SpecEyes: Accelerating Agentic Multimodal LLMs via Speculative Perception and Planning

Haoyu Huang[1*], Jinfa Huang[2*], Zhongwei Wan[3], Xiawu Zheng[1], Rongrong Ji[1†], and Jiebo Luo[2†]

[1] Xiamen University, Xiamen, Fujian, China
[2] University of Rochester, Rochester, NY, USA
[3] The Ohio State University, Columbus, OH, USA
{huanghaoyu@stu.,rrji@}xmu.edu.cn, {jhuang90@ur,jluo@cs}.rochester.edu

Code: https://github.com/MAC-AutoML/SpecEyes

Abstract. Agentic multimodal large language models (MLLMs) (*e.g.*, OpenAI o3 [41] and Gemini Agentic Vision [10]) achieve remarkable reasoning capabilities through the iterative invocation of visual tools. However, the cascaded perception, reasoning, and tool-calling loops introduce significant sequential overhead. This overhead, termed agentic depth, incurs prohibitive latency and seriously limits system-level concurrency. To this end, we propose SpecEyes, an *agentic-level* speculative acceleration framework that breaks this sequential bottleneck. Our key insight is that a lightweight MLLM can plan a tool-free execution path to directly answer many queries, bypassing the expensive tool-use loop. To regulate this speculative planning, we introduce a cognitive gating mechanism based on answer separability, which quantifies the model's confidence in self-verification without requiring oracle labels. Furthermore, we design a heterogeneous parallel funnel that exploits the small model's stateless concurrency to mask the large model's stateful serial execution, thereby maximizing system throughput. Extensive experiments on V* Bench, HR-Bench, and POPE demonstrate that SpecEyes achieves $\mathbf{1.1-3.35\times}$ speedup over the baseline while preserving or even improving accuracy, thereby boosting serving throughput under concurrent workloads.

Keywords: Agentic MLLM · Efficient Reasoning · Speculative Decoding

1 Introduction

Multimodal large language models (MLLMs) have undergone a paradigm shift, from static, single-pass visual perception to dynamic, *agentic* interaction with the visual world. Early MLLMs encode an image once and generate a response in a single forward pass, treating vision as a passive input channel. Recent breakthroughs [16,18,34,47,71,75] fundamentally alter this design: models actively invoke external perception tools (*e.g.* zoom-in, crop, OCR) during reasoning,

* Equal Contributors † Corresponding Authors

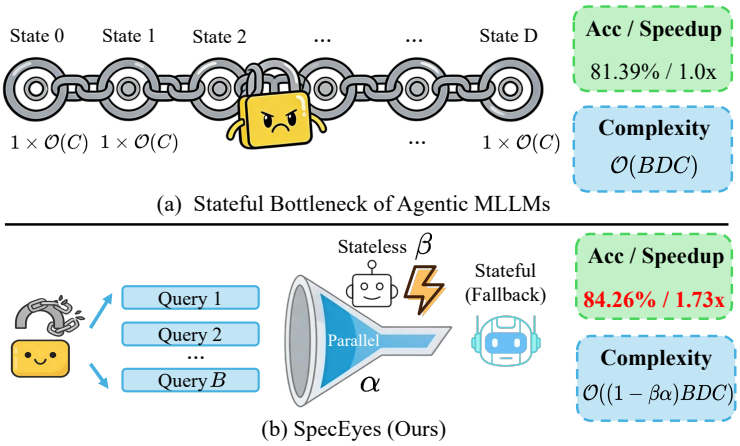


Fig. 1: Motivation and overview of SpecEyes. Top: Agentic MLLMs evaluate each query via a Markovian sequence of stateful tool invocations of depth D. This strict causal dependency prohibits parallelization, imposing a serving complexity of $\mathcal{O}(BDC)$ for B queries, where C denotes the tool per-step inference cost. **Bottom:** SpecEyes enables agentic-level speculative bypass with a stateless small model and an answer-separability gate. Here, β is the fraction of tool-free candidates after screening (Sec. 3.4) and α is the acceptance rate of speculative answers among them (Secs. 3.2 and 3.3), averaging 80% and 71% across all benchmarks, respectively. All reported accuracy and speedup values are averaged across V* [57], HR-Bench [55], and POPE [29].

forming iterative loops of perception, reasoning, and tool calling that progressively refine their understanding. This agentic paradigm has demonstrated remarkable capabilities on challenging visual tasks that require fine-grained inspection, multi-step compositional reasoning, and active information seeking [8, 25, 39, 40, 60, 67].

However, the mechanism that empowers agentic MLLMs simultaneously introduces a severe *efficiency crisis*. As shown in Fig. 1, each query triggers a cascade of tool-calling steps, a quantity we term the *agentic depth* D, in which each step depends on the observation from the previous step. This strict data dependency inflicts a **dual disaster** on system performance: **(i) Latency explosion**: the end-to-end response time for a single query grows linearly with D, since each reasoning-and-tool cycle must complete before the next can begin; **(ii) Concurrency collapse**: because each query's tool-use chain mutates a per-query state, batching efficiency is severely bottlenecked, the agentic model can only advance one step at a time per query, leaving massive hardware parallelism idle. Therefore, these effects render agentic MLLMs orders of magnitude slower than non-agentic counterparts, posing a fundamental barrier to real-world deployment.

Existing approaches to efficient reasoning fall short of addressing this bottleneck. Token-level speculative decoding [21, 42] accelerates individual generation steps by letting a small draft model propose tokens for a larger model to verify. However, these methods still operate *within* a fixed reasoning trajectory: the *agentic pipeline itself*, *i.e.*, the multi-turn loop of perception and reasoning, remains fully

serial and every tool must still be invoked in sequence. Moreover, the additional draft/verification interaction often expands the generated traces (longer token sequences and extra turns), introducing non-trivial overhead that can offset the per-step speedup in practice. Similarly, multimodal token pruning [11, 17, 28, 54] and temporal compression [13, 20] reduce per-step compute within a fixed model, yet they do not eliminate the repeated tool invocations that dominate agentic latency. In short, all prior methods operate *within* the agentic loop, none question whether the loop itself is necessary for every query.

In this paper, we make a conceptual leap: we lift the speculative paradigm from the token/semantic level to the **agentic level**. Our key observation is that a large fraction of queries directed at agentic MLLMs do *not* actually require deep tool-assisted reasoning. Instead, a lightweight, tool-free vision model can answer them correctly using only the original image, provided we can reliably identify which queries fall into this category. This motivates a heterogeneous "*think fast, think slow*" architecture: a small non-agentic model rapidly generates speculative answers via intuition (fast thinking), while the large agentic model is reserved for queries that genuinely require multi-step tool interaction (slow thinking).

We instantiate this idea by introducing SpecEyes, an *agentic-level* speculative acceleration framework for multimodal reasoning. It comprises three tightly integrated components: **(1) A four-phase speculative pipeline** (Sec. 3.2) that routes each query through heuristic tool-use judgment, small-model speculation, confidence-based switching, and agentic fallback. **(2) Cognitive gating** (Sec. 3.3) via a novel *answer separability* metric S_{sep} that measures the competitive margin among top-K logits, providing a calibration-free, scale-invariant decision boundary for trusting the small model's output. **(3) A heterogeneous parallel serving architecture** (Sec. 3.4) that runs the stateless small model concurrently and forwards only low-confidence queries to the agentic model, converting the speculative acceptance rate into multiplicative throughput gains. Extensive experiments on V* Bench, HR-Bench, and POPE show that SpecEyes preserves the full accuracy of the agentic pipeline while substantially reducing latency and improving throughput. Overall, the main contributions are as follows:

- We identify and formalize the *stateful bottleneck* of agentic MLLMs, showing that data dependency inherent in tool-use chains imposes a fundamental barrier to both per-query latency and system-level concurrency.
- We propose SpecEyes, the first framework that lifts speculative acceleration from the token level to the *agentic level*, bypassing entire tool-use loop for queries that do not require it while preserving full accuracy.
- We introduce *cognitive gating* based on answer separability among top-K logits, providing a label-free, scale-invariant criterion for the small model to decide when to trust its own output versus escalating to the agentic model.

– We design a *heterogeneous parallel funnel* that exploits the stateless nature of the small model to achieve concurrent query processing, yielding multiplicative throughput improvements proportional to the speculative acceptance rate.

2 Related Work

Agentic Multimodal Large Language Models. Agentic reasoning in language models originates from tool-augmented frameworks that interleave action generation with external feedback [44, 46, 68, 69]. Building on this, multimodal large language models (MLLMs) have adopted a similar agentic paradigm, enabling active interleaving of perception and reasoning through external visual tools rather than relying on passive single-pass encoding. Early large-scale MLLMs [1, 2, 9, 27, 48] established the backbone architectures upon which agentic extensions are built. DeepEyes [75] demonstrates that reinforcement learning can train models to call perception tools during reasoning; subsequent work enables executable reasoning via code generation and visual manipulation [16, 18, 19, 47, 49, 61, 71–73], and further scales agentic depth through multi-turn interaction, self-reflection, and reinforcement-learning-based agent optimization [8, 25, 33, 43, 67]. Despite their effectiveness, these methods rely on deeply sequential perception–reasoning tool loops, incurring substantial latency and limited concurrency, a system-level bottleneck that prior work largely overlooks.

Efficient Reasoning. Token-level speculative decoding [4, 5, 26, 30–32, 58, 59, 64, 65, 70] accelerates generation by having a small draft model propose tokens for a larger model to verify. Recent extensions apply this idea to collaborative reasoning: SpecReason [42] delegates simpler steps to a lightweight model verified via semantic consistency; RelayLLM [21] dynamically invokes a stronger expert at critical steps; ATTS [63] further explores asynchronous test-time scaling by dynamically allocating computation under uncertainty; DSP [15] speculatively drafts and verifies agent actions *within* text-only LLM-agent trajectories via online reinforcement learning, and SpecTemp [20] and Lin et al. [35, 36] reduce redundant visual processing in multimodal and interactive settings. Adaptive computation and early-exit methods [7, 12, 23, 38, 51, 76] further bypass layers for easier inputs. In contrast to methods that only optimize steps *within* a fixed trajectory, our SpecEyes speculatively bypasses the agentic loop entirely, breaking sequential bottlenecks to unlock parallel execution.

Efficient Multimodal Perception. A parallel line of work reduces the per-step computational burden of multimodal perception. Frequency-based compression truncates high-frequency visual signals [54]; token pruning retains visually salient tokens via attention scores or multimodal relevance [11, 28, 45, 62, 66]; and dynamic sparsification optimizes retention across layers [17]. Token merging [3, 22, 56] reduces sequence length by combining redundant representations, and temporal redundancy across frames is exploited to merge or prune spatial tokens in video settings [6, 13]. KV-cache compression [37, 52, 53] additionally reduces memory and decoding cost by evicting cached visual keys and values.

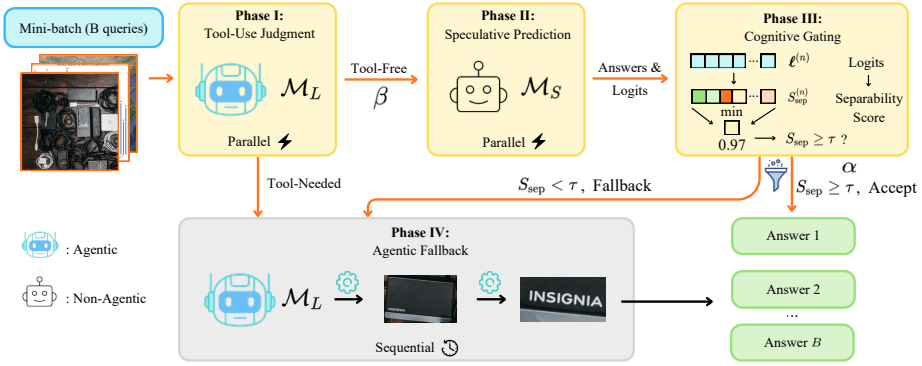


Fig. 2: Pipeline overview of SpecEyes. A batch of B queries passes through a four-phase funnel. **I:** $\mathcal{M}_L$ screens tool necessity, splitting queries into tool-free (g=0) and tool-required (g=1). **II:** A stateless $\mathcal{M}_S$ speculatively answers all tool-free queries with token-level logits. **III:** An answer separability score S_{sep} gates each answer; those above τ are accepted directly. **IV:** Remaining queries fall back to the full agentic loop. The funnel yields $\approx 1/(1-\beta\alpha)\times$ throughput speedup.

Despite these gains, all such methods operate within a monolithic model and leave the sequential agentic pipeline intact, as the large model must still execute the full perception–reasoning loop. In contrast, SpecEyes targets efficiency at the *agentic level*: rather than accelerating individual operations, it speculatively bypasses entire tool-use loops via a cognitively gated lightweight model, breaking sequential bottlenecks to enable high-throughput parallel execution.

3 Methodology

We begin by formalizing the stateful bottleneck inherent in agentic multimodal reasoning (Sec. 3.1), then introduce SpecEyes, our four-phase speculative acceleration framework (Sec. 3.2). We detail the cognitive gating mechanism that governs speculative bypass (Sec. 3.3), and finally describe the heterogeneous parallel architecture that maximizes system throughput (Sec. 3.4).

3.1 Modeling the Stateful Bottleneck of Agentic MLLMs

Preliminaries. We formalize an agentic multimodal large language model (MLLM) as a stateful reasoning system $\mathcal{A} = (\mathcal{S}, \mathcal{T}, \pi)$, where $\mathcal{S}$ denotes the state space, $\mathcal{T} = \{t_1, \ldots, t_N\}$ is a set of perception tools (*e.g.* `Zoom-in`, `Crop`), and π is policy that jointly selects tool invocations and generates reasoning tokens.

Given a query q and an input image I, the model maintains a state trajectory $\{s_0, s_1, \ldots, s_D\}$ over D reasoning steps. The initial state is $s_0 = (q, I)$. At each step d, the policy produces an action $a_d = \pi(s_d)$ that either invokes a tool $t \in \mathcal{T}$ or emits a final answer. When a tool is invoked, the state transitions as:

$$s_{d+1} = f(s_d, t_d(s_d)), \tag{1}$$

where $t_d(s_d)$ applies the selected tool t_d to the current visual context (*e.g.* cropping a region of interest from I) and f fuses the resulting observation into the next state. We refer to D as the *agentic depth* of the query.

State Dependency and Sequential Bottleneck. A critical property of Eq. (1) is that subsequent tool selections depend causally on prior observations. Concretely, let $t_{d+1} \sim \pi(\cdot \mid s_{d+1})$ be tool chosen at step $d{+}1$. Since s_{d+1} contains the output of t_d, Markov chain $(s_0, a_0, s_1, a_1, \ldots)$ forms a strict data dependency:

$$p(a_{d+1} \mid s_0, a_0, \ldots, s_d) = p(a_{d+1} \mid s_d, t_d(s_d)) \neq p(a_{d+1} \mid s_0). \tag{2}$$

This dependency renders the agentic pipeline inherently *sequential*: step $d{+}1$ cannot begin until step d completes. Consequently, the end-to-end latency for a single query scales linearly with agentic depth:

$$L_{\text{agent}}(q) = \sum_{d=0}^{D(q)} \Big(\underbrace{c_{\text{llm}}}_{\text{reasoning}} + \underbrace{c_{\text{tool}}(t_d)}_{\text{perception}} \Big), \tag{3}$$

where c_{llm} and $c_{\text{tool}}(t_d)$ denote the latency of LLM inference and tool execution at step d, respectively.

Throughput Implication. At the system level, this strict serialization limits concurrency even under continuous batching (*e.g.*, vLLM [24]). While LLM forward passes for different queries at different agentic steps can share a GPU batch, the stateful tool-use loop for each query still proceeds sequentially: step $d{+}1$ of query i cannot begin until step d completes. Consequently, within a batch of B queries, the batch-level wall-clock time is dominated by the *slowest* (deepest) trajectory, as queries with heavy-tailed agentic depth stall the entire batch. The effective throughput under batched serving is:

$$\Theta_{\text{agent}}^{\text{batched}} \approx \frac{B}{\max_{i \in [B]} L_{\text{agent}}(q_i) + c_{\text{sched}}}, \tag{4}$$

where c_{sched} is a small scheduling overhead. This bound tightens as agentic depth variance grows, since a single long-tail query forces the entire batch to wait. Speculatively converting a fraction $\beta\alpha$ of queries into stateless single-pass inferences directly shrinks the effective $\max_i L_{\text{agent}}(q_i)$, motivating our approach.

3.2 SpecEyes: Agentic-Level Speculative Reasoning

Our key insight is that not all queries require deep agentic reasoning. For a substantial fraction of inputs, a small *non-agentic* MLLM, denoted $\mathcal{M}_S$, can produce a correct answer *without any tool invocation*, directly from the original image I. SpecEyes exploits this observation through a four-phase pipeline (Fig. 2) that speculatively bypasses expensive tool chains whenever $\mathcal{M}_S$ is sufficiently confident, and falls back to the full agentic model $\mathcal{M}_L$ otherwise.

We denote the small non-agentic model as $\mathcal{M}_S$ and the large agentic MLLM as $\mathcal{M}_L = \mathcal{A}$. The four phases are detailed below.

Phase I: Heuristic Tool-Use Judgment. Given a query q and image I, the large agentic model $\mathcal{M}_L$ first determines whether tool invocation is necessary. We prompt $\mathcal{M}_L$ with a lightweight binary classification head:

$$g(q, I) = \mathcal{M}_L(q, I;\ \mathcal{P}_{\text{judge}}) \in \{0, 1\}, \tag{5}$$

where $\mathcal{P}_{\text{judge}}$ is a prompt instructing the model to assess tool necessity, $g = 0$ indicates that $\mathcal{M}_L$ judges the query to be answerable from the global image alone, and $g = 1$ indicates a potential need for tool-assisted perception. Queries with $g = 0$ proceed directly to Phase II; queries with $g = 1$ are immediately forwarded to Phase IV (agentic fallback). Although Phase I is executed by $\mathcal{M}_L$, it generates only a single binary token with no tool invocation, incurring negligible overhead. We use $\mathcal{M}_L$ rather than $\mathcal{M}_S$ because its tool-calling capability makes it a more reliable judge of tool necessity, yielding more accurate screening.

Phase II: Speculative Prediction. For queries passing Phase I (*i.e.*, $g = 0$), $\mathcal{M}_S$ directly generates an answer $\hat{y}_S$ along with the full output logit distribution:

$$\hat{y}_S,\ \{\boldsymbol{\ell}^{(n)}\}_{n=1}^{|\hat{y}_S|} = \mathcal{M}_S(q, I), \tag{6}$$

where $\boldsymbol{\ell}^{(n)} \in \mathbb{R}^{|\mathcal{V}|}$ is the logit vector over the vocabulary $\mathcal{V}$ for the n^{th} generated token. Crucially, this inference is *stateless*: it requires no tool execution and can be performed concurrently for all queries in the batch.

Phase III: Small MLLM Confidence Switching. The logits from Phase II are passed to a *cognitive gating* function S_{sep} (detailed in Sec. 3.3) that quantifies the answer confidence of $\mathcal{M}_S$ without requiring ground-truth labels. We compute a scalar separability score for the speculative answer $\hat{y}_S$:

$$\text{decision} = \begin{cases} \texttt{accept}\ \hat{y}_S, & \text{if } S_{\text{sep}}(\hat{y}_S) \geq \tau, \\ \texttt{fallback to}\ \mathcal{M}_L, & \text{if } S_{\text{sep}}(\hat{y}_S) < \tau, \end{cases} \tag{7}$$

where τ is a threshold selected from a coarse operating-point grid. Accepted answers are returned immediately, completely bypassing the agentic pipeline; rejected queries proceed to Phase IV.

Phase IV: Agentic Fallback. Queries that fail confidence switching are routed to the full agentic model $\mathcal{M}_L$, which executes the complete stateful perception-reasoning loop:

$$\hat{y}_L = \mathcal{M}_L(q, I) = \pi\big(s_0 \xrightarrow{t_0} s_1 \xrightarrow{t_1} \cdots \xrightarrow{t_{D-1}} s_D\big). \tag{8}$$

The agentic model retains full access to all tools $\mathcal{T}$ and performs multi-step reasoning at the cost of sequential latency $L_{\text{agent}}(q)$. By design, Phase IV serves as a *safety net*: routing low-confidence queries back to the full agentic pipeline

substantially mitigates potential accuracy loss, even if a marginal gap may remain due to the imperfect nature of the gating mechanism.

End-to-End Latency. Let $\beta \in [0,1]$ denote the tool-free screening ratio from Phase I and $\alpha \in [0,1]$ the cognitive gate acceptance rate from Phase III. All queries incur the judgment cost c_J, only the β fraction passing Phase I additionally incurs the small model cost c_S, the remaining $(1-\beta\alpha)$ fraction forwarded to M_L pays the full agentic cost L_{agent}. Therefore, the expected per-query latency is:

$$\mathbb{E}[L_{\text{SpecEyes}}] = c_J + \beta\, c_S + \big(1-\beta\alpha\big)\, L_{\text{agent}}, \tag{9}$$

where $c_J + \beta c_S \ll L_{\text{agent}}$. When $\beta\alpha$ is large (*e.g.* $\beta\alpha > 0.6$), the expected latency is dominated by the lightweight front-end cost, yielding substantial speedups over the purely agentic baseline.

3.3 Small MLLM Cognitive Gating via Answer Separability

The effectiveness of SpecEyes hinges critically on the quality of the confidence switching mechanism in Phase III. We now introduce the *answer separability* score S_{sep} that serves as the cognitive gate.

Limitations of Probability-Based Confidence. A common probability-based confidence for sequence generation aggregates per-token max-softmax probabilities via the geometric mean [74]. Concretely, for the n-th generated token with logits $\boldsymbol{\ell}^{(n)}$, we define the maximum softmax probability $p_{\max}^{(n)}$ as:

$$p_{\max}^{(n)} = \max_{v\in\mathcal{V}} \sigma(\boldsymbol{\ell}^{(n)})_v, \tag{10}$$

where $\sigma(\cdot)$ denotes the softmax operator and $\mathcal{V}$ is the vocabulary. The overall confidence is computed as:

$$S_{\log}(\hat{y}_S) = \exp\left(\frac{1}{|\hat{y}_S|}\sum_{n=1}^{|\hat{y}_S|}\log p_{\max}^{(n)}\right), \tag{11}$$

which corresponds to the geometric mean of $\{p_{\max}^{(n)}\}$. However, $S_{\log}$ remains unreliable for gating: (1) it inherits the well-known miscalibration of softmax, where large logit magnitudes can yield overconfident probabilities; (2) token-wise $p_{\max}^{(n)}$ can be spuriously high for low-entropy or nearly-deterministic positions (e.g., punctuation, formatting tokens), and the geometric aggregation does not explicitly measure how well the top prediction is separated from strong competitors. These issues increase the risk of false acceptance in our speculative bypass.

Answer Separability Score. Instead of relying on the raw softmax probability, we design a metric that measures the *decision margin* between the top prediction and its competitors. For the n^{th} generated token with logit vector $\boldsymbol{\ell}^{(n)}$, let

$\ell_{[1]}^{(n)} \geq \ell_{[2]}^{(n)} \geq \cdots \geq \ell_{[|\mathcal{V}|]}^{(n)}$ be the sorted logits in descending order. We define the *token-level separability* as:

$$S_{\text{sep}}^{(n)} = \frac{\ell_{[1]}^{(n)} - \mu_K^{(n)}}{\sigma_K^{(n)} + \epsilon}, \tag{12}$$

where $\mu_K^{(n)}$ and $\sigma_K^{(n)}$ are the mean and standard deviation of the top-K logits $\{\ell_{[1]}^{(n)}, \ldots, \ell_{[K]}^{(n)}\}$, and $\epsilon > 0$ is a small constant for numerical stability. Intuitively, $S_{\text{sep}}^{(n)}$ quantifies how far the leading logit stands apart from its nearest competitors: a large value indicates a clear decision boundary, while a small value signals ambiguity among top candidates.

Compared to softmax probability, $S_{\text{sep}}^{(n)}$ offers two key advantages: (i) it is *scale-invariant*, since both the numerator and denominator scale linearly with logit magnitude, neutralizing the calibration artifacts of softmax; (ii) it explicitly models the *competitive landscape* among top candidates via the variance term $\sigma_K^{(n)}$, providing a more informative confidence signal.

Token-to-Answer Aggregation. The token-level score $S_{\text{sep}}^{(n)}$ must be aggregated across all $|\hat{y}_S|$ generated tokens to obtain an answer-level confidence. We consider three natural aggregation strategies:

$$S_{\text{sep}}^{\text{mean}} = \frac{1}{|\hat{y}_S|} \sum_{n=1}^{|\hat{y}_S|} S_{\text{sep}}^{(n)}, \quad S_{\text{sep}}^{\text{min}} = \min_{n \in [|\hat{y}_S|]} S_{\text{sep}}^{(n)}, \quad S_{\text{sep}}^{\text{bottom-}r} = \frac{1}{|\mathcal{B}|} \sum_{n \in \mathcal{B}} S_{\text{sep}}^{(n)}, \tag{13}$$

where $\mathcal{B}$ is the index set of the bottom-r fraction of tokens with the smallest $S_{\text{sep}}^{(n)}$ values, *i.e.*, $|\mathcal{B}| = \lceil r\,|\hat{y}_S| \rceil$ for a ratio $r \in (0, 1)$ chosen empirically. The aggregated score is then normalized via a sigmoid function. We adopt min aggregation as the default strategy, based on the following risk-theoretic argument.

Proposition 1. *Let $\hat{y}_S = (y_1, \ldots, y_{|\hat{y}_S|})$ be the speculative answer. Define the answer-level error event $\mathcal{E} = \bigcup_n \mathcal{E}_n$, where $\mathcal{E}_n$ denotes the event that token y_n is incorrect. Then:*

$$P(\mathcal{E}) = P\left(\bigcup_n \mathcal{E}_n\right) \leq \sum_n P(\mathcal{E}_n). \tag{14}$$

Under the assumption that each $P(\mathcal{E}_n)$ is monotonically decreasing in $S_{sep}^{(n)}$, a plausible working assumption, thresholding on $\min_n S_{sep}^{(n)}$ ensures every token exceeds the confidence threshold, providing the tightest bound on $P(\mathcal{E})$. We therefore treat $S_{\text{sep}}^{\text{min}}$ as an *empirically motivated* default: it consistently achieves the highest matched-speed accuracy among all four aggregation variants across every benchmark . Intuitively, the min strategy acts as a *worst-case guard*: it triggers fallback whenever *any* token in the answer exhibits low separability. This is conservative by design, prioritizing precision (*i.e.*, avoiding false acceptances) to preserve the accuracy guarantee of the agentic pipeline.

3.4 Heterogeneous Parallelism for Throughput Acceleration

Beyond per-query latency reduction, SpecEyes enables system-level throughput gains by organizing the four phases into a heterogeneous parallel funnel.

Batch-Parallel Front-End. We serve requests in batches of size B. Let $\beta \in [0,1]$ be the fraction of queries that Phase I screens as tool-free (g=0) and $\alpha \in [0,1]$ be the acceptance rate of the cognitive gate among those candidates. Both screening (Phase I, latency c_J) and speculative inference (Phase II, latency c_S) are stateless single-turn forward passes and therefore fully batch-parallelizable, giving a parallel front-end cost of $c_J + c_S$.

Funnel-Shaped Serving. Accepted queries (α, β, B) are returned immediately; the remaining *residual set* $\mathcal{R}$, consisting of gating-rejected and tool-required queries, falls back to sequential agentic execution:

$$\begin{aligned}
&\underbrace{B}_{\text{batch}} \xrightarrow{\mathcal{M}_L \text{ screen (par.)}} \underbrace{\beta B}_{g=0} + \underbrace{(1-\beta)B}_{g=1} \\
&\underbrace{\beta B}_{g=0} \xrightarrow{\mathcal{M}_S \text{ speculate (par.)}} \underbrace{\alpha\beta B}_{\text{accept}} + \underbrace{(1-\alpha)\beta B}_{\text{reject}} \\
&\underbrace{(1-\beta)B + (1-\alpha)\beta B}_{\mathcal{R}} \xrightarrow{\mathcal{M}_L \text{ agentic (seq.)}} \underbrace{(1-\beta\alpha)B}_{\text{fallback}}.
\end{aligned} \tag{15}$$

Under continuous batching (*e.g.*, vLLM [24]), throughput scales approximately linearly with the number of queries served through the agentic path. SpecEyes converts a $\beta\alpha$ fraction of queries into stateless single-pass inferences that bypass the agentic loop entirely, reducing the effective agentic residual from B to $(1-\beta\alpha)B$. Since $c_J + c_S \ll \bar{L}_{\text{agent}}$, the per-batch cost is dominated by the agentic fallback on $|\mathcal{R}| = (1-\beta\alpha)B$ residual queries, yielding a throughput speedup is:

$$\Theta_{\text{SpecEyes}} \,/\, \Theta_{\text{agent}} \;\approx\; 1/(1-\beta\alpha), \tag{16}$$

jointly governed by the screening ratio β and the gate acceptance rate α. Importantly, continuous batching multiplies throughput proportionally for *both* the baseline and SpecEyes, so it does not alter the speedup ratio.

4 Experiment

4.1 Experiment Setups

Benchmarks and Baselines. We evaluate SpecEyes on three multimodal benchmarks spanning fine-grained perception, high-resolution understanding, and hallucination robustness. V* [57] provides two multiple-choice subsets: Direct Attributes (115 questions) for attribute recognition and Relative Position (76 questions) for spatial reasoning. HR-Bench [55] tests high-resolution perception

Table 1: Main results on V*, HR-Bench, and POPE datasets. Spd. = wall-clock speedup over each base model (green: faster; red: slower). Bold indicates the best accuracy within each group, and highlighted rows represent our recommended variants. SpecEyes (min) provides the best speedup-accuracy trade-off across both agentic backbones. *Backbone* (w/o tools) uses the same large model with tools disabled.

Method	V*				HR-Bench				POPE						Avg.	
	Attr.		Pos.		4K		8K		Adv.		Pop.		Rand.			
	Acc.	Spd.	Acc.	Spd.	Acc.	Spd.	Acc.	Spd.	Acc.	Spd.	Acc.	Spd.	Acc.	Spd.	Acc.	Spd.
Qwen3-VL-2B (draft only)	77.39	5.44×	82.89	5.31×	71.38	3.20×	68.00	2.90×	82.56	4.20×	83.80	3.78×	86.47	4.07×	78.93	4.13×
Based on DeepEyes [75]																
DeepEyes (w tools)	90.43	1.00×	82.89	1.00×	75.85	1.00×	71.43	1.00×	78.43	1.00×	81.90	1.00×	88.83	1.00×	81.39	1.00×
DeepEyes (w/o tools)	80.87	4.08×	73.68	4.18×	75.25	2.71×	72.00	2.53×	46.90	3.78×	49.33	3.60×	48.20	3.81×	63.75	3.53×
SpecReason [42]	80.19	0.61×	73.91	0.38×	80.43	0.44×	72.54	0.42×	49.10	0.38×	51.55	0.38×	60.20	0.37×	66.85	0.43×
▷ SpecEyes (log)	83.48	2.06×	88.16	2.05×	73.71	1.35×	69.67	1.28×	83.97	1.89×	86.70	1.95×	**90.50**	2.05×	82.31	1.80×
▷ SpecEyes (mean)	78.26	2.89×	84.21	3.35×	71.62	1.88×	67.38	1.77×	**85.13**	2.06×	**87.00**	2.10×	90.13	2.14×	80.53	2.31×
▷ SpecEyes (bottom-r)	83.48	2.13×	84.21	2.12×	75.22	1.20×	71.18	1.04×	**85.13**	2.08×	**87.00**	2.08×	90.13	2.11×	82.34	1.82×
▷ **SpecEyes (min)**	**90.43**	1.53×	**89.47**	1.90×	**75.85**	1.13×	**71.80**	1.08×	**85.13**	2.13×	**87.00**	2.15×	90.13	2.19×	**84.26**	1.73×
Based on Thyme [71]																
Thyme (w tools)	86.96	1.00×	82.89	1.00×	77.72	1.00×	72.43	1.00×	81.32	1.00×	84.53	1.00×	90.17	1.00×	82.29	1.00×
Thyme (w/o tools)	84.35	2.81×	76.32	2.56×	74.25	1.85×	69.88	1.97×	77.77	3.51×	78.17	3.32×	79.93	2.99×	77.24	2.72×
SpecReason [42]	89.57	0.48×	75.00	0.53×	80.01	0.52×	81.02	0.51×	84.62	0.46×	85.97	0.43×	90.27	0.46×	83.78	0.48×
▷ SpecEyes (log)	80.87	1.82×	**82.89**	1.45×	74.97	1.13×	70.84	1.06×	85.76	1.68×	87.80	1.67×	**91.47**	1.59×	82.09	1.49×
▷ SpecEyes (mean)	77.39	2.34×	80.26	1.83×	72.62	1.27×	68.00	1.21×	**85.89**	1.78×	**88.30**	1.80×	91.27	1.65×	80.53	1.70×
▷ SpecEyes (bottom-r)	78.26	2.18×	80.26	1.84×	77.35	1.05×	72.31	0.99×	**85.89**	1.81×	**88.30**	1.81×	91.27	1.73×	81.95	1.63×
▷ **SpecEyes (min)**	**87.83**	1.32×	**82.89**	1.42×	**78.47**	1.01×	**73.31**	0.95×	85.87	1.77×	**88.30**	1.78×	91.27	1.70×	**83.99**	1.42×

with 4K and 8K subsets (800 questions each). POPE [29] is a yes/no hallucination probe with Adversarial, Popular, and Random splits (3 000 questions each). All benchmarks are evaluated by accuracy. The small non-agentic model M_S is Qwen3-VL-2B [50], the large agentic model M_L is instantiated with DeepEyes [75] and Thyme [71], both capped at 5 tool-use steps per query.

Implementation Details. All models use greedy decoding (temperature 0), and all reported latencies include tool execution time. For cognitive gating (Sec. 3.3), we set K=64, ϵ=10^{-6}, and adopt min-token aggregation; for the bottom aggregation variant, we set the bottom fraction to r=0.2, inspired by [14]. The gating threshold τ is selected as follows: we run $\mathcal{M}_S$ once on a random 10% sample of each benchmark solely to *visualize* the empirical S_{sep} distribution and define a coarse search range. We then evenly sample a fixed grid of operating points from this range, the reported τ for each variant is chosen from this grid with *no per-sample or per-benchmark optimization*, so any overlap with the test set is immaterial and the results do not constitute an upper bound. All experiments run on a single NVIDIA A100 40GB GPU.

4.2 Main Results

Tab. 1 compares SpecEyes against the agentic baselines and SpecReason [42] across all seven evaluation splits, using two agentic backbones (DeepEyes [75] and Thyme [71]) paired with Qwen3-VL-2B [50] as the tool-free speculative model. For each SpecEyes variant, we report the result at the best operating-point threshold that preserves the baseline level accuracy. Among the four confidence aggregation

Table 2: Comparison of draft-only baselines and **SpecEyes (min)** on visual benchmarks. We report accuracy (%) and speedup (×) for two draft models (M_s).

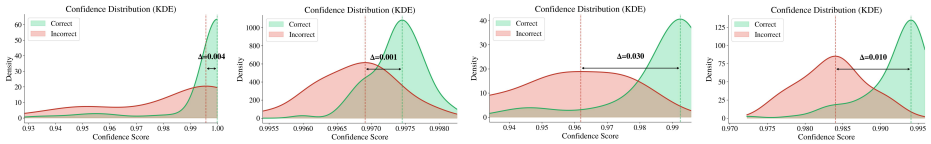
Dataset		M_s = **Qwen3-VL-8B**						M_s = **Qwen2.5-VL-7B**					
		Draft-only		**SpecEyes (min)**				Draft-only		**SpecEyes (min)**			
				DeepEyes		Thyme				DeepEyes		Thyme	
		Acc.	Spd.	Acc.	Spd.	Acc.	Spd.	Acc.	Spd.	Acc.	Spd.	Acc.	Spd.
V*	Attr.	81.74	4.23×	92.17	1.47×	90.43	1.28×	79.13	3.95×	90.43	0.92×	87.83	0.93×
	Pos.	78.95	1.72×	80.26	2.69×	80.26	1.57×	71.05	1.68×	78.95	1.63×	78.95	1.25×
HR-Bench	4K	77.50	2.46×	78.49	1.05×	78.38	0.96×	74.88	2.75×	75.97	1.02×	77.72	0.85×
	8K	69.90	1.54×	74.06	1.06×	74.22	0.94×	66.87	1.72×	71.43	0.98×	72.31	0.83×
POPE	Adv.	84.47	2.48×	84.23	1.75×	85.52	1.60×	79.66	2.68×	80.60	1.20×	82.82	1.06×
	Pop.	86.67	2.69×	86.47	1.80×	87.27	1.48×	81.23	3.01×	84.47	1.27×	86.20	1.17×
	Rand.	91.33	3.06×	89.70	1.85×	90.80	1.56×	88.79	3.55×	89.77	1.23×	91.10	1.19×
Avg.		81.51	2.60×	83.63	1.67×	83.84	1.34×	77.37	2.76×	81.66	1.18×	82.42	1.04×

strategies, SpecEyes (min) consistently delivers the strongest accuracy–speed profile, validating the worst-case guard design in Sec. 3.3; we focus the discussion on this variant below. With DeepEyes, SpecEyes (min) achieves a 1.73× average speedup while *improving* average accuracy from 81.39% to 84.26%. On V* Bench [57], it matches the baseline on Direct Attributes (90.43%, 1.53×) and boosts the Relative Position subset from 82.89% to 89.47% at 1.90×. POPE benefits most (2.13–2.19×) with accuracy consistently above baseline (*e.g.* Adversarial: 78.43% → 85.13%), suggesting that bypassing unnecessary tool trajectories can also reduce hallucination errors. HR-Bench yields moderate speedups (1.08–1.13×) because queries increasingly demand fine-grained, tool-assisted inspection.

Replacing the backbone with Thyme confirms generalization: SpecEyes (min) yields a 1.42× average speedup while raising accuracy from 82.29% to 83.99%. The per-benchmark pattern mirrors DeepEyes: POPE benefits most (1.70–1.78×), V* enjoys solid gains (1.32–1.42×), and HR-Bench remains the bottleneck (0.95–1.01×). The marginal sub-1× speedup on HR-Bench 8K arises because high-resolution inputs suppress both β and α, keeping $\beta\alpha$ low; in this regime the fixed cost of running M_S slightly exceeds any savings, consistent with Eq. (9).

In contrast, SpecReason [42] consistently *decelerates* inference (0.37–0.61× with DeepEyes; 0.43–0.53× with Thyme), as the small model lacