcrafted procedure with no learnable parameters. Thus* g_{eng} *is an instance of* g_{fix}.

Prompt engineering focuses on designing human-crafted visual prompts to express intent or highlight specific areas within visual data effectively. Such prompts provide efficient ways to convey attention or instructions directly to MLLMs.

Challenge 5.1.1 (Attention Guidance) Simple techniques like drawing red circles effectively guide model attention (Shtedritski et al., 2023), but enriching prompts with detailed visual evidence remains a challenge. MIVPG (Zhong et al., 2024) addresses this by leveraging instance correlations within images or patches, enhancing the model's focus and comprehension of the highlighted areas.

Challenge 5.1.2 (Naturalness) Recent advancements have emphasized making prompt engineering more intuitive and interactive. ViP-Bench (Cai et al., 2024) evaluates MLLMs on these natural cues, setting benchmarks for region-specific comprehension tasks. Frameworks like *Draw-and-Understand* (Lin et al., 2025) unify the understanding of diverse visual prompts (points, boxes, scribbles) through training on large-scale, multi-domain instruction data. Emerging benchmarks such as the MARS2 2025 Challenge (Xu et al., 2025c) focus on evaluating multimodal reasoning in complex, real-world scenarios, steering prompt engineering toward more practical and robust applications.

Limitation 5.1 (Transferability) While prompt engineering techniques have improved interactivity and precision, challenges persist in generalizing across domains and handling ambiguous visual prompts. For instance, domain-specific frameworks like CityLLaVA (Duan et al., 2024) tailor engineered visual prompts for urban scenarios, but their transferability to other contexts is limited. A promising direction is to integrate prompt engineering with *prompt learning* paradigms: methods like ModalPrompt (Zeng et al., 2025) learn task-specific soft prompts for continual instruction tuning, mitigating catastrophic forgetting and enhancing cross-task generalization.

5.2 Visual Segmentation

Definition 5.2 (Segmentation-Based Generator) *A* segmentation-based generator $g_{\text{seg}} : \mathcal{T} \times \mathcal{I} \to \Psi_{\text{S}}$ *instantiates* g_{fix} *via a frozen segmenter* $s : \mathcal{I} \times \mathcal{T} \to 2^{[H]\times[W]}$ *that maps an image-instruction pair to a set of task-relevant pixel coordinates:*

$$g_{\text{seg}}(\boldsymbol{t},\, \boldsymbol{x}) \;=\; \big(s(\boldsymbol{x}, \boldsymbol{t}),\, \boldsymbol{f}\big), \qquad \pi\big(\boldsymbol{x};\, g_{\text{seg}}(\boldsymbol{t}, \boldsymbol{x})\big) \;=\; \boldsymbol{p}_{\text{pixel}}\big(\boldsymbol{x};\, s(\boldsymbol{x}, \boldsymbol{t}),\, \boldsymbol{f}\big), \tag{12}$$

where the pixel transformation $\boldsymbol{f}$ *specifies how the masked region is rendered (e.g., color overlay, blur, semantic label).*

Segmentation-based generators produce pixel-level prompts that delineate task-relevant regions. The Segment Anything Model (SAM) (Kirillov et al., 2023) is the workhorse segmenter throughout this subsection; we cite it once here and refer to it by name thereafter. Alternative segmenters include OpenSeeD (Zhang et al., 2023b), SegFormer (Xie et al., 2021), and the subsequent scaling of SAM to over 270K concepts in SAM 3 (Carion et al., 2025).

Challenge 5.2.1 (Fine-Grained Information) Segmentation methods refine the granularity of visual prompts to enhance contextual understanding. Yang et al. (2024) leverage pixel-level annotations derived from image inpainting (Bar et al., 2022) for fine-grained prompting. Lin et al. (2024) embed segmentation knowledge into spatial embedding maps through instruction tuning, improving MLLMs' comprehension of detailed visual contexts. Extending analysis to the temporal dimension, recent work on video reasoning segmentation emphasizes the critical role of motion and temporal tokens for dynamic scenes (Gong et al., 2025).

Challenge 5.2.2 (Grounding) Segmentation supports grounding by linking visual elements to task-specific requirements. VAP (Chen et al., 2024a) integrates SAM into navigation tasks, using affordance-based prompts to guide decision-making. DOrA (Wu et al., 2024e) extends this by incorporating 3D spatial and contextual information, enabling segmentation for 3D visual grounding tasks. A promising direction unifies segmentation and reasoning by using reinforcement learning to refine region proposals based on chain-of-thought reasoning, thereby improving semantic grounding (Zhu et al., 2025a).

Challenge 5.2.3 (Multilevel Integration) Integrating segmentation at multiple levels enhances MLLMs' visual perception and reasoning. OMG-LLaVA (Zhang et al., 2024f) combines coarse-to-fine visual prompts for comprehensive scene understanding. Liu et al. (2023b) improve processing of low-level structures, while He et al. (2024) utilize segmentation prompts during fine-tuning to enhance fine-grained perception. This integration is advanced by models designed for multi-round, region-grounded visual reasoning (Sun et al., 2026) and by scalable architectures that unify segmentation across diverse tasks and datasets through query-fusion mechanisms (Wang et al., 2025a).

Limitation 5.2 (Multi-Object) While segmentation techniques have achieved notable success, handling complex multi-object interactions remains challenging. CoLLaVO (Lee et al., 2024) introduces a crayon prompting technique, augmenting segmentation with panoptic maps and inpainting-based color overlays to improve multi-object discrimination. Recent efforts to tackle occlusion and appearance variation include detection models with multi-time inquiry mechanisms for better object disambiguation (Nan et al., 2025) and view-consistent learning strategies that improve robustness in open-world settings (Zhang et al., 2025a). Ensuring the robustness of these methods across diverse visual domains requires further investigation.

5.3 Object Detection

Definition 5.3 (Detection-Based Generator) *A* detection-based generator $g_{\text{det}} : \mathcal{T} \times \mathcal{I} \to \Psi_{\text{B}}$ *instantiates* g_{fix} *via an object detector* $d : \mathcal{I} \times \mathcal{T} \to 2^{\mathbb{R}^4}$ *that returns a set of bounding-box coordinates:*

$$g_{\text{det}}(\boldsymbol{t}, \boldsymbol{x}) \;=\; d(\boldsymbol{x}, \boldsymbol{t}) \;\subseteq\; \mathbb{R}^4, \qquad \pi\big(\boldsymbol{x};\, g_{\text{det}}(\boldsymbol{t}, \boldsymbol{x})\big) \;=\; \boldsymbol{p}_{\text{bbox}}\big(\boldsymbol{x};\, d(\boldsymbol{x}, \boldsymbol{t})\big). \tag{13}$$

Detection-based generators produce bounding-box prompts via a frozen detector. Representative detectors include SoM (Yang et al., 2023a), Fast RCNN (Girshick, 2015), and Omni3D (Brazil et al., 2023).

Challenge 5.3.1 (Visual-Text Alignment) Aligning visual objects with textual descriptions remains a significant challenge. SoM-LLaVA (Yan et al., 2024a) uses numeric tags to map visual objects to textual descriptions, enabling accurate listing and description of objects and strengthening the model's ability to reason about and follow visual instructions.

Challenge 5.3.2 (Context Awareness) Extending object detection to 3D environments introduces the challenge of spatial reasoning and understanding object interactions in three dimensions. Cho et al. (2024) enhance MLLM spatial awareness by integrating 3D object interactions, enabling better comprehension of object relationships within complex scenes. Fundamental techniques for understanding and editing the intrinsic properties of 3D scenes, such as albedo and shading, provide a lower-level foundation for such contextual reasoning (Moazeni et al., 2024).

Limitation 5.3 (Domain Adaptation) Although object detection methods have improved MLLM capabilities, adapting these techniques across diverse domains and handling cluttered, multi-object environments remains a limitation. Insights from offline learning paradigms, which focus on leveraging diverse demonstrations to improve generalization, may offer valuable directions for addressing this adaptation challenge (Yue et al., 2024). Robust solutions for transitions between 2D and 3D contexts require further development.

5.4 Visual Prompt Toolchain

Definition 5.4 (Toolchain Generator) *A* toolchain generator *composes* n *component generators into a multi-step pipeline:*

$$g_{\text{tool}}(\boldsymbol{t}, \boldsymbol{x}) \;=\; \big(g_n \circ g_{n-1} \circ \cdots \circ g_1\big)(\boldsymbol{t}, \boldsymbol{x}), \qquad g_i \in \{g_{\text{eng}},\, g_{\text{seg}},\, g_{\text{det}},\, g_{\text{fix}},\, g_{\text{learn}}\}, \tag{14}$$

where each g_i may consume both the original $(\boldsymbol{t}, \boldsymbol{x})$ and the partial output of $g_{1:i-1}$ to produce intermediate prompt parameters, and the composition is realized as the reasoning-step recurrence of Equation (5) *with $\theta_{\pi,t}$ provided by g_t. The chain is an instance of g_{fix} when none of the g_i are learned, and of g_{learn} otherwise.*

Toolchain generators compose multiple of the above mechanisms – prompt engineering, segmenters, detectors, even other MLLMs – into a multi-step pipeline that produces prompt parameters θ_π. These pipelines enable MLLMs to invoke specific tools for individual reasoning sub-tasks, facilitating structured multimodal reasoning (Wu et al., 2024g).

Challenge 5.4.1 (Stepwise Reasoning) A key challenge in toolchain development is guiding MLLMs through sequential reasoning processes. Zhou et al. (2024b) address this with the Image-of-Thought framework, which automates the selection of visual-information extraction methods for each reasoning step. Implementing these steps as visual prompts, the model follows a defined reasoning path that enables systematic multimodal reasoning. An extension, Render-of-Thought (RoT) (Wang et al., 2026), compresses and visualizes entire textual reasoning chains into compact images, making the latent reasoning process more explicit and efficient.

Challenge 5.4.2 (Task Generalization) Developing generalizable and transferable visual prompts is another challenge. Sheng et al. (2024) propose a unified in-context learning framework that integrates diverse contextual visual prompts into a unified representational space, enhancing adaptability across tasks. Similarly, MineDreamer (Zhou et al., 2024a) introduces a versatile visual-prompt generation method for imaginary visual scenes, aligning visual prompts with decision-making intentions to guide next-step goals.

Limitation 5.4 (Scalability) While visual-prompt toolchains improve structured reasoning and task adaptability, scalability remains a limitation. Complex tasks such as open-world grasping (Tziafas & Kasaei, 2024) require integrating multiple prompting methods – segmentation and object grounding among them – which can introduce inefficiencies and inconsistencies in large-scale applications.

5.5 Learnable and Soft Visual Prompt

Definition 5.5 (Learnable Generator) *A* learnable generator $g_{\text{learn}} : \mathcal{T} \times \mathcal{I} \to \Psi_{\text{soft}}$ *uses a neural network parameterized by $\boldsymbol{\varphi}$ to produce soft-prompt parameters:*

$$g_{\text{learn}}(\boldsymbol{t},\, \boldsymbol{x};\, \boldsymbol{\varphi}) \;=\; \boldsymbol{\psi}(\boldsymbol{t},\, \boldsymbol{x};\, \boldsymbol{\varphi}), \qquad \pi\big(\boldsymbol{x};\, g_{\text{learn}}(\boldsymbol{t}, \boldsymbol{x}; \boldsymbol{\varphi})\big) \;=\; \boldsymbol{p}_{\text{soft}}\big(\boldsymbol{x};\, \boldsymbol{\psi}(\boldsymbol{t}, \boldsymbol{x}; \boldsymbol{\varphi})\big). \tag{15}$$

Special cases set $\boldsymbol{\psi}$ to be input-independent (a universal prompt) or replace $\boldsymbol{p}_{\text{soft}}$ with the injection variant π_{inject}; for example, MemVP (Jie et al., 2024) injects $\boldsymbol{\psi}$ into the FFN weights.

Learnable and soft visual prompts instantiate $g_{\text{learn}}(\,\cdot\,; \boldsymbol{\varphi})$ – the second mode of Definition 3.1 – where prompt parameters are produced by a neural network with its own learnable parameters $\boldsymbol{\varphi}$ rather than by a frozen tool. These techniques play a pivotal role in multimodal instruction tuning, where visual instructions are tailored for downstream tasks.

Challenge 5.5.1 (Attention Alignment) Learning visual prompts to guide attention mechanisms within Vision Transformers (ViT) is a key focus area. Rezaei et al. (2024) explores methods to train visual prompts that effectively steer attention in ViT models. Li et al. (2023a) fine-tune MLLMs to follow zero-shot demonstrative instructions using learnable visual prompts, enhancing their adaptability in unseen scenarios.

Challenge 5.5.2 (Label Mapping) Improving the mapping of visual inputs to corresponding labels is another critical challenge. Chen et al. (2023a) focuses on better aligning visual prompts with task-specific labels through learned prompting techniques. For domain-specific problems, Ren et al. (2024a) develops a learnable visual prompt for image watermarking, enabling the identification of ownership within visual data.

Challenge 5.5.3 (Transferability) Transferability of visual prompts across MLLMs and downstream tasks remains an active area of research. VPGTrans (Zhang et al., 2024a) introduces a transferable visual-prompt generator that adapts pre-trained source MLLMs to target MLLMs with minimal training data and computational cost. Memory-space Visual Prompt (Jie et al., 2024) (π_{inject}) injects learnable prompts into the key and value layers of vision-transformer architectures, facilitating efficient vision-language fine-tuning.

Challenge 5.5.4 (Few-Shot Learning) Soft visual prompts are also employed to enhance multimodal information composition in few-shot scenarios. Wu et al. (2023) use soft visual tokens as compositional operations, enabling models to integrate multimodal inputs effectively with limited examples. Black-Box Visual Prompting (Oh et al., 2023) focuses on robust transfer learning, allowing visual prompts to adapt models to new tasks and domains without requiring access to underlying model parameters.

Limitation 5.5 (Generalization) Challenges remain in scaling learnable visual prompts to broader tasks and ensuring generalization across diverse domains. Reliance on task-specific tuning and potential inefficiency in highly dynamic environments highlight the need for further research into scalable and universally applicable prompting techniques.

6 In-context and Few-shot Learning

This section surveys the two training-light points of Axis 3 (adaptation paradigm): *in-context learning* (ICL), where $\Theta_{\text{trainable}} = \emptyset$ and $\mathcal{D}_{\text{demo}}$ is absorbed into the prompt (Section 6.1), and *few-shot learning*, where a small support set $\mathcal{D}_{\text{support}}$ is used to update a low-dimensional $\boldsymbol{\psi}$ while $\boldsymbol{\theta}$ remains frozen (Section 6.2). Both are instances of the paradigm signature $(\Theta_{\text{trainable}}, \mathcal{D}_{\text{adapt}}, \Pi_{\text{search}})$ introduced in Section 3.5; ICL sets $\Theta_{\text{trainable}} = \emptyset$ and $\mathcal{D}_{\text{adapt}} = \mathcal{D}_{\text{demo}}$, while few-shot sets $\Theta_{\text{trainable}} = \{\boldsymbol{\psi}\}$ and $\mathcal{D}_{\text{adapt}} = \mathcal{D}_{\text{support}}$.

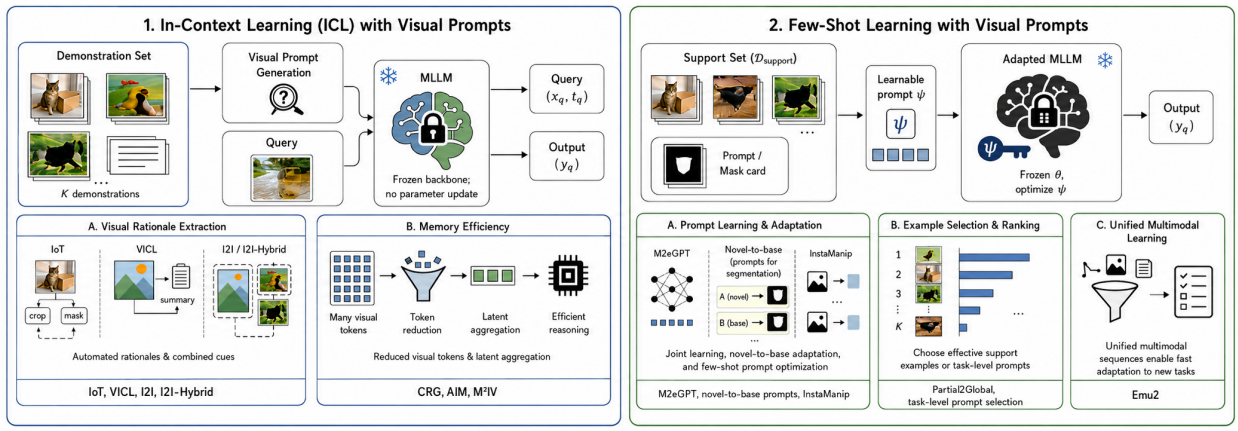

Figure 6: In-context and few-shot learning with visual prompts. Left: ICL inserts K demonstrations $\mathcal{D}_{\text{demo}}$ as a visual-textual context preceding the query; no parameters are updated. Right: few-shot learning optimizes a small set of soft-prompts $\boldsymbol{\psi}$ on a support set $\mathcal{D}_{\text{support}}$ while the MLLM backbone $\boldsymbol{\theta}$ remains frozen.

6.1 In-context Learning

In-context learning with visual prompts enables MLLMs to learn task patterns from visual demonstrations without any parameter updates. Visual prompts in demonstration examples highlight key regions, making it easier for models to identify and replicate task-specific patterns in the query.

Definition 6.1 (In-context Learning with Visual Prompts) *Given a query $(\boldsymbol{x}_q, \boldsymbol{t}_q)$ and a demonstration set $\mathcal{D}_{\text{demo}} = \{(\boldsymbol{x}_i, \boldsymbol{t}_i, \boldsymbol{y}_i)\}_{i=1}^{K}$ with corresponding visual-prompt parameters $\{\theta_{\pi,i}\}_{i=1}^{K}$, in-context learning produces*

$$\boldsymbol{y}_q \;=\; f_{\boldsymbol{\theta}}\Big(\big[\,\pi(\boldsymbol{x}_1;\theta_{\pi,1}),\, \boldsymbol{t}_1,\, \boldsymbol{y}_1;\; \ldots;\; \pi(\boldsymbol{x}_K;\theta_{\pi,K}),\, \boldsymbol{t}_K,\, \boldsymbol{y}_K;\; \pi(\boldsymbol{x}_q;\theta_{\pi,q}),\, \boldsymbol{t}_q\,\big]\Big), \tag{16}$$

where $[\,\cdot\,;\,\cdot\,]$ denotes multimodal-sequence concatenation. $\boldsymbol{\theta}$ is frozen and no gradient flows through any $\theta_{\pi,i}$; the prompt operators are typically constructed by g_{fix} (Section 5). This is the $T = K + 1$ specialization of the compositional reasoning recurrence in Equation (5).

Challenge 6.1.1 (Visual Rationale Extraction) Extracting meaningful visual rationales from demonstration examples is crucial for effective in-context learning. Image-of-Thought (IoT) prompting (Zhou et al., 2024b)

is a training-free approach that enhances MLLMs for visual question-answering tasks by integrating discrete image-processing actions. IoT enables MLLMs to automatically design and extract visual rationales step-by-step, combining them with textual rationales to improve both accuracy and interpretability. I2L (Wang et al., 2024a) combines demonstrations, visual cues, and reasoning into a single image to enhance multimodal models' performance on complex tasks through ICL; I2L-Hybrid extends this by automatically selecting between I2L and other in-context learning methods for each task instance. VICL (Zhou et al., 2024e) retrieves relevant demonstrations, generates *intent-oriented* image summaries via task-specific visual parsing, and composes demonstrations by partially replacing demo images with these summaries to reduce cross-modal interaction burden and improve interpretability.

Challenge 6.1.2 (Memory Efficiency) Managing memory costs while maintaining effective visual demonstrations is a key challenge in visual ICL. CRG (Wan et al., 2024) is a training-free method that improves visual grounding in MLLMs by contrasting model outputs with and without specific image regions masked, guiding models to focus on relevant image areas. AIM (Gao et al., 2024) enables any MLLM to perform efficient ICL by aggregating image information from demonstrations into the latent space of corresponding textual labels, reducing memory costs by discarding visual tokens after aggregation. Beyond aggregating visual tokens into text labels, M^2IV (Li et al., 2025c) replaces explicit token-level multimodal demonstrations with a small set of learnable *Multimodal In-context Vectors* injected into MLLM residual streams, substantially reducing token overhead and enabling more scalable many-shot settings.

Limitation 6.1 (Demonstration Quality and Selection) The effectiveness of in-context learning depends heavily on the quality and relevance of demonstration examples. Methods like IoT (Zhou et al., 2024b) and I2L (Wang et al., 2024a) require carefully curated demonstrations to achieve optimal performance. The selection of appropriate visual prompts for demonstrations remains a challenge, as suboptimal prompts can mislead the model or fail to capture essential task patterns. Additionally, the scalability of ICL approaches is limited by context-window constraints, making it difficult to use large numbers of demonstrations effectively.

6.2 Few-shot Learning

Unlike ICL, *few-shot learning* updates a small set of learnable parameters on a labelled support set $\mathcal{D}_{\text{support}}$ while keeping the MLLM backbone $\boldsymbol{\theta}$ frozen. Concretely, the paradigm signature instantiates as $(\{\boldsymbol{\psi}\}, \mathcal{D}_{\text{support}}, \nabla_{\boldsymbol{\psi}}\mathcal{L})$, with the soft-prompt operator π from Definition 2.2.

Definition 6.2 (Few-shot Learning with Visual Prompts) *Given a support set* $\mathcal{D}_{\text{support}} = \{(\boldsymbol{x}_i, \boldsymbol{t}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$ *and learnable soft-prompt parameters* $\boldsymbol{\psi} \in \Psi_{\text{soft}}$*, few-shot learning solves*

$$\boldsymbol{\psi}^* = \underset{\boldsymbol{\psi}\in\Psi_{\text{soft}}}{\arg\min} \sum_{i=1}^{N} \mathcal{L}\big(f_{\boldsymbol{\theta}}\big(\boldsymbol{p}_{\text{soft}}(\boldsymbol{x}_i;\, \boldsymbol{\psi}),\, \boldsymbol{t}_i\big),\, \boldsymbol{y}_i\big), \tag{17}$$

where $\boldsymbol{p}_{\text{soft}}(\boldsymbol{x}_i;\, \boldsymbol{\psi})$ *adapts the model to the support distribution while* $\boldsymbol{\theta}$ *remains frozen, enabling generalization to novel examples with similar visual patterns.*

Challenge 6.2.1 (Prompt Learning and Adaptation) Learning effective visual prompts from few examples requires balancing task-specific adaptation against generalization. CoMM (Chen et al., 2024b) proposes a high-quality, coherent interleaved image-text dataset designed to enhance the generation capabilities of MLLMs and probe their in-context learning ability. M2oEGPT (Sheng et al., 2024) proposes an ICL framework using multimodal quantization and unified embeddings to enable joint learning of multimodal data in a general token-embedding space, combining an autoregressive transformer with a Mixture of Experts (MoE) for stable multi-task co-training. Hossain et al. (2024) introduce learnable visual prompts for both base and novel classes in semantic segmentation, along with a novel-to-base causal-attention mechanism that lets novel prompts be contextualized by base prompts without degrading base-class performance. Few-shot adaptation with visual prompts has also been explored for image manipulation: InstaManip (Lai et al., 2025) learns an unseen editing operation from a small set of exemplar image pairs plus text guidance via in-context learning, and introduces group self-attention to disentangle the "learning" and "applying" phases.

Challenge 6.2.2 (Example Selection and Ranking) Selecting optimal examples from available demonstrations is crucial for few-shot learning performance. Partial2Global (Xu et al., 2024a) selects optimal in-context

examples in visual ICL through a transformer-based list-wise ranker that compares multiple alternative samples, plus a consistency-aware ranking aggregator for globally consistent ranking. To reduce the inference-time cost of per-query prompt search, task-level prompt selection observes that many samples in a task share the same near-optimal demonstration set and proposes Top-k and greedy strategies to identify a single task-level prompt with substantially lower search cost (Zhu et al., 2025b).

Challenge 6.2.3 (Unified Multimodal Learning) Unified frameworks that handle diverse multimodal sequences and adapt quickly to new tasks are essential for practical few-shot learning. Emu2 (Sun et al., 2024a) is an MLLM trained to predict the next element in diverse multimodal sequences. Its unified architecture enables strong multimodal in-context learning abilities, allowing rapid adaptation to new tasks with just a few examples and demonstrating the potential of unified architectures for few-shot learning.

Limitation 6.2 (Generalization and Domain Shift) Few-shot learning methods face challenges in generalizing to domains significantly different from the support set. While approaches like M2oEGPT (Sheng et al., 2024) and Emu2 (Sun et al., 2024a) show promise, they may struggle when the distribution of novel examples differs substantially from the few-shot support. The learned soft prompts may also overfit to the support set, limiting applicability to diverse scenarios. The computational cost of learning prompts for each new task, though lower than full fine-tuning, can still be prohibitive for real-time applications or when dealing with many different tasks.

7 Training and Alignment for Visual Prompting

This section instantiates Axis 3 (adaptation paradigm) of the taxonomy beyond the training-light setting of Section 6, surveying the three procedures that update the MLLM parameters $\boldsymbol{\theta}$ (or a subset thereof): pre-training (Section 7.1), supervised fine-tuning (Section 7.2), and reinforcement learning with preference optimization (Section 7.3). Each procedure is a specialization of the paradigm signature $(\Theta_{\text{trainable}}, \mathcal{D}_{\text{adapt}}, \Pi_{\text{search}})$ from Section 3.5: pre-training expands $\Theta_{\text{trainable}}$ to the full $\boldsymbol{\theta}$ on large grounded corpora, SFT updates $\boldsymbol{\theta}$ (or LoRA-style subsets) on task-specific instruction data, and RL replaces direct loss minimization with reward-shaped or preference-based objectives.

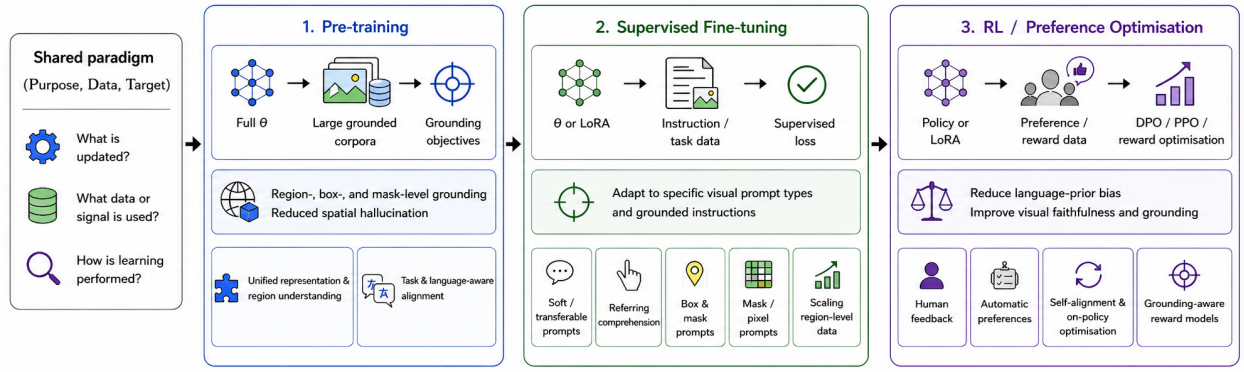

Figure 7: Training and alignment pipeline for visual prompting.

7.1 Pre-training

Pre-training instantiates the paradigm signature of Equation (6) as

$$(\boldsymbol{\theta}, \mathcal{D}_{\text{train}}^{\text{ground}}, \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{ground}}), \tag{18}$$

where $\mathcal{D}_{\text{train}}^{\text{ground}}$ is a large-scale grounded corpus (image-region-text triples or region-mask annotations) and $\mathcal{L}_{\text{ground}}$ aggregates language-modeling, mask-prediction, and spatial-alignment losses. Pre-training objectives

that ground language to spatial regions, bounding boxes, or pixel-level masks improve MLLMs' ability to interpret visual prompts at the region level and reduce spatial hallucinations.

Challenge 7.1.1 (Unified Segmentation and Region-Level Understanding) Unifying image-level, object-level, and pixel-level reasoning within a single model enables flexible interaction through various visual and text prompts. PSALM (Zhang et al., 2024j) extends MLLM capabilities to multiple image segmentation tasks by incorporating a mask decoder and a flexible input schema that unifies generic, referring, interactive, and open-vocabulary segmentation. OMG-LLaVA (Zhang et al., 2024f) bridges image-level, object-level, and pixel-level reasoning within a single model by combining a universal segmentation method as the visual encoder with an LLM, enabling flexible interaction through various visual and text prompts. VisionLLM v2 (Wu et al., 2024a) introduces an end-to-end generalist MLLM that unifies visual perception, understanding, and generation through a "super link" mechanism connecting the LLM with task-specific decoders. UrbanVLP (Hao et al., 2024) proposes a vision-language pre-training framework for urban region profiling that integrates multi-granularity information from both satellite and street-view imagery.

Challenge 7.1.2 (Explicit Spatial Reference Alignment) Explicitly aligning language with spatial references during pre-training improves fine-grained grounding and reduces the need for task-specific adaptation. Kosmos-2 (Peng et al., 2024) represents referring expressions as hyperlinked text-location pairs with special location tokens, pre-trained on GrIT, a large-scale grounded image-text dataset constructed from web data. Ferret (You et al., 2024) proposes a hybrid region representation that combines discrete coordinates with continuous visual features, together with a spatial-aware visual sampler that supports points, bounding boxes, and free-form shapes as input. Ferret-v2 (Zhang et al.) extends Ferret with any-resolution grounding and multi-granularity visual encoding by combining CLIP and DINOv2 features through a three-stage training pipeline with a dedicated high-resolution dense alignment stage. Shikra (Chen et al., 2023b) handles spatial coordinate references directly in natural language without additional vocabularies or external modules, and introduces Grounding Chain-of-Thought to reduce spatial hallucinations.

Challenge 7.1.3 (Dense Region and Mask-Level Alignment) Pre-training with dense region, mask, and segmentation supervision enables models to ground language in object-level, part-level, and arbitrary-shape visual structures. GLaMM (Rasheed et al., 2024) is the first model to generate natural language responses interleaved with object segmentation masks, pre-trained on the GranD dataset containing 810M region-level annotations with hierarchical scene, object, and part-level descriptions. GROUNDHOG (Zhang et al., 2024i) grounds LLMs to holistic segmentation by introducing grounding tokens and a masked feature extraction module, trained on a multi-granularity multi-task grounded dataset aggregated from 27 existing sources. RegionGPT (Guo et al., 2024) employs task-guided instruction prompts and a mask-pooling-based feature refinement module to extract semantic embeddings for arbitrary-shape regions, supported by an automated region captioning pipeline. Groma (Ma et al., 2024a) embeds localization directly into the visual tokenizer through a region proposer that encodes image regions into region tokens, which are then integrated into both the instruction and response of the LLM.

Limitation 7.1 (Data Scale and Domain Coverage) Pre-training for visual prompt understanding relies heavily on large-scale grounded or region-level data; scaling to long-tail concepts, diverse domains, and consistent annotation quality remains challenging. Balancing multi-granularity objectives without task interference also requires further investigation.

7.2 Supervised Fine-tuning

Supervised fine-tuning instantiates the paradigm signature as

$$(\Theta_{\text{SFT}} \subseteq \{\boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\psi}\}, \mathcal{D}^{\text{instr}}_{\text{train}}, \nabla_{\Theta_{\text{SFT}}} \mathcal{L}_{\text{SFT}}), \tag{19}$$

where Θ_{SFT} is the subset of MLLM, generator, and soft-prompt parameters updated – often LoRA-style adapters or the projection layer alone – and $\mathcal{D}^{\text{instr}}_{\text{train}}$ is a task-specific or visual-instruction dataset. Supervised

fine-tuning adapts pre-trained MLLMs to interpret specific types of visual prompts (boxes, markers, masks) and to follow region-level or grounded instructions through task-specific and instruction-tuning objectives.

Challenge 7.2.1 (Transferability and Soft Visual Prompts) Improving the transferability of soft visual prompts across MLLMs and tasks enables efficient adaptation with limited data. Transferable Visual Prompting (TVP) (Zhang et al., 2024h) improves the transferability of soft visual prompts across different MLLMs by introducing feature consistency alignment and task semantics enrichment. Fine-grained external knowledge such as OCR and segmentation can be integrated into MLLMs through visual prompts that embed information directly into a spatial embedding map (Lin et al., 2024). Explicit Visual Prompting (EVP) (Liu et al., 2023b) proposes a unified approach for low-level structure segmentation with a frozen pre-trained ViT backbone, introducing task-specific soft prompts derived from frozen patch embeddings and high-frequency image components. BlackVIP (Oh et al., 2023) adapts large pre-trained models using a Coordinator that generates soft visual prompts and SPSA-GC for efficient gradient estimation, enabling few-shot adaptation in a black-box setting. ILM-VP (Chen et al., 2023a) improves soft visual prompting by jointly optimizing input perturbation patterns and label mapping through bi-level optimization. MemVP (Jie et al., 2024) injects visual information directly into the feed-forward network weights of MLLMs, treating visual prompts as additional factual knowledge for efficient vision-language adaptation. VPG-C (Li et al., 2023a) extends Visual Prompt Generators with LLM-guided, context-aware visual feature extraction to complete missing visual details for better comprehension of demonstrative instructions.

Challenge 7.2.2 (Multi-Modal and Referring Comprehension) Training MLLMs to handle diverse visual prompt modalities and referring formats improves object-level and region-level comprehension. CoLLaVO (Lee et al., 2024) enhances MLLMs' object-level image understanding through Crayon Prompt, a visual prompt derived from panoptic color maps generated by a panoptic segmentation model. CityLLaVA (Duan et al., 2024) introduces an efficient fine-tuning framework for urban scenarios incorporating visual prompt engineering techniques including bounding box guidance, view selection, and global-local joint views. ViP-LLaVA (Cai et al., 2024) is trained to understand arbitrary visual prompts by directly overlaying visual markers onto images and processing the combined input with a visual encoder. ImageBrush (Yang et al., 2023b) introduces a framework for exemplar-based image manipulation that learns visual in-context instructions without language prompts, formulated as a diffusion-based inpainting problem. AnyRef (He et al., 2024) introduces a unified referring representation that enables the MLLM to handle diverse visual prompts (text, bounding boxes, images, audio) through instruction tuning with special formatting tokens and a refocusing mechanism for mask embeddings.

Challenge 7.2.3 (Bounding Box and Marker Prompts) Spatial instruction tuning and multi-domain visual prompt data improve alignment between box/marker inputs and language. For training with bounding box and marker prompts, GPT4RoI (Zhang et al., 2023d) proposes spatial instruction tuning that replaces bounding box references with RoI visual features extracted via RoIAlign, which are interleaved with language embeddings as input to the LLM. Draw-and-Understand (Lin et al., 2025) proposes SPHINX-V, an end-to-end MLLM equipped with a visual prompt encoder that supports points, bounding boxes, and free-form shapes, trained on MDVP-Data, a multi-domain dataset of 1.6M image-visual prompt-text triplets spanning natural images, documents, screenshots, and remote sensing. SoM-LLaVA (Yan et al., 2024a) addresses the failure of Set-of-Mark prompting in open-source MLLMs by generating synthetic tagged-image instruction data that teaches models to associate numerical tags with corresponding objects. Pink (Xuan et al., 2024) enhances referential comprehension by representing bounding boxes as normalized coordinate text tokens and employing self-consistent bootstrapping to expand dense annotations into referring-expression instruction pairs. Griffon v2 (Zhan et al., 2025) introduces a plug-and-play visual tokenizer for visual-language co-referring that processes high-resolution inputs and supports localization across images, text, and coordinate modalities through a three-stage fine-tuning pipeline.

Challenge 7.2.4 (Pixel-Level and Mask Prompts) Enabling MLLMs to consume and produce pixel-level or mask-level visual prompts supports reasoning segmentation and fine-grained grounding. For training with pixel-level and mask prompts, LISA (Lai et al., 2024) proposes the embedding-as-mask paradigm, where a special token's hidden embedding is decoded into a segmentation mask via SAM's mask decoder, enabling reasoning segmentation through end-to-end training with joint text generation and mask prediction

losses. Osprey (Yuan et al., 2024b) introduces a mask-aware visual extractor that accepts fine-grained mask regions as input, trained on Osprey-724K, a GPT-4-generated mask-text instruction dataset with object-level and part-level annotations. PixelLM (Ren et al., 2024b) produces pixel-level segmentation through a lightweight decoder with a segmentation codebook that eliminates the need for an external segmentation model, supporting multi-target reasoning segmentation via target-specific token refinement.

Challenge 7.2.5 (Scaling Region-Level Instruction Data) Scaling region-level instruction data is critical to SFT for visual prompting and cross-task generalization. SVIT (Zhao et al., 2023a) provides 4.2 million visual instruction tuning samples generated from Visual Genome (Krishna et al., 2017) annotations via GPT-4, including 1.0 million referring QA pairs with explicit region references. The All-Seeing Project (Wang et al., 2024c) creates AS-1B, a dataset of 1.2 billion region-level annotations covering 3.5 million concepts, constructed through a semi-automatic data-human-model annotation loop for panoptic visual recognition and understanding. All-Seeing V2 (Wang et al., 2024b) extends this framework to general relation comprehension with 127K relation-aware region conversation samples for understanding inter-object relationships.

Limitation 7.2 (Task Specificity and Data Quality) SFT for visual prompts often remains task- or modality-specific; transfer across prompt types and domains is limited. Data quality and consistency of region-level annotations, as well as the risk of overfitting to narrow instruction distributions, remain open challenges.

7.3 Reinforcement Learning

Reinforcement learning and preference optimization replace direct loss minimization with a reward-shaped objective. With $\theta_\pi = g(\boldsymbol{t}, \boldsymbol{x})$, $\boldsymbol{x}' = \pi(\boldsymbol{x}; \theta_\pi)$, $\boldsymbol{y} \sim f_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{x}', \boldsymbol{t})$, and a reference policy $f_{\boldsymbol{\theta}_{\text{ref}}}$, the RL paradigm-signature instantiation maximizes

$$\boldsymbol{\theta}^* \;=\; \arg\max_{\boldsymbol{\theta}} \; \mathbb{E}_{\substack{(\boldsymbol{x},\boldsymbol{t})\sim\mathcal{D}^{\text{pref}}_{\text{train}} \\ \boldsymbol{y}\sim f_{\boldsymbol{\theta}}(\cdot\mid\boldsymbol{x}',\boldsymbol{t})}} \big[\, r(\boldsymbol{x}',\, \boldsymbol{t},\, \boldsymbol{y}) \;-\; \beta\, D_{\text{KL}}\big(f_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{x}', \boldsymbol{t}) \,\|\, f_{\boldsymbol{\theta}_{\text{ref}}}(\cdot \mid \boldsymbol{x}', \boldsymbol{t})\big) \,\big], \tag{20}$$

where r is a reward function – instantiated by human preferences (RLHF), automatically constructed preference pairs (DPO and variants), self-rewarding signals, or grounding-aware reward models – and $\beta \geq 0$ controls the KL anchor to the reference policy. This objective addresses a key failure mode in visual prompting: models relying on language priors rather than faithfully attending to visual prompt regions, leading to hallucinated or spatially incorrect responses.

Challenge 7.3.1 (Human-Annotated Preference Feedback) Human feedback provides direct supervision for reducing multimodal hallucination and improving grounding. Early efforts rely on human-annotated preferences to train reward models. LLaVA-RLHF (Sun et al., 2024b) is the first to apply RLHF to MLLMs, training a reward model on 10K human-annotated hallucination-aware preference pairs and using Proximal Policy Optimization (PPO) to reduce multimodal hallucination. RLHF-V (Yu et al., 2024) introduces segment-level correctional human feedback with Dense Direct Preference Optimization (DDPO), which provides fine-grained supervision at the sentence segment level rather than the full-response level.

Challenge 7.3.2 (Automatic Preference Construction) Constructing preference pairs automatically reduces annotation cost while targeting language-prior bias and visual unfaithfulness. To reduce the cost of human annotation, subsequent works construct preference pairs automatically for Direct Preference Optimization (DPO). HA-DPO (Zhao et al., 2023b) applies DPO with hallucination-aware preference pairs constructed by using GPT-4 to detect and correct hallucinated object descriptions, producing style-consistent positive and negative response pairs. POVID (Zhou et al., 2024c) constructs preference data automatically by combining visual modality perturbations (e.g., Gaussian noise on images) with text-based hallucination injection, generating paired preferred and dispreferred responses without human annotation. V-DPO (Xie et al., 2024b) integrates Classifier-Free Guidance into the DPO objective and constructs both response-contrast and image-contrast preference pairs to improve fine-grained visual context learning. BPO (Pi et al., 2024) generates bootstrapped negative samples from distorted images and text-LLM-injected errors to expose and suppress the language-prior bias inherited from LLM pre-training during DPO.

Challenge 7.3.3 (Self-Alignment and On-Policy Optimization) Self-alignment and on-policy methods eliminate dependence on external annotators and can improve calibration with visual constraints. More recently,

self-alignment and on-policy methods eliminate the dependence on external annotators entirely. CSR (Zhou et al., 2024d) proposes Calibrated Self-Rewarding, where the MLLM evaluates its own responses using CLIP-based visual constraints as calibration signals for iterative self-improvement via DPO. RLAIF-V (Yu et al., 2025) achieves fully open-source self-alignment by using divide-and-conquer atomic claim evaluation to generate AI preference feedback, followed by iterative DPO without reliance on proprietary models. OPA-DPO (Yang et al., 2025b) identifies on-policy data as the critical factor for effective preference optimization, using a pipeline where the model's own generations are edited by a strong critic model before LoRA-SFT and DPO stages.

Challenge 7.3.4 (Grounding-Aware Reward Models) Training grounding-aware reward models offers an alternative to preference-only optimization for visual faithfulness. Beyond preference optimization, reward-based methods offer an alternative by directly training grounding-aware reward models. ViGoR (Yan et al., 2024b) improves visual grounding through a fine-grained reward model that combines automated detection rewards from GroundingDINO (Liu et al., 2024c) with human-annotated spatial grounding preference data for rejection sampling.

Limitation 7.3 (Reward Quality and Scalability) RL and preference optimization for visual prompting depend on the quality of reward or preference signals; noisy or misaligned feedback can reinforce language-prior bias or superficial grounding. Scaling self-alignment to diverse tasks and domains while maintaining calibration remains an open direction.

8 Visual Perception

This section surveys the perception leaf of Axis 4 (Inference and Reasoning), focusing on how visual prompts improve fine-grained grounding and referring – the core building blocks for cross-modal alignment. By associating textual references with visual cues (Chen et al., 2023b), visual prompting reduces visual hallucinations (Huang et al., 2024d) and grounding deficiencies (Yan et al., 2024a); recent benchmarks (Yan et al., 2025) confirm that cross-modal reasoning remains a primary failure mode, motivating the visual-text alignment metrics surveyed here.

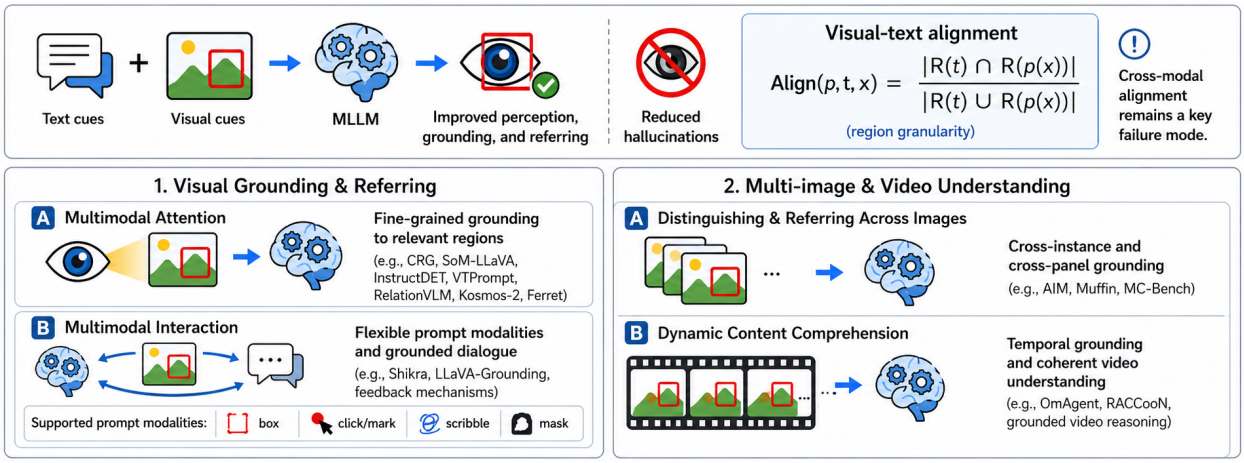

Figure 8: Visual perception with visual prompts. An image $\boldsymbol{x}$ is augmented by one or more prompt operators (e.g., bbox, marker, mask) and consumed by the MLLM to produce grounded or referring outputs at multiple granularities: pixel/region-level (grounding and referring), multi-image (cross-instance referring), and 3D. Each branch is rated by an alignment functional $\mathrm{Align}_\gamma(\,\cdot\,,\,\cdot\,)$ specialized to its native granularity γ (Definition 3.2).

Definition 8.1 (Region-Granularity Visual-Text Alignment) *Specializing Definition 3.2 to the region granularity, let $\mathcal{R}_{\boldsymbol{x}}(\boldsymbol{t}) \subseteq [H] \times [W]$ denote the set of image regions in $\boldsymbol{x}$ semantically referenced by $\boldsymbol{t}$ and*

$\mathcal{R}(\pi(\boldsymbol{x};\theta_\pi)) \subseteq [H] \times [W]$ *the regions modified or highlighted by the prompt. Then*

$$\mathrm{Align}_{\mathrm{region}}(\pi(\boldsymbol{x};\theta_\pi),\, \boldsymbol{t}) \;=\; \frac{\big|\mathcal{R}_{\boldsymbol{x}}(\boldsymbol{t}) \cap \mathcal{R}\big(\pi(\boldsymbol{x};\theta_\pi)\big)\big|}{\big|\mathcal{R}_{\boldsymbol{x}}(\boldsymbol{t}) \cup \mathcal{R}\big(\pi(\boldsymbol{x};\theta_\pi)\big)\big|}. \tag{21}$$

Pixel and token granularities give the analogous IoU and embedding-cosine specializations from Definition 3.2.

8.1 Visual Grounding and Referring

Problem 8.1 (Visual Grounding and Referring) *Given a multimodal input $(\boldsymbol{x}, \boldsymbol{t}) \in \mathcal{I} \times \mathcal{T}$, the visual-grounding/referring problem seeks a prompt parameter $\theta_\pi^* \in \Psi$ – equivalently, a target region set $\hat{\mathcal{R}} \subseteq [H] \times [W]$ – that maximizes the region-granularity alignment of Definition 8.1:*

$$\theta_\pi^* \;=\; \underset{\theta_\pi \in \Psi}{\arg\max}\; \mathrm{Align}_{\mathrm{region}}(\pi(\boldsymbol{x};\theta_\pi),\, \boldsymbol{t})\,, \qquad \hat{\mathcal{R}} \;=\; \mathcal{R}\big(\pi(\boldsymbol{x};\theta_\pi^*)\big). \tag{22}$$

Grounding corresponds to retrieving $\hat{\mathcal{R}}$ from $\boldsymbol{t}$; referring corresponds to producing $\boldsymbol{t}$ from $\hat{\mathcal{R}}$. Both reduce to maximizing the same alignment functional.

Challenge 8.1.1 (Multimodal Attention) Fine-grained visual grounding guides the MLLM's attention to relevant regions within images, enhancing regional referring abilities. CRG (Wan et al., 2024) uses contrastive regional guidance to direct the model's attention to specific areas of interest within an image, without model fine-tuning. SoM-LLaVA (Yan et al., 2024a) employs the Set-of-Mark (SoM) model to tag all the objects in the image and ask the model to list all the items. InstructDET (Dang et al., 2023) and VTPrompt (Jiang et al., 2024) further enable multimodal grounding by extracting object entities from text together with their bounding boxes. RelationVLM (Huang et al., 2024d) leverages visual prompts to enhance the MLLM's understanding and reasoning about objects' spatial relations. Complementary to masking or object tagging, Kosmos-2 (Peng et al., 2024)introduces an explicit span-to-region grounding format, representing referring expressions as text spans linked to bounding boxes via discrete location tokens, while Ferret (You et al., 2024) extends region prompting beyond boxes to points and free-form shapes through a hybrid discrete-continuous region representation; Ferret-v2 (Zhang et al.) further improves fine-grained perception with higher-resolution handling and multi-granularity visual encoding. Beyond prompt engineering, recent training-free analyses show that only a few specialized attention heads in frozen MLLMs can act as "localization heads" whose attention maps can be assembled into bounding-box or mask predictions for referring queries (Kang et al., 2025).

Challenge 8.1.2 (Multimodal Interaction) Methods integrate diverse visual prompts at varying levels of granularity, focusing on more fine-grained and flexible multimodal interactions. Free-form visual prompt inputs (Lin et al., 2025) and feedback mechanisms (Liao et al., 2024) refine visual understanding. Shikra (Chen et al., 2023b) extends these capabilities to visual dialogue systems, where MLLMs respond to referential cues within a conversational context, achieving more precise and context-aware interactions. LLaVA-Grounding (Zhang et al., 2024b) further unifies grounded dialogue with object- and pixel-level grounding by connecting a grounding model to an LMM, supporting multiple prompt modalities (click/mark, box, scribble).

8.2 Multi-image and Video Understanding

Definition 8.2 (Multi-image / Video Visual Prompting) *A multi-image input is a tuple $\boldsymbol{x}_{1:K} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K) \in \mathcal{I}^K$ paired with a single textual instruction $\boldsymbol{t}$. Each image $\boldsymbol{x}_k$ carries its own prompt parameter $\theta_{\pi,k} \in \Psi$, yielding a per-image augmented sequence $\big(\pi(\boldsymbol{x}_k;\theta_{\pi,k})\big)_{k=1}^{K}$, and the MLLM output is*

$$\boldsymbol{y} \;=\; f_{\boldsymbol{\theta}}\Big(\big[\pi(\boldsymbol{x}_1;\theta_{\pi,1});\, \ldots;\, \pi(\boldsymbol{x}_K;\theta_{\pi,K})\big],\, \boldsymbol{t}\Big), \tag{23}$$

where $[\,\cdot\,;\,\cdot\,]$ denotes multimodal-sequence concatenation. For video, $\boldsymbol{x}_{1:K}$ is a temporally ordered frame sequence and we use a temporal *alignment functional* $\mathrm{Align}_{\mathrm{temporal}}(\boldsymbol{x}'_{1:K}, \boldsymbol{t})$ *extending Definition 3.2 to the multi-image setting.*

Challenge 8.2.1 (Distinguishing and Referring Across Images) Accurately distinguishing and referring to objects across multiple images requires precise alignment between textual descriptions and visual elements. Fan et al. (2024c) introduce a benchmark dataset with multipanel images, focusing on evaluating MLLM capabilities in distinguishing objects across panels and navigating visual relationships – a critical resource for assessing other methods. To more directly test multi-image *instance-level* grounding, MC-Bench (Xu et al., 2025d) formulates multi-context visual grounding (localizing target instances across multiple images using open-ended text prompts). In contrast, Pan et al. (2024) leverage morph-token auto-encoding to enhance intrinsic representation learning for visual grounding across multiple images, emphasizing model-architecture improvements. Li et al. (2023a) employ fine-tuning to teach MLLMs to follow in-context demonstrative instructions for multi-image tasks, which, while effective, relies on curated datasets and lacks the adaptability of AIM (Gao et al., 2024). AIM dynamically adjusts grounding and referring abilities to new visual contexts, offering a flexible solution compared to the static nature of fine-tuning or the benchmark-focused approach of Fan et al. (2024c).

Challenge 8.2.2 (Dynamic Content Comprehension) Handling dynamic video content involves identifying regions of interest and maintaining coherence across temporal visual features. OmAgent (Zhang et al., 2024d) develops a visual-prompting method to enable task division in video understanding by annotating a series of visual features. RACCooN (Yoon et al., 2024) uses visual prompts to guide MLLMs in identifying target regions in videos for manipulation. Wu et al. (2024c) ground objects across videos, enabling the model to comprehend and refer to objects in dynamic scenes.

Limitation 8 (Scalability and Robustness) Current methods often struggle with scalability and robustness when processing extensive datasets of multi-image and video inputs. Techniques such as fine-tuning show promise but require significant computational resources (Fan et al., 2024c; Li et al., 2023a). Dynamic adaptability in new contexts is also limited, as models may experience distribution shift (Wu et al., 2024c) in rapidly changing scenarios or when presented with unseen visual relationships. Recent analyses identify strong *position bias* in multi-image MLLMs: predictions can shift substantially when the same images are permuted, motivating training-free attention interpolation (SoFA) to mitigate this bias and improve multi-image reasoning robustness (Tian et al., 2025).

9 Compositional Reasoning

This section surveys the compositional-reasoning leaf of Axis 4: how visual prompts support multi-step, multimodal reasoning by chaining prompt operators across time. We examine visual planning, step-by-step reasoning, grounded perception-and-action, and visual imagination as reasoning substrate. Each setting instantiates the compositional-reasoning recurrence $\boldsymbol{y}_t = f_{\boldsymbol{\theta}}(\pi(\boldsymbol{x}_t; \theta_{\pi,t}),\, \boldsymbol{t}_t,\, \boldsymbol{h}_{t-1})$ introduced in Equation (5), differing in how $\theta_{\pi,t}$ and $\boldsymbol{x}_t$ are produced.

Definition 9.1 (Compositional Visual Reasoning) *Compositional visual reasoning is the T-step specialization of Definition 2.2 obtained by unrolling Equation* (5)*: given prompt parameters $\{\theta_{\pi,t}\}_{t=1}^{T}$, textual queries $\{\boldsymbol{t}_t\}_{t=1}^{T}$, and an initial visual state $\boldsymbol{x}_1$, the model produces intermediate outputs $\boldsymbol{y}_t = f_{\boldsymbol{\theta}}(\pi(\boldsymbol{x}_t; \theta_{\pi,t}),\, \boldsymbol{t}_t,\, \boldsymbol{h}_{t-1})$ with running history $\boldsymbol{h}_t = [\boldsymbol{h}_{t-1};\, (\pi(\boldsymbol{x}_t; \theta_{\pi,t}), \boldsymbol{t}_t, \boldsymbol{y}_t)]$, where $\boldsymbol{x}_{t+1}$ may be the original image, a tool-processed update, or a model-generated visual artifact (visual imagination, Section 9.1).*

9.1 Visual Planning

Definition 9.2 (Visual Planning Recurrence) *The* visual planning recurrence *specializes Equation* (5) *by allowing the prompt parameter at each step to be produced by either an external tool or the MLLM itself:*

$$\theta_{\pi,t} \;=\; \begin{cases} g_{\text{fix}}(\boldsymbol{t}_t,\, \boldsymbol{x}_t) & \textit{(tool-driven branch: step-by-step / grounded action)}, \\ f_{\boldsymbol{\theta}}\big(\boldsymbol{x}'_{t-1},\, \boldsymbol{t}_t,\, \boldsymbol{h}_{t-1}\big) & \textit{(self-generated branch: visual imagination)}, \end{cases} \tag{24}$$

with $\boldsymbol{y}_t = f_{\boldsymbol{\theta}}\big(\pi(\boldsymbol{x}_t; \theta_{\pi,t}),\, \boldsymbol{t}_t,\, \boldsymbol{h}_{t-1}\big)$ and $\boldsymbol{h}_t = [\boldsymbol{h}_{t-1};\, (\pi(\boldsymbol{x}_t; \theta_{\pi,t}),\, \boldsymbol{t}_t,\, \boldsymbol{y}_t)]$. The tool-driven branch covers step-by-step visual reasoning (Challenge 9.1.1) and grounded perception and action (Challenge 9.1.2); the self-generated branch covers visual imagination (Challenge 9.1.3).

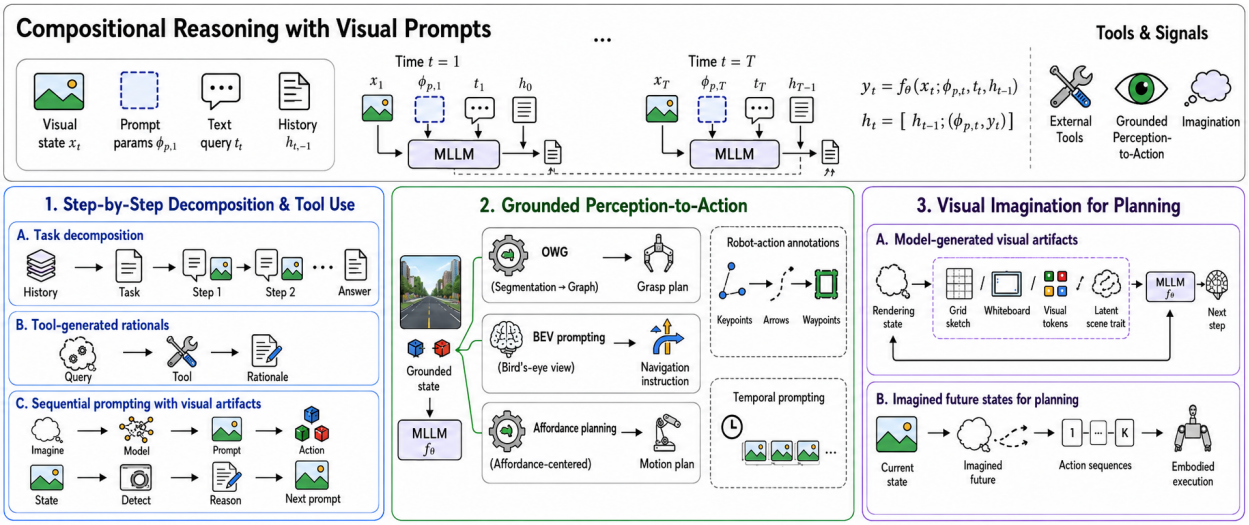

Figure 9: Compositional reasoning with visual prompts. A reasoning timeline $t = 1 \ldots T$ alternates textual queries $\boldsymbol{t}_t$ with visual-prompt parameters $\theta_{\pi,t}$ that are either retrieved from external tools (planning, grounded action) or *generated* by the model itself (visual imagination). The running history $\boldsymbol{h}_{t-1}$ feeds the MLLM $f_{\boldsymbol{\theta}}$ at every step, instantiating Equation (5).

Visual prompting enhances visual planning by integrating reasoning, segmentation, and motion planning into multimodal frameworks, enabling MLLMs to interpret complex visual reasoning tasks and generate actionable plans. We organize this body of work around the three branches of Definition 9.2: step-by-step visual reasoning that decomposes complex tasks into grounded intermediate steps (Challenge 9.1.1), grounded perception and action that connects visual understanding with physical or navigational planning (Challenge 9.1.2), and visual imagination for planning that generates novel visual artifacts as reasoning substrate (Challenge 9.1.3).

Challenge 9.1.1 (Step-by-Step Visual Reasoning) Visual planning often requires breaking down complex tasks into sequential visual and textual steps (Luan et al., 2024). Zhou et al. (2024b) propose the Image-of-Thought (IoT) prompting method, which compels MLLMs to automatically design visual and textual steps. This method leverages external image processing tools to generate a multimodal rationale series, guiding MLLMs through step-by-step processes for complex visual reasoning tasks. Zhou et al. (2024a) introduce the Chain-of-Imagination (CoI) method and develop an embodied agent named MineDreamer in Minecraft. This approach envisions the sequential execution of instructions using an LLM-enhanced diffusion model, translating imagined steps into precise visual prompts to generate accurate agent actions. Wu et al. (2024g) proposes DetToolChain to unlock the potential of MLLMs in object detection task. This method involves using a "detection prompting toolkit," which includes visual processing and detection reasoning prompts, to reason the sequential implementation of the detection prompts.

A growing line of work grounds each reasoning step in explicit visual regions, moving beyond text-only rationales. VisCoT (Shao et al., 2024) introduces a large-scale dataset where each question-answer pair is annotated with intermediate bounding boxes highlighting key image regions; the MLLM first outputs bounding boxes as visual chain-of-thought, then zooms into those regions to produce answers through a multi-turn focus-then-answer pipeline. VoCoT (Li et al., 2025d) extends this idea to an object-centric formulation, where each reasoning step revolves around objects represented as tuples of textual description, coordinates, and visual representation. A RefBind mechanism incorporates visual representations of detected objects into the reasoning chain, creating multimodal interleaved reasoning paths. Argus (Man et al., 2025) further advances the paradigm by introducing goal-directed visual attention grounding, where the model predicts regions of interest during reasoning and re-engages with those image regions by extracting additional visual tokens via re-encoding, mimicking how humans shift focus during multi-step reasoning. UV-CoT (Zhao

et al., 2025a) addresses the supervision bottleneck of these methods by generating bounding-box regions as intermediate visual reasoning steps without annotated chain-of-thought data, instead using preference optimization with a separate MLLM judge to learn region-based reasoning preferences.

Rather than grounding reasoning in image regions, some methods construct structured compositional representations as intermediate steps. CCoT (Mitra et al., 2024) proposes a zero-shot two-step approach where the MLLM first generates a scene graph capturing objects, attributes, and relationships, then uses this structured representation as compositional context for answering, improving performance without fine-tuning. Aurora (Bigverdi et al., 2025) takes a different approach by tokenizing intrinsic image representations – depth maps, bounding boxes, and instance-segmentation masks – as auxiliary perception tokens. The model generates these perception tokens as intermediate reasoning steps before producing final answers.

As visual reasoning tasks grow more complex, several methods address the challenge of maintaining coherence across extended reasoning chains. LLaVA-CoT (Xu et al., 2025a) introduces a four-stage autonomous reasoning pipeline – summary, visual caption, logical reasoning, and conclusion – with stage-wise beam search that enables self-correction through backtracking during multi-step visual reasoning. Insight-V (Dong et al., 2025) addresses long-chain reasoning through a multi-agent system with dedicated reasoning and summarization agents, showing that decomposing these roles enables more reliable extended chains. VCTP (Chen et al., 2024e) implements a three-stage see-think-confirm pipeline for knowledge-based visual reasoning, where each stage produces visual annotations that ground the next stage's reasoning in external knowledge.

Challenge 9.1.2 (Grounded Perception and Action) Grounded world understanding involves connecting visual perception with actionable planning and decision-making. OWG (Tziafas & Kasaei, 2024) focuses on combining segmentation and grasp synthesis models, enabling robust segmentation, grasp planning, and ranking for a comprehensive understanding of the physical environment. Compared to OWG, BEVInstructor (Fan et al., 2024a) takes a different approach by incorporating Bird's Eye View (BEV) representations as visual prompts into MLLMs, emphasizing spatial reasoning to generate navigation instructions. While OWG excels in tasks involving fine-grained interaction with physical objects, BEVInstructor prioritizes high-level spatial understanding and textual reasoning for navigation tasks. AO-Planner (Chen et al., 2024a) further diverges by adopting an affordance-based framework for motion planning. Unlike OWG and BEVInstructor, AO-Planner integrates VAP methods with a PathAgent to facilitate affordance-oriented action decision-making, focusing on actionable possibilities within the environment rather than static spatial or object representations.

A recent trend is the use of visual annotations – keypoints, numbered markers, and trajectory overlays – as the dominant interface between VLM reasoning and continuous robotic action spaces. PIVOT (Nasiriany et al., 2024) establishes the template for this paradigm: candidate robot actions such as arrows, numbered keypoints, and trajectories are annotated directly on camera images, and a VLM selects among proposals through iterative refinement, enabling zero-shot robotic navigation and manipulation without robot-specific training. MOKA (Liu et al., 2024a) adapts Set-of-Mark prompting for open-world manipulation, annotating RGB images with candidate keypoints and grid cells, then querying GPT-4V to select affordance points and motion waypoints, effectively converting manipulation planning into a series of visual question-answering problems. Building on the iterative prompting paradigm, several works develop increasingly sophisticated keypoint representations for robotic control. ReKep (Huang et al., 2024c) proposes relational keypoint constraints, where a visual foundation model proposes keypoint candidates overlaid on the scene image, and GPT-4o generates Python constraint functions over 3D keypoints that are optimized hierarchically for real-time robot trajectories, enabling multi-stage, bimanual, and reactive manipulation. KUDA (Liu et al., 2025) unifies dynamics learning and visual prompting through keypoints as a shared intermediate representation, where VLM-generated cost functions are optimized jointly with a neural dynamics model, achieving 30% improvement over baselines on diverse objects including deformable and granular materials. CrayonRobo (Li et al., 2025b) introduces crayon visual prompts – simple 2D drawn markers on key-frame images indicating contact points, gripper directions, and movement directions for each step in a task sequence – predicting accurate contact poses and SE(3) movement directions for long-horizon manipulation. RoboPoint (Yuan et al., 2024a) complements these zero-shot approaches as a VLM fine-tuned to predict spatial affordance keypoints, outperforming GPT-4o in affordance prediction.

Extending beyond structured manipulation, recent methods scale visual prompting to open-world and cross-domain settings. ROCKET-1 (Cai et al., 2025) introduces visual-temporal context prompting for open-world agents, where VLMs generate object segmentation masks that serve as visual prompts to guide a low-level policy, with different mask colours representing different interaction types; SAM-2 provides real-time tracking in Minecraft's dynamic environment. Magma (Yang et al., 2025a) proposes a foundation model pretrained with two complementary visual prompting strategies: Set-of-Mark (SoM) labels actionable objects with numbered bounding boxes, and Trace-of-Mark (ToM) labels object movement trajectories in videos, enabling a single model to perform UI navigation, robot manipulation, and visual planning.

Challenge 9.1.3 (Visual Imagination for Planning) A distinctive emerging direction involves models that *generate* intermediate visual outputs – images, sketches, diagrams, or visual tokens – as part of their reasoning process, forming a visual imagination loop. Unlike Challenges 9.1.1 and 9.1.2, where visual prompts annotate or act on existing images, these methods create novel visual artifacts that did not exist in the input and then reason over them. This is the $\boldsymbol{x}_{t+1} = \pi(\boldsymbol{x}_t; \theta_{\pi,t})$ branch of Definition 9.1 in which $\theta_{\pi,t}$ is itself produced by the MLLM rather than by an external tool.

Early efforts in this direction leverage text or code to produce visual reasoning artifacts. VoT (Wu et al., 2024f) elicits spatial reasoning by having LLMs generate text-based grid-world visualizations at each intermediate reasoning step, adding a visuospatial sketchpad inspired by the human "Mind's Eye" to each chain-of-thought step and significantly outperforming standard chain-of-thought on multi-hop spatial tasks. WoT (Menon et al., 2024) extends this idea by providing MLLMs a metaphorical whiteboard to draw intermediate reasoning steps as images through Turtle graphics or Matplotlib code; the rendered images are fed back to the model for further reasoning. More recent methods train models to generate visual tokens as an intrinsic part of the reasoning process, eliminating the need for external rendering. MVoT (Li et al., 2025a) extends VoT to multimodal LLMs by training them to jointly autoregress interleaved verbal and visual tokens, directly generating diagrammatic mental models as part of planning. Latent Sketchpad (Zhang et al., 2025b) takes a more architecture-agnostic approach, equipping pretrained MLLMs with an internal visual scratchpad that generates continuous visual latents during reasoning rather than pixel-level images. A Context-Aware Vision Head and a pretrained Sketch Decoder allow this approach to generalize across architectures without retraining from scratch.

Visual imagination also extends to embodied settings, where models predict future visual states to guide physical actions. CoT-VLA (Zhao et al., 2025b) autoregressively predicts future image frames as visual subgoals before generating action sequences, showing that visual-subgoal imagination directly improves physical task execution. MineDreamer (Zhou et al., 2024a) similarly generates visual goal-state prompts via diffusion to guide embodied agents through its Chain-of-Imagination method. VisuoThink (Wang et al., 2025b) combines vision-text interleaved generation with look-ahead tree search, including rollout simulation and self-voting selection – representing the convergence of visual imagination with test-time compute scaling, suggesting that visual reasoning can benefit from the same inference-time scaling strategies that have proven effective for text-only reasoning.

Limitation 9 (Generalization and Faithfulness) Despite advancements, current methods face two coupled challenges. *Task generalization* remains difficult: approaches like OWG (Tziafas & Kasaei, 2024) and AO-Planner (Chen et al., 2024a), while effective in specific domains, struggle to adapt to dynamic or unseen scenarios without retraining or significant fine-tuning. Addressing this requires more flexible and scalable frameworks capable of self-adapting to previously unseen environments. *Faithfulness* is the second challenge: methods that generate intermediate visual reasoning artifacts – bounding boxes in VisCoT (Shao et al., 2024), perception tokens in Aurora (Bigverdi et al., 2025), imagined future frames in CoT-VLA (Zhao et al., 2025b) – risk artifacts that do not accurately reflect the true visual content, leading to error propagation across the reasoning chain. UV-CoT (Zhao et al., 2025a) partially addresses this through preference optimization, and VisuoThink (Wang et al., 2025b) employs tree search with self-voting, but systematic approaches to verify and correct intermediate visual reasoning steps remain underdeveloped.

10 Evaluation

This section presents the evaluation landscape for visual prompting in MLLMs. We first organize the available datasets and benchmarks by modality and annotation strategy (Section 10.1), then compare representative methods across vision-language benchmarks and finally distil cross-task patterns and their root causes back to the four-axis taxonomy of Section 3.1 (Section 10.2).

10.1 Dataset and Benchmark

The development and evaluation of state-of-the-art vision-language models rely on diverse, comprehensive datasets that challenge models across a variety of tasks. These datasets span visual question answering, panoptic segmentation, and video understanding, providing complementary benchmarks for assessing robustness, generalization, and task-specific performance. Table 2 organizes them by modality, prompt type, and annotation strategy.

Table 2: **Benchmark and training-dataset overview.** Each row summarizes a dataset, the visual-prompt types it supports – Semantic (SP), Textual (TP), Generative (GP) – the modalities covered (Image, Video, Audio), and whether visual prompts are produced manually, automatically, or both.

Reference	SP	TP	GP	Image	Video	Audio	Manual	Automatic
Image-only benchmarks								
MDVP-Bench (Lin et al., 2025)	✓	✓	✓	✓			✓	✓
A3VLM (Huang et al., 2024a)	✓			✓				✓
GPT-4V Challenger (Fu et al., 2023)	✓		✓	✓				
EarthMarker (Zhang et al., 2024g)	✓	✓	✓	✓			✓	
MultipanelVQA (Fan et al., 2024c)			✓	✓				
Groma (Ma et al., 2024a)	✓	✓		✓			✓	✓
Image + Video benchmarks								
VLM Feedback (Li et al., 2023c)	✓			✓	✓		✓	✓
GLEE (Wu et al., 2024c)	✓	✓		✓	✓			
RACCooN (Yoon et al., 2024)			✓		✓		✓	✓
Audio-inclusive multimodal benchmarks								
Safety of MLLMs (Liu et al., 2024d)			✓	✓	✓	✓		
AutoAD-Zero (Xie et al., 2024a)	✓		✓	✓		✓		✓
MM-Vid (Lin et al., 2023)	✓	✓	✓	✓	✓	✓	✓	

10.2 Cross-Task Analysis and Root Causes

While raw benchmark scores indicate task-level performance, they do not explain why methods differ. To address this, we conduct a cross-task analysis that identifies recurring performance patterns and traces their root causes back to the four-axis taxonomy of Section 3.1.

A. Spatial grounding deficits. The strongest signal is that dense panoptic and mask-prediction tasks (COCO Panoptic, ADE20K-150) admit only a small set of competitive methods, all of which produce *pixel-aligned* prompt parameters in Ψ_{S}: OMG-LLaVA, Groma, and VisionLLM v2 occupy the top of the panoptic column, while bounding-box- and marker-only methods such as SoM-LLaVA, Shikra, and VTPrompt trail by several points despite comparable backbones. The root cause is a granularity mismatch on Axis 1: when label granularity is pixel-level, prompt parameters living in Ψ_{B} or Ψ_{M} inherit a structural precision ceiling that cannot be closed by larger backbones or longer training. The actionable implication is that the prompt-representation subspace should be chosen to match the supervision granularity of the downstream task, rather than treated as a free design choice.

B. Cross-modal sensitivity and variance. Open-ended VQA benchmarks (VQAv2, GQA, OKVQA) show the widest cross-method variance despite all relying on instruction-tuned MLLMs. The variance correlates with the choice of generator g (Axis 2): methods that route prompt generation through an explicit detector g_{det} consistently outperform methods that rely on implicit attention. The combination of high VQA variance with stable detection-driven gains points to two coupled root causes: prompt-phrasing sensitivity under g_{fix} heuristics and incomplete visual grounding under attention-only generation. Standardizing g_{fix}

Table 3: **Diagnostic patterns and root causes of visual-prompting performance.** Each row pairs an observed performance pattern derived from the empirical results with the representative methods/benchmarks where it manifests, the structural root cause expressed in the four-axis taxonomy (Section 3.1), and a concise actionable takeaway. Rows are grouped into five diagnostic clusters spanning spatial precision, cross-modal sensitivity, modality coverage, 3D transfer, and the manual-vs-automatic trade-off. Axis tags: A1 = representation granularity, A2 = generation mechanism, A3 = adaptation paradigm, A4 = capability enabled.

Observed pattern	Representative evidence	Root cause (axis)	Takeaway
A. Spatial grounding deficits	segmentation, dense prediction, region-level QA		
Pixel-aligned prompts dominate dense panoptic / mask tasks	OMG-LLaVA, Groma, VisionLLM v2 on COCO Panoptic and ADE20K-150 (Zhang et al., 2024f; Ma et al., 2024a; Wu et al., 2024a)	Bounding-box / marker prompts lack pixel precision; only Ψ_{S} matches mask-level supervision **(A1)**	For dense segmentation, pixel/mask is the only Ψ subspace that reaches SOTA; bbox/marker variants saturate early.
Detection-style markers underperform on dense panoptic segmentation	SoM-LLaVA, Shikra, VTPrompt trail OMG-LLaVA / Groma on panoptic IoU (Yan et al., 2024a; Chen et al., 2023b; Jiang et al., 2024)	Granularity mismatch: prompt granularity coarser than label granularity **(A1)**	Match prompt granularity to label granularity; coarse prompts inherit a precision ceiling regardless of backbone scale.
B. Cross-modal sensitivity and variance	VQA, compositional / referring reasoning		
High variance on open-ended VQA benchmarks	VQAv2 / GQA / OKVQA spread across ViP, MineDreamer, A3VLM, VPG-C (Cai et al., 2024; Zhou et al., 2024a; Huang et al., 2024a; Li et al., 2023a)	Sensitivity to prompt phrasing and incomplete visual grounding **(A2 + A4)**	Standardize g_{fix} templates and report alignment-score (Definition 3.2) alongside accuracy to attribute the variance.
Detection-based generators close the compositional gap	A3VLM (RefCOCO 80.3), SoM-LLaVA (RefCOCO 79.1) vs. VisionLLM v2 (63.9) (Huang et al., 2024a; Yan et al., 2024a; Wu et al., 2024a)	Explicit g_{det} outperforms implicit attention for object-level referring **(A2)**	For referring and compositional reasoning, route generation through an explicit detector; implicit attention is brittle.
C. Modality and temporal coverage gaps	video, audio, multi-image		
Sparse method coverage on video benchmarks	Only OmAgent, MM-Vid, RACCooN, DetToolChain populate MovieNet; many methods leave it blank (Zhang et al., 2024d; Lin et al., 2023; Yoon et al., 2024; Wu et al., 2024g)	π rarely typed over a frame sequence $\boldsymbol{x}_{1:K}$; few temporal prompt operators in Ψ **(A1 + A4)**	Temporal prompt operators (per-frame markers, trajectory overlays, key-frame masks) are an open frontier for Ψ.
Audio-inclusive multimodal evaluation is underexplored	Only MM-Vid, AutoAD-Zero, Safety-of-MLLMs span audio in Table 2 (Lin et al., 2023; Xie et al., 2024a; Liu et al., 2024d)	$\mathcal{I}$ defined over image space; no canonical audio-visual prompt operator **(A1)**	Extending $\mathcal{I}$ and π to audio-visual co-prompts is a clean opportunity for the next generation of benchmarks.
D. 3D and domain transfer	PartNet-Mobility, 3D reasoning, cross-domain		
2D-prompt methods underperform on 3D reasoning	VisionLLM v2 / Groma absent on PartNet-Mobility; only A3VLM (76.4), RelationVLM (81.7), InstructDET (84.3) competitive (Wu et al., 2024a; Ma et al., 2024a; Huang et al., 2024a;d; Dang et al., 2023)	2D prompt parameter space mismatched with 3D geometry; no native $\Psi_{\mathrm{B}}^{\mathrm{3D}}$ analogue **(A1)**	3D-native subspaces – 3D bboxes, BEV markers, viewpoint prompts – are required; 2D upcasting saturates quickly.
Domain-specific prompts do not transfer	CityLLaVA strong in urban scenes but blank on most other benchmarks (Duan et al., 2024)	g_{fix} encodes domain priors that do not survive distribution shift **(A2 + A3)**	Couple domain-specific g_{fix} with a small learnable g_{learn} head for cross-domain transfer rather than retraining from scratch.
E. Manual vs automatic prompting trade-off	scalability, precision, alignment noise		
Manual prompting peaks per-benchmark but does not scale	ViP (OKVQA 85.3), CityLLaVA (COCO Cap. 75.5) hit local peaks but cover few benchmarks (Cai et al., 2024; Duan et al., 2024)	Per-image manual annotation overhead bounds dataset breadth **(A2)**	Treat manual prompting as a per-task ceiling estimate; production systems require g_{learn} or toolchain composition.
Automatic / learnable prompting broadens coverage but introduces alignment noise	VPGTrans, TVP, MemVP populate many benchmarks but trail manual peaks on each (Zhang et al., 2024a;h; Jie et al., 2024)	LLM-generated prompts drift semantically; learnable $\boldsymbol{\psi}$ overfits the support distribution **(A2 + A3)**	Hybrid pipelines (auto-propose + grounding-aware reward, Equation (20)) recover the ceiling while preserving breadth.

templates and reporting the alignment score $\mathrm{Align}_{\gamma}(\pi(\boldsymbol{x};\theta_{\pi}),\, \boldsymbol{t})$ from Definition 3.2 alongside task accuracy would let future evaluations attribute variance to its true source.

C. Modality and temporal coverage gaps. Video benchmarks (MovieNet) are sparsely populated and audio-inclusive evaluation in Table 2 is dominated by MM-Vid, AutoAD-Zero, and the safety benchmark of Liu et al. (2024d). The structural cause is that the visual-prompt operator π in Definition 2.2 is typed over a single image $\boldsymbol{x} \in \mathcal{I}$; extending it to frame sequences $\boldsymbol{x}_{1:K}$ (Definition 8.2) or to audio-visual co-prompts requires a temporal or cross-modal prompt parameter subspace largely absent from the surveyed works.

The coverage gap is therefore not an annotation-cost problem alone but a representation-space problem on Axis 1, which makes per-frame markers, trajectory overlays, and key-frame masks a clean opportunity for the next generation of Ψ extensions.

D. 3D and domain transfer. Performance on PartNet-Mobility separates 3D-native methods (A3VLM, RelationVLM, InstructDET) from 2D-prompt methods (VisionLLM v2, Groma). Likewise, domain-specialized methods such as CityLLaVA peak in their target domain. Both patterns trace to Axis 1 – no native $\Psi_{\mathrm{B}}^{\mathrm{3D}}$ analogue for 2D bounding-box prompts – and Axis 2 – g_{fix} generators that encode domain priors do not survive distribution shift. 3D-native prompt subspaces (3D bounding boxes, BEV markers, viewpoint prompts) are the principled fix; coupling a domain-specific g_{fix} with a small learnable g_{learn} head is a lower-cost intermediate route for cross-domain transfer.

E. Manual vs automatic prompting trade-off. Manual prompting methods (ViP, CityLLaVA) achieve per-benchmark peaks but cover only a handful of columns, while learnable or transferable prompting (VPG-Trans, TVP, MemVP) populates many benchmarks but trails the manual peaks on each. The mechanism is direct: per-image manual annotation overhead bounds the breadth of g_{fix}, while LLM-generated prompts and learnable $\boldsymbol{\psi}$ drift semantically (Axis 2) or overfit the support distribution (Axis 3). Manual prompting should therefore be read as a per-task ceiling estimate, while production deployment requires hybrid pipelines in which an automatic proposer is followed either by a manual verifier or by a grounding-aware reward as in the RL objective of Equation (20).

Synthesis. Across the five clusters, the same three structural levers recur: (i) match the prompt-representation subspace within Ψ to the task's supervision granularity (A1); (ii) route generation through an explicit, axis-aware g rather than relying on implicit attention (A2); and (iii) close the manual–automatic gap with hybrid pipelines anchored by an alignment functional or a grounding-aware reward (A2 + A3). Each lever maps to a concrete operator in Sections 2 and 3, so the gaps surfaced by this diagnosis correspond to specific design choices in the formalism rather than to missing engineering effort alone.

11 Frontier Applications

The five frontier areas surveyed here populate Axis 4 (Inference and Reasoning) beyond the canonical perception/reasoning leaves of Sections 8 and 9. Jailbreaking and safety (Section 11.1) and debiasing (Section 11.3) map onto the *Safety & Debias* leaf, hallucination mitigation (Section 11.2) bridges perception and safety, and visual generation (Section 11.4) and 3D understanding (Section 11.5) populate the *Generation & Control* leaf. Table 4 summarizes representative works per area. Each subsection is short by design: these are open frontiers rather than mature subfields.

11.1 Jailbreaking and Safety

While visual prompting enables fine-grained instructions to MLLMs for better response generation, it can also be intentionally designed to expose critical safety issues (Liu et al., 2024d; Ni et al., 2024). Instead of feeding harmful textual instructions directly, Gong et al. (2023) converts them into images through typography and feeds them to MLLMs as visual prompts; the results show that visual prompting opens a new jailbreaking surface even when the underlying LLM has been safety-aligned.

To further expose safety problems for red-teaming, multimodal jailbreaking prompts combining textual and visual instructions have also been studied. Ying et al. (2024) first embed a harmful perturbation in the visual prompt and then optimize the textual prompt through LLM reasoning on the harmful intent in the image. Liu et al. (2024e) use a red-team MLLM and a red-team LLM guided by reinforcement learning to automatically generate visual and textual jailbreaking prompts, respectively. Their results suggest that multimodal prompts can mount stronger attacks on MLLMs that fuse multimodal input features. Furthermore, Gu et al. (2024) observe a more severe issue of *infectious jailbreaking* in multi-agent MLLM environments: a single adversarial image jailbreaks one agent and, with no further intervention, harmful behaviour spreads through the multi-agent interaction at an exponential rate.

Table 4: **Frontier applications of visual prompting.** Each row summarizes an application area, representative works, the core prompting mechanism, and a concise takeaway.

Area	Representative works	Visual prompting mechanism	Takeaway
Jailbreaking & Safety		(Liu et al., 2024d; Ni et al., 2024)	
Text-as-image jailbreak	Gong et al. (2023)	Render harmful text as an image (typography) to bypass text filters	Visual channels create new attack surfaces even for safety-aligned LLMs.
Optimized multimodal jailbreaks	Ying et al. (2024); Liu et al. (2024e)	Embed harmful visual perturbations and co-optimize text+image prompts (incl. RL red-teaming)	Joint prompts can amplify attacks, especially for early-fusion MLLMs.
Infectious jailbreak in multi-agent settings	Gu et al. (2024)	Jailbreak one agent via an adversarial image; harmful behavior propagates via interaction	Safety must be evaluated at the multi-agent *system* level.
Hallucination			
Decoding-time grounding	Favero et al. (2024)	Mutual-information decoding to upweight vision-conditioned signals	Hallucination can be reduced without retraining via decoding control.
Detector-assisted prompting	Jiang et al. (2024)	Use detectors to highlight relevant objects/regions as visual prompts	External perception modules can sharpen object-level grounding.
Multi-object hallucination probing	Chen et al. (2024c)	Use visual referring prompts to test grounding over multiple target objects	Multi-object focus increases risk; object-wise prompting can help.
Debiasing			
Grounding against spurious correlations	(Ye et al., 2024)	Emphasize essential visual evidence (objects/relations) to reduce shortcut reliance	Visual evidence can mitigate bias by reducing spurious textual cues.
Causal robustness via controlled evidence	(Conceptual direction)	Prompt with minimal/counterfactual visual edits to stress-test invariances	Visual prompts enable causal-style auditing by manipulating evidence.
Visual Generation			
Spatially controlled generation	Zhang et al. (2023c); Mou et al. (2024)	Use structural conditions (e.g., edges/depth/pose) as visual prompts	Visual prompts provide high-bandwidth spatial control beyond text.
In-context diffusion generation	Wang et al. (2023)	Condition on exemplar vision-language prompts after multi-task finetuning	Visual prompting can unlock in-context behavior for diffusion models.
Example-driven image manipulation	Yang et al. (2023b)	Use before/after exemplar pairs as demonstrations to specify edit intent	Demonstrations reduce language ambiguity for precise edits.
Restoration and 2D→3D prompting	Ai et al. (2024); Chen et al. (2024d)	Leverage diffusion priors; use 2D prompts to guide restoration / text-to-3D	Visual prompts transfer 2D priors to restoration and 3D pipelines.
Interactive multimodal creation systems	Wang et al. (2024e)	Iterative refinement with multimodal prompts in a human-in-the-loop loop	Interfaces and iteration policies are key to practical controllability.
3D Visual Understanding			
3D instruction/prompt tuning	Li et al. (2024)	3D instruction–response data; efficient tuning without heavy alignment	Prompt/tuning design is central for 3D-language grounding at scale.
Order-aware 3D referring	Wu et al. (2024e)	Use LLM-inferred object order to guide progressive feature refinement	Procedural (ordered) prompts can stabilize 3D grounding.
Pretraining + viewpoint/axes prompts	Cho et al. (2024); Zhang et al. (2024e); Liu et al. (2023a)	Use BEV/viewpoint selection and explicit 3D axes/coordinates as prompts	View/coordinate prompts externalize spatial priors for reliable 3D reasoning.

11.2 Hallucination

The fine-grained visual contexts provided by visual prompting are also useful for multimodal hallucination mitigation. To address the issue that MLLMs' textual outputs are often not grounded in the reference image, Favero et al. (2024) propose a mutual-information decoding strategy that amplifies the influence of visual prompts on model generation. To reduce object hallucination and enhance fine-grained understanding in object-oriented perception tasks, Jiang et al. (2024) develop a prompting strategy jointly utilizing visual and textual prompts: a specialized detection model highlights relevant visual objects based on key concepts extracted from the textual prompt. While previous works mostly focus on single-object hallucination, Chen et al. (2024c) utilize visual referring prompts to evaluate multi-object hallucination of MLLMs and find that MLLMs hallucinate more when asked to focus on multiple objects simultaneously, suggesting that probing objects individually via visual prompts improves performance.

11.3 Debiasing

MLLMs are known to exploit spurious correlations between input and target variables, which can manifest as social biases on topics such as gender and race (Ye et al., 2024). As visual prompting enables fine-

grained understanding of visual objects and their relationships, it offers a plausible direction for mitigating these biases by grounding outputs in essential visual information rather than non-essential confounders. Visual prompts may also strengthen the model's causal understanding – both within a modality and across modalities – supporting more robust and grounded responses.

11.4 Visual Generation

Text-to-image diffusion models (Rombach et al., 2022) can be considered MLLMs in a broad sense, and visual prompting plays a central role in controlling their generation. ControlNet (Zhang et al., 2023c) and T2I-Adapter (Mou et al., 2024) take various visual prompts for spatial control in image generation. Our scope here is visual prompting specifically, not controllable generation in general (Cao et al., 2024). Prompt Diffusion (Wang et al., 2023) proposes a diffusion-based generative model that takes vision-language prompts and outputs target images, unlocking in-context generation after fine-tuning on six visual tasks. InstructBrush (Yang et al., 2023b) achieves adaptive image manipulation under the instruction of a pair of exemplar demonstrations, addressing language ambiguity in image editing. MPerceiver (Ai et al., 2024) introduces a multimodal prompt-learning approach that leverages generative priors of diffusion models for all-in-one image restoration. Chen et al. (2024d) propose VP3D, which leverages rich knowledge in 2D visual prompts to improve text-to-3D generation quality and supports stylized text-to-3D generation. PromptCharm (Wang et al., 2024e) provides an interaction system for text-to-image creation via multimodal prompting and image refinement, illustrating the necessity of visual prompting for better image creation.

11.5 3D Visual Understanding

Recent works use visual prompting for better 3D understanding. Li et al. (2024) construct an extensive dataset of instruction-response pairs for 3D scenes and introduce 3DMIT for efficient prompt tuning while eliminating the alignment stage between 3D scenes and language. DOrA (Wu et al., 2024e) proposes a 3D visual grounding framework with Order-Aware Referring, leveraging an LLM to infer an ordered object series that guides progressive feature refinement. Cho et al. (2024) construct the LV3D dataset and introduce Cube-LLM, an MLLM pre-trained on it. Zhang et al. (2024e) propose Agent3D-Zero, which introduces visual prompts based on bird's-eye-view images and viewpoint selection to unlock the MLLM's ability to observe 3D scenes. 3DAP (Liu et al., 2023a) develops a visual-prompting method that creates a 3D coordinate system with additional annotations to enable GPT-4V to complete 3D spatial tasks.

12 Conclusion

We presented the first comprehensive survey of visual prompting in multimodal large language models. We organized the field along four orthogonal axes – representation granularity, generation mechanism, adaptation paradigm, and inference and reasoning (Figure 3) – and formalized this structure via a unified visual-prompt operator family π (Definition 2.2), an alignment functional family (Definition 3.2), and an adaptation-paradigm signature (Section 3.5). Within this framework, we surveyed how visual prompts are categorized (Section 4), generated (Section 5), exploited for perception (Section 8) and compositional reasoning (Section 9), and learned through in-context, supervised, or reinforcement procedures (Sections 6 and 7). The empirical comparison and root-cause analysis of Section 10 surfaced consistent trade-offs between manual precision and learned scalability; the frontier-applications survey of Section 11 mapped jailbreaking, hallucination, debiasing, controllable generation, and 3D understanding onto the same four-axis taxonomy.

Open directions follow naturally from the axes: *(i)* unifying π specializations into a single trainable operator that exposes Axis 1 as a continuous interpolation rather than four disjoint subspaces; *(ii)* closing the gap between g_{fix} and g_{learn} via tool-augmented, differentiable generators that retain interpretability while scaling; *(iii)* grounding-aware reward design for Axis 3, particularly for self-alignment in the absence of human preferences; and *(iv)* faithfulness-preserving visual imagination for compositional reasoning, where the model itself generates the next $\theta_{\pi,t}$. Progress on these directions will determine whether visual prompting matures from a collection of effective tricks into a principled interface between language models and the visual world.

References

Yuang Ai, Huaibo Huang, Xiaoqiang Zhou, Jiexiang Wang, and Ran He. Multimodal prompt perceiver: Empower adaptiveness generalizability and fidelity for all-in-one image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25432–25444, 2024.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

Xiaoyi Bao, Chenwei Xie, Hao Tang, Tingyu Weng, Xiaofeng Wang, Yun Zheng, and Xingang Wang. Dyn-img: Key frames with visual prompts are good representation for multi-modal video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23678–23688, 2025.

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017, 2022.

Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G Shapiro, and Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3836–3845, 2025.

Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13154–13164, 2023.

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12914–12923, 2024.

Shaofei Cai, Zihao Wang, Kewei Lian, Zhancun Mu, Xiaojian Ma, Anji Liu, and Yitao Liang. Rocket-1: Mastering open-world interaction with visual-temporal context prompting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12122–12131, 2025.

Pu Cao, Feng Zhou, Qing Song, and Lu Yang. Controllable generation with text-to-image diffusion models: A survey. *arXiv preprint arXiv:2403.04279*, 2024.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL `https://arxiv.org/abs/2511.16719`.

Aochuan Chen, Yuguang Yao, Pin-Yu Chen, Yihua Zhang, and Sijia Liu. Understanding and improving visual prompting: A label-mapping perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19133–19143, 2023a.

Jiaqi Chen, Bingqian Lin, Xinmin Liu, Xiaodan Liang, and Kwan-Yee K Wong. Affordances-oriented planning using foundation models for continuous vision-language navigation. *arXiv preprint arXiv:2407.05890*, 2024a.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023b.

Wei Chen, Lin Li, Yongqi Yang, Bin Wen, Fan Yang, Tingting Gao, Yu Wu, and Long Chen. Comm: A coherent interleaved image-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2406.10462*, 2024b.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David F Fouhey, and Joyce Chai. Multi-object hallucination in vision-language models. *arXiv preprint arXiv:2407.06192*, 2024c.

Yang Chen, Yingwei Pan, Haibo Yang, Ting Yao, and Tao Mei. Vp3d: Unleashing 2d visual prompt for text-to-3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4896–4905, 2024d.

Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuang Gan. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, pp. 1254–1262, 2024e.

Jang Hyun Cho, Boris Ivanovic, Yulong Cao, Edward Schmerling, Yue Wang, Xinshuo Weng, Boyi Li, Yurong You, Philipp Krähenbühl, Yan Wang, et al. Language-image models with 3d understanding. *arXiv preprint arXiv:2405.03685*, 2024.

Ronghao Dang, Jiangyan Feng, Haodong Zhang, Chongjian Ge, Lin Song, Lijun Gong, Chengju Liu, Qijun Chen, Feng Zhu, Rui Zhao, et al. Instructdet: Diversifying referring object detection with generalized instructions. *arXiv preprint arXiv:2310.05136*, 2023.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9062–9072, 2025.

Zhizhao Duan, Hao Cheng, Duo Xu, Xi Wu, Xiangxie Zhang, Xi Ye, and Zhen Xie. Cityllava: Efficient fine-tuning for vlms in city scenario. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7180–7189, 2024.

Sheng Fan, Rui Liu, Wenguan Wang, and Yi Yang. Navigation instruction generation with bev perception and large language models. *arXiv preprint arXiv:2407.15087*, 2024a.

Yue Fan, Lei Ding, Ching-Chen Kuo, Shan Jiang, Yang Zhao, Xinze Guan, Jie Yang, Yi Zhang, and Xin Eric Wang. Read anywhere pointed: Layout-aware gui screen reading with tree-of-lens grounding. *arXiv preprint arXiv:2406.19263*, 2024b.

Yue Fan, Jing Gu, Kaiwen Zhou, Qianqi Yan, Shan Jiang, Ching-Chen Kuo, Xinze Guan, and Xin Eric Wang. Muffin or chihuahua? challenging large vision-language models with multipanel vqa. *arXiv preprint arXiv:2401.15847*, 2024c.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14303–14312, 2024.

Chaoyou Fu, Renrui Zhang, Haojia Lin, Zihan Wang, Timin Gao, Yongdong Luo, Yubo Huang, Zhengye Zhang, Longtian Qiu, Gaoxiang Ye, et al. A challenger to gpt-4v? early explorations of gemini in visual expertise. *arXiv preprint arXiv:2312.12436*, 2023.

Jun Gao, Qian Qiao, Ziqiang Cao, Zili Wang, and Wenjie Li. Aim: Let any multi-modal large language models embrace efficient in-context learning. *arXiv preprint arXiv:2406.07588*, 2024.

Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

Sitong Gong, Yunzhi Zhuge, Lu Zhang, Zongxin Yang, Pingping Zhang, and Huchuan Lu. The devil is in temporal token: High quality video reasoning segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 29183–29192. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02717. URL `https://openaccess.thecvf.com/content/CVPR2025/html/Gong_The_Devil_is_in_Temporal_Token_High_Quality_Video_Reasoning_CVPR_2025_paper.html`.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts. *arXiv preprint arXiv:2311.05608*, 2023.

Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023.

Xiangming Gu, Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Ye Wang, Jing Jiang, and Min Lin. Agent smith: A single image can jailbreak one million multimodal llm agents exponentially fast. *arXiv preprint arXiv:2402.08567*, 2024.

Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. Regiongpt: Towards region understanding vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13583–13592, 2024.

Xixuan Hao, Wei Chen, Yibo Yan, Siru Zhong, Kun Wang, Qingsong Wen, and Yuxuan Liang. Urbanvlp: A multi-granularity vision-language pre-trained foundation model for urban indicator prediction. *arXiv preprint arXiv:2403.16831*, 2024.

Junwen He, Yifan Wang, Lijun Wang, Huchuan Lu, Jun-Yan He, Jin-Peng Lan, Bin Luo, and Xuansong Xie. Multi-modal instruction tuned llms with fine-grained visual perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13980–13990, 2024.

Mir Rayat Imtiaz Hossain, Mennatullah Siam, Leonid Sigal, and James J Little. Visual prompting for generalized few-shot segmentation: A multi-scale approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23470–23480, 2024.

Siyuan Huang, Haonan Chang, Yuhan Liu, Yimeng Zhu, Hao Dong, Peng Gao, Abdeslam Boularias, and Hongsheng Li. A3vlm: Actionable articulation-aware vision language model. *arXiv preprint arXiv:2406.07549*, 2024a.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. Visual hallucinations of multi-modal large language models. *arXiv preprint arXiv:2402.14683*, 2024b.

Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv preprint arXiv:2409.01652*, 2024c.

Zhipeng Huang, Zhizheng Zhang, Zheng-Jun Zha, Yan Lu, and Baining Guo. Relationvlm: Making large vision-language models understand visual relations. *arXiv preprint arXiv:2403.12801*, 2024d.

Xiaofeng Ji, Faming Gong, Nuanlai Wang, Yanpu Zhao, Yuhui Ma, and Zhuang Shi. Pixel-level semantic parsing in complex industrial scenarios using large vision-language models. *Information Fusion*, 116: 102794, 2025.

Mengzhao Jia, Zhihan Zhang, Wenhao Yu, Fangkai Jiao, and Meng Jiang. Describe-then-reason: Improving multimodal mathematical reasoning through visual comprehension training. *arXiv preprint arXiv:2404.14604*, 2024.

Songtao Jiang, Yan Zhang, Chenyi Zhou, Yeying Jin, Yang Feng, Jian Wu, and Zuozhu Liu. Joint visual and text prompting for improved object-centric perception with multimodal large language models. *arXiv preprint arXiv:2404.04514*, 2024.

Shibo Jie, Yehui Tang, Ning Ding, Zhi-Hong Deng, Kai Han, and Yunhe Wang. Memory-space visual prompting for efficient vision-language fine-tuning. *arXiv preprint arXiv:2405.05615*, 2024.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9339–9350, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Bolin Lai, Felix Juefei-Xu, Miao Liu, Xiaoliang Dai, Nikhil Mehta, Chenguang Zhu, Zeyi Huang, James M Rehg, Sangmin Lee, Ning Zhang, et al. Unleashing in-context learning of autoregressive models for few-shot image manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18346–18357, 2025.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589, 2024.

Byung-Kwan Lee, Beomchan Park, Chae Won Kim, and Yong Man Ro. Collavo: Crayon large language and vision model. *arXiv preprint arXiv:2402.11248*, 2024.

Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. *arXiv preprint arXiv:2402.12058*, 2024a.

Yiming Lei, Jingqi Li, Zilong Li, Yuan Cao, and Hongming Shan. Prompt learning in computer vision: a survey. *Frontiers of Information Technology & Electronic Engineering*, 25(1):42–63, 2024b.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025a.

Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. In *The Twelfth International Conference on Learning Representations*, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023b.

Xiaoqi Li, Lingyun Xu, Mingxu Zhang, Jiaming Liu, Yan Shen, Iaroslav Ponomarenko, Jiahui Xu, Liang Heng, Siyuan Huang, Shanghang Zhang, et al. Crayonrobo: Object-centric prompt-driven vision-language-action model for robotic manipulation. *arXiv preprint arXiv:2505.02166*, 2025b.

Yanshu Li, Yi Cao, Hongyang He, Qisen Cheng, Xiang Fu, Xi Xiao, Tianyang Wang, and Ruixiang Tang. M²IV: Towards efficient and fine-grained multimodal in-context learning via representation engineering. In *Second Conference on Language Modeling*, 2025c. URL `https://openreview.net/forum?id=9ffYcEiNw9`.

Yinheng Li. A practical survey on zero-shot prompt design for in-context learning. *arXiv preprint arXiv:2309.13205*, 2023.

Zeju Li, Chao Zhang, Xiaoyan Wang, Ruilong Ren, Yifan Xu, Ruifei Ma, and Xiangde Liu. 3dmit: 3d multi-modal instruction tuning for scene understanding. *arXiv preprint arXiv:2401.03201*, 2024.

Zejun Li, Ruipu Luo, Jiwen Zhang, Minghui Qiu, Xuan-Jing Huang, and Zhongyu Wei. Vocot: Unleashing visually grounded multi-step reasoning in large multi-modal models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3769–3798, 2025d.

Zongjie Li, Chaozheng Wang, Chaowei Liu, Pingchuan Ma, Daoyuan Wu, Shuai Wang, and Cuiyun Gao. Vrptest: Evaluating visual referring prompting in large multimodal models. *arXiv preprint arXiv:2312.04087*, 2023c.

Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. Can feedback enhance semantic grounding in large vision-language models? *arXiv preprint arXiv:2404.06510*, 2024.

Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, et al. Mm-vid: Advancing video understanding with gpt-4v (ision). *arXiv preprint arXiv:2310.19773*, 2023.

Weifeng Lin, Xinyu Wei, Ruichuan An, Gao Peng, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. In *International Conference on Learning Representations*, volume 2025, pp. 46374–46403, 2025.

Yuanze Lin, Yunsheng Li, Dongdong Chen, Weijian Xu, Ronald Clark, Philip Torr, and Lu Yuan. Rethinking visual prompting for multimodal large language models with external knowledge. *arXiv preprint arXiv:2407.04681*, 2024.

Dingning Liu, Xiaomeng Dong, Renrui Zhang, Xu Luo, Peng Gao, Xiaoshui Huang, Yongshun Gong, and Zhihui Wang. 3daxiesprompts: Unleashing the 3d spatial task capabilities of gpt-4v. *arXiv preprint arXiv:2312.09738*, 2023a.

Fangchen Liu, Kuan Fang, Pieter Abbeel, and Sergey Levine. Moka: Open-vocabulary robotic manipulation through mark-based visual prompting. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024c.

Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19434–19445, 2023b.

Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. Safety of multimodal large language models on images and text. *arXiv preprint arXiv:2402.00357*, 2024d.

Yi Liu, Chengjun Cai, Xiaoli Zhang, Xingliang Yuan, and Cong Wang. Arondight: Red teaming large vision language models with auto-generated multi-modal jailbreak prompts. *arXiv preprint arXiv:2407.15050*, 2024e.

Zixian Liu, Mingtong Zhang, and Yunzhu Li. Kuda: Keypoints to unify dynamics learning and visual prompting for open-vocabulary robotic manipulation. *arXiv preprint arXiv:2503.10546*, 2025.

Bozhi Luan, Hao Feng, Hong Chen, Yonghui Wang, Wengang Zhou, and Houqiang Li. Textcot: Zoom in for enhanced multimodal text-rich image understanding. *arXiv preprint arXiv:2404.09797*, 2024.

Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision (ECCV)*, 2024a.

Huan Ma, Yan Zhu, Changqing Zhang, Peilin Zhao, Baoyuan Wu, Long-Kai Huang, Qinghua Hu, and Bingzhe Wu. Invariant test-time adaptation for vision-language model generalization. *arXiv preprint arXiv:2403.00376*, 2024b.

Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. Argus: Vision-centric reasoning with grounded chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14268–14280, 2025.

Sachit Menon, Richard Zemel, and Carl Vondrick. Whiteboard-of-thought: Thinking step-by-step across modalities. *arXiv preprint arXiv:2406.14562*, 2024.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14420–14431, 2024.

Alireza Moazeni, Shichong Peng, and Ke Li. Intrinsic papr for point-level 3d scene albedo and shading editing, 2024. URL `https://arxiv.org/abs/2407.00500`.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, pp. 4296–4304, 2024.

Zhixiong Nan, Xianghong Li, Jifeng Dai, and Tao Xiang. MI-DETR: an object detection model with multi-time inquiries mechanism. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 4703–4712. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00443. URL `https://openaccess.thecvf.com/content/CVPR2025/html/Nan_MI-DETR_An_Object_Detection_Model_with_Multi-time_Inquiries_Mechanism_CVPR_2025_paper.html`.

Soroush Nasiriany, Fei Xia, Wenhao Yu, Ted Xiao, Jacky Liang, Ishita Dasgupta, Annie Xie, Danny Driess, Ayzaan Wahid, Zhuo Xu, et al. Pivot: Iterative visual prompting elicits actionable knowledge for vlms. *arXiv preprint arXiv:2402.07872*, 2024.

Minheng Ni, Yeli Shen, Lei Zhang, and Wangmeng Zuo. Responsible visual editing. *arXiv preprint arXiv:2404.05580*, 2024.

Changdae Oh, Hyeji Hwang, Hee-young Lee, YongTaek Lim, Geunyoung Jung, Jiyoung Jung, Hosik Choi, and Kyungwoo Song. Blackvip: Black-box visual prompting for robust transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24224–24235, 2023.

Kaihang Pan, Siliang Tang, Juncheng Li, Zhaoyu Fan, Wei Chow, Shuicheng Yan, Tat-Seng Chua, Yueting Zhuang, and Hanwang Zhang. Auto-encoding morph-tokens for multimodal llm. *arXiv preprint arXiv:2405.01926*, 2024.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. Strengthening multimodal large language model with bootstrapped preference optimization. In *European Conference on Computer Vision (ECCV)*, 2024.

Leigang Qu, Haochuan Li, Tan Wang, Wenjie Wang, Yongqi Li, Liqiang Nie, and Tat-Seng Chua. Unified text-to-image generation and retrieval. *arXiv preprint arXiv:2406.05814*, 2024.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13009–13018, 2024.

Huali Ren, Anli Yan, Chong-zhi Gao, Hongyang Yan, Zhenxin Zhang, and Jin Li. Are you copying my prompt? protecting the copyright of vision prompt for vpaas via watermark. *arXiv preprint arXiv:2405.15161*, 2024a.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26374–26383, 2024b.

Razieh Rezaei, Masoud Jalili Sabet, Jindong Gu, Daniel Rueckert, Philip Torr, and Ashkan Khakzar. Learning visual prompts for guiding the attention of vision transformers. *arXiv preprint arXiv:2406.03303*, 2024.

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

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, et al. The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*, 2024.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.

Dianmo Sheng, Dongdong Chen, Zhentao Tan, Qiankun Liu, Qi Chu, Jianmin Bao, Tao Gong, Bin Liu, Shengwei Xu, and Nenghai Yu. Towards more unified in-context visual understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13362–13372, 2024.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11987–11997, 2023.

Minghong Sun, Lingye Zhao, and Luojun Lin. Probabilistic visual prompt tuning. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 432–445. Springer, 2025.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14398–14409, 2024a.

Wenfang Sun, Hao Chen, Yingjun Du, Yefeng Zheng, and Cees G. M. Snoek. Regionreasoner: Region-grounded multi-round visual reasoning, 2026. URL `https://arxiv.org/abs/2602.03733`.

Zhiqing Sun, Sheng Shen, Qinhong Zhou, Hongxin Zhang, Zengvhian Chen, David Cox, Yiming Yang, and Chuang Gan. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics (ACL Findings)*, 2024b.

Wei Tang, Yanpeng Sun, Qinying Gu, and Zechao Li. Visual position prompt for mllm based visual grounding, 2025. URL `https://arxiv.org/abs/2503.15426`.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and mitigating position bias of multi-image vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10599–10609, 2025.

Georgios Tziafas and Hamidreza Kasaei. Towards open-world grasping with large vision-language models. *arXiv preprint arXiv:2406.18722*, 2024.

David Wan, Jaemin Cho, Elias Stengel-Eskin, and Mohit Bansal. Contrastive region guidance: Improving grounding in vision-language models without training. *arXiv preprint arXiv:2403.02325*, 2024.

Lei Wang, Wanyu Xu, Zhiqiang Hu, Yihuai Lan, Shan Dong, Hao Wang, Roy Ka-Wei Lee, and Ee-Peng Lim. All in an aggregated image for in-image learning. *arXiv preprint arXiv:2402.17971*, 2024a.

Pei Wang, Zhaowei Cai, Hao Yang, Ashwin Swaminathan, R. Manmatha, and Stefano Soatto. Scaling up image segmentation across data and tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 4573–4583. Computer Vision Foundation / IEEE, 2025a. doi: 10.1109/CVPR52734.2025.00431. URL `https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Scaling_up_Image_Segmentation_across_Data_and_Tasks_CVPR_2025_paper.html`.

Weiyun Wang, Yiming Ren, Haowen Luo, Zhou Li, Zhe Chen, Zhenhang Huang, Zhiguo Cao, Wenhai Wang, Xizhou Zhu, Lewei Lu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. In *European Conference on Computer Vision (ECCV)*, 2024b.

Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024c.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36, 2024d.

Yifan Wang, Shiyu Li, Peiming Li, Xiaochen Yang, Yang Tang, and Zheng Wei. Render-of-thought: Rendering textual chain-of-thought as images for visual latent reasoning, 2026. URL `https://arxiv.org/abs/2601.14750`.

Yikun Wang, Siyin Wang, Qinyuan Cheng, Zhaoye Fei, Liang Ding, Qipeng Guo, Dacheng Tao, and Xipeng Qiu. Visuothink: Empowering lvlm reasoning with multimodal tree search. *arXiv preprint arXiv:2504.09130*, 2025b.

Zhendong Wang, Yifan Jiang, Yadong Lu, Pengcheng He, Weizhu Chen, Zhangyang Wang, Mingyuan Zhou, et al. In-context learning unlocked for diffusion models. *Advances in Neural Information Processing Systems*, 36:8542–8562, 2023.

Zhijie Wang, Yuheng Huang, Da Song, Lei Ma, and Tianyi Zhang. Promptcharm: Text-to-image generation through multi-modal prompting and refinement. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–21, 2024e.

Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Wenhai Wang, Zhe Chen, Xizhou Zhu, Lewei Lu, Tong Lu, et al. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *arXiv preprint arXiv:2406.08394*, 2024a.

Junda Wu, Rui Wang, Handong Zhao, Ruiyi Zhang, Chaochao Lu, Shuai Li, and Ricardo Henao. Few-shot composition learning for image retrieval with prompt tuning. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, pp. 4729–4737, 2023.

Junda Wu, Xintong Li, Tong Yu, Yu Wang, Xiang Chen, Jiuxiang Gu, Lina Yao, Jingbo Shang, and Julian McAuley. Commit: Coordinated instruction tuning for multimodal large language models. *arXiv preprint arXiv:2407.20454*, 2024b.

Junfeng Wu, Yi Jiang, Qihao Liu, Zehuan Yuan, Xiang Bai, and Song Bai. General object foundation model for images and videos at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3783–3795, 2024c.

Mingrui Wu, Xinyue Cai, Jiayi Ji, Jiale Li, Oucheng Huang, Gen Luo, Hao Fei, Xiaoshuai Sun, and Rongrong Ji. Controlmllm: Training-free visual prompt learning for multimodal large language models. *arXiv preprint arXiv:2407.21534*, 2024d.

Tung-Yu Wu, Sheng-Yu Huang, and Yu-Chiang Frank Wang. Dora: 3d visual grounding with order-aware referring. *arXiv preprint arXiv:2403.16539*, 2024e.

Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind's eye of llms: visualization-of-thought elicits spatial reasoning in large language models. *Advances in Neural Information Processing Systems*, 37:90277–90317, 2024f.

Yixuan Wu, Yizhou Wang, Shixiang Tang, Wenhao Wu, Tong He, Wanli Ouyang, Jian Wu, and Philip Torr. Dettoolchain: A new prompting paradigm to unleash detection ability of mllm. *arXiv preprint arXiv:2403.12488*, 2024g.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Junyu Xie, Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. Autoad-zero: A training-free framework for zero-shot audio description. *arXiv preprint arXiv:2407.15850*, 2024a.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024b.

Chengming Xu, Chen Liu, Yikai Wang, and Yanwei Fu. Towards global optimal visual in-context learning prompt selection. *arXiv preprint arXiv:2405.15279*, 2024a.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2087–2098, 2025a.

Mingjie Xu, Jinpeng Chen, Yuzhi Zhao, Jason Chun Lok Li, Yue Qiu, Zekang Du, Mengyang Wu, Pingping Zhang, Kun Li, Hongzheng Yang, Wenao Ma, Jiaheng Wei, Qinbin Li, Kangcheng Liu, and Wenqiang Lei. Vp-bench: A comprehensive benchmark for visual prompting in multimodal large language models, 2025b. URL `https://arxiv.org/abs/2511.11438`.

Peng Xu, Shengwu Xiong, Jiajun Zhang, Yaxiong Chen, Bowen Zhou, Chen Change Loy, David A. Clifton, Kyoung Mu Lee, Luc Van Gool, Ruiming He, Ruilin Yao, Xinwei Long, Jirui Huang, Kai Tian, Sa Yang, Yihua Shao, Jin Feng, Yue Zhong, Jiakai Zhou, Cheng Tang, Tianyu Zou, Yifang Zhang, Junming Liang, Guoyou Li, Zhaoxiang Wang, Qiang Zhou, Yichen Zhao, Shili Xiong, Hyeongjin Nam, Jaerin Lee, Jaeyoung Chung, JoonKyu Park, Junghun Oh, Kanggeon Lee, Wooseok Lee, Juneyoung Ro, Turghun Osman, Can Hu, Chaoyang Liao, Cheng Chen, Chengcheng Han, Chenhao Qiu, Chong Peng, Cong Xu, Dailin Li, Feiyu Wang, Feng Gao, Guibo Zhu, Guopeng Tang, Haibo Lu, Han Fang, Han Qi, Hanxiao Wu, Haobo Cheng, Hongbo Sun, Hongyao Chen, Huayong Hu, Hui Li, Jiaheng Ma, Jiang Yu, Jianing Wang, Jie Yang, Jing He, Jinglin Zhou, Jingxuan Li, Josef Kittler, Lihao Zheng, Linnan Zhao, Mengxi Jia, Muyang Yan, Nguyen Thanh Thien, Pu Luo, Qi Li, Shien Song, Shijie Dong, Shuai Shao, Shutao Li, Taofeng Xue, Tianyang Xu, Tianyi Gao, Tingting Li, Wei Zhang, Weiyang Su, Xiaodong Dong, Xiao-Jun Wu, Xiaopeng Zhou, Xin Chen, Xin Wei, Xinyi You, Xudong Kang, Xujie Zhou, Xusheng Liu, Yanan Wang, Yanbin Huang, Yang Liu, Yang Yang, Yanglin Deng, Yashu Kang, Ye Yuan, Yi Wen, Yicen Tian, Yilin Tao, Yin Tang, Yipeng Lin, Yiqing Wang, Yiting Xi, Yongkang Yu, Yumei Li, Yuxin Qin, Yuying Chen, Yuzhe Cen, Zhaofan Zou, Zhaohong Liu, Zhehao Shen, Zhenglin Du, Zhengyang Li, Zhenni Huang, Zhenwei Shao, Zhilong Song, Zhiyong Feng, Zhiyu Wang, Zhou Yu, Ziang Li, Zihan Zhai, Zijian Zhang, Ziyang Peng, Ziyun Xiao, and Zongshu Li. Mars2 2025 challenge on multimodal reasoning: Datasets, methods, results, discussion, and outlook, 2025c. URL `https://arxiv.org/abs/2509.14142`.

Xin Xu, Yue Liu, Panupong Pasupat, Mehran Kazemi, et al. In-context learning with retrieved demonstrations for language models: A survey. *arXiv preprint arXiv:2401.11624*, 2024b.

Yunqiu Xu, Linchao Zhu, and Yi Yang. Mc-bench: A benchmark for multi-context visual grounding in the era of mllms. In *International Conference on Computer Vision (ICCV)*, 2025d.

Shiyu Xuan, Qingpei Guo, Ming Yang, and Shiliang Zhang. Pink: Unveiling the power of referential comprehension for multi-modal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13838–13848, 2024.

An Yan, Zhengyuan Yang, Junda Wu, Wanrong Zhu, Jianwei Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Julian McAuley, Jianfeng Gao, et al. List items one by one: A new data source and learning paradigm for multimodal llms. *arXiv preprint arXiv:2404.16375*, 2024a.

Qianqi Yan, Yue Fan, Hongquan Li, Shan Jiang, Yang Zhao, Xinze Guan, Ching-Chen Kuo, and Xin Eric Wang. Multimodal inconsistency reasoning (mmir): A new benchmark for multimodal reasoning models. In *Annual Meeting of the Association for Computational Linguistics*, 2025. URL `https://api.semanticscholar.org/CorpusID:276576222`.

Siming Yan, Yu Yang, Xiaojian Ma, Haoran Zhang, Xinyu Chen, Yichun Zhang, Peiyuan Zhang, Yanyuan Zhang, Yichen Gong, and Qiang Zhang. Vigor: Improving visual grounding of large vision language models with fine-grained reward modeling. In *European Conference on Computer Vision (ECCV)*, 2024b.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023a.

Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, et al. Magma: A foundation model for multimodal ai agents. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14203–14214, 2025a.

Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting. *Advances in Neural Information Processing Systems*, 36, 2024.

Sihan Yang, Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. Opa-dpo: On-policy alignment for direct preference optimization in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Yifan Yang, Houwen Peng, Yifei Shen, Yuqing Yang, Han Hu, Lili Qiu, Hideki Koike, et al. Imagebrush: Learning visual in-context instructions for exemplar-based image manipulation. *Advances in Neural Information Processing Systems*, 36:48723–48743, 2023b.

Wenqian Ye, Guangtao Zheng, Yunsheng Ma, Xu Cao, Bolin Lai, James M Rehg, and Aidong Zhang. Mm-spubench: Towards better understanding of spurious biases in multimodal llms. *arXiv preprint arXiv:2406.17126*, 2024.

Zonghao Ying, Aishan Liu, Tianyuan Zhang, Zhengmin Yu, Siyuan Liang, Xianglong Liu, and Dacheng Tao. Jailbreak vision language models via bi-modal adversarial prompt. *arXiv preprint arXiv:2406.04031*, 2024.

Jaehong Yoon, Shoubin Yu, and Mohit Bansal. Raccoon: Remove, add, and change video content with auto-generated narratives. *arXiv preprint arXiv:2405.18406*, 2024.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=2msbbX3ydD`.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yunjie Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13807–13816, 2024.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721*, 2024a.

Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28202–28211, 2024b.

Sheng Yue, Jiani Liu, Xingyuan Hua, Ju Ren, Sen Lin, Junshan Zhang, and Yaoxue Zhang. How to leverage diverse demonstrations in offline imitation learning, 2024. URL `https://arxiv.org/abs/2405.17476`.

Fanhu Zeng, Fei Zhu, Haiyang Guo, Xu-Yao Zhang, and Cheng-Lin Liu. ModalPrompt: Towards efficient multimodal continual instruction tuning with dual-modality guided prompt. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12126–12141, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.609. URL `https://aclanthology.org/2025.emnlp-main.609/`.

Yufei Zhan, Shurong Zheng, Yousong Zhu, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon v2: Advancing multimodal perception with high-resolution scaling and visual-language co-referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. Vpgtrans: Transfer visual prompt generator across llms. *Advances in Neural Information Processing Systems*, 36, 2024a.

Chang-Bin Zhang, Jinhong Ni, Yujie Zhong, and Kai Han. v-clr: View-consistent learning for open-world instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 20307–20317. Computer Vision Foundation / IEEE, 2025a. doi: 10.1109/CVPR52734.2025.01891. URL `https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_v-CLR_View-Consistent_Learning_for_Open-World_Instance_Segmentation_CVPR_2025_paper.html`.

Chaoning Zhang, Fachrina Dewi Puspitasari, Sheng Zheng, Chenghao Li, Yu Qiao, Taegoo Kang, Xinru Shan, Chenshuang Zhang, Caiyan Qin, Francois Rameau, et al. A survey on segment anything model (sam): Vision foundation model meets prompt engineering. *arXiv preprint arXiv:2306.06211*, 2023a.

Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1020–1031, 2023b.

Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang, Chunyuan Li, et al. Llava-grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024b.

Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. In *First Conference on Language Modeling*.

Huanyu Zhang, Wenshan Wu, Chengzu Li, Ning Shang, Yan Xia, Yangyu Huang, Yifan Zhang, Li Dong, Zhang Zhang, Liang Wang, et al. Latent sketchpad: Sketching visual thoughts to elicit multimodal reasoning in mllms. *arXiv preprint arXiv:2510.24514*, 2025b.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024c.

Lu Zhang, Tiancheng Zhao, Heting Ying, Yibo Ma, and Kyusong Lee. Omagent: A multi-modal agent framework for complex video understanding with task divide-and-conquer. *arXiv preprint arXiv:2406.16620*, 2024d.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023c.

Sha Zhang, Di Huang, Jiajun Deng, Shixiang Tang, Wanli Ouyang, Tong He, and Yanyong Zhang. Agent3d-zero: An agent for zero-shot 3d understanding. *arXiv preprint arXiv:2403.11835*, 2024e.

Shilong Zhang, Peize Pei, SY Yi, Kaicheng Xu, Yida Wang, Ding Wang, Haibo Duan, Xinggang Wang, et al. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023d.

Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *arXiv preprint arXiv:2406.19389*, 2024f.

Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. Earthmarker: A visual prompt learning framework for region-level and point-level remote sensing imagery comprehension. *arXiv preprint arXiv:2407.13596*, 2024g.

Yichi Zhang, Yinpeng Dong, Siyuan Zhang, Tianzan Min, Hang Su, and Jun Zhu. Exploring the transferability of visual prompting for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26562–26572, 2024h.

Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. Groundhog: Grounding large language models to holistic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14227–14238, 2024i.

Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multimodal model. *arXiv preprint arXiv:2403.14598*, 2024j.

Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023a.

Kesen Zhao, Beier Zhu, Qianru Sun, and Hanwang Zhang. Unsupervised visual chain-of-thought reasoning via preference optimization. *arXiv preprint arXiv:2504.18397*, 2025a.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1702–1713, 2025b.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023b.

Wenliang Zhong, Wenyi Wu, Qi Li, Rob Barton, Boxin Du, Shioulin Sam, Karim Bouyarmane, Ismail Tutar, and Junzhou Huang. Enhancing multimodal large language models with multi-instance visual prompt generator for visual representation enrichment. *arXiv preprint arXiv:2406.02987*, 2024.

Enshen Zhou, Yiran Qin, Zhenfei Yin, Yuzhou Huang, Ruimao Zhang, Lu Sheng, Yu Qiao, and Jing Shao. Minedreamer: Learning to follow instructions via chain-of-imagination for simulated-world control. *arXiv preprint arXiv:2403.12037*, 2024a.

Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. Image-of-thought prompting for visual reasoning refinement in multimodal large language models. *arXiv preprint arXiv:2405.13872*, 2024b.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024c.

Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. Calibrated self-rewarding vision language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024d.

Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. Visual in-context learning for large vision-language models. In *Annual Meeting of the Association for Computational Linguistics*, 2024e. URL https://api.semanticscholar.org/CorpusID:267750174.

Lianghui Zhu, Bin Ouyang, Yuxuan Zhang, Tianheng Cheng, Rui Hu, Haocheng Shen, Longjin Ran, Xiaoxin Chen, Li Yu, Wenyu Liu, and Xinggang Wang. Lens: Learning to segment anything with unified reinforced reasoning, 2025a. URL https://arxiv.org/abs/2508.14153.

Yan Zhu, Huan Ma, and Changqing Zhang. Exploring task-level optimal prompts for visual in-context learning. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, pp. 11031–11039, 2025b.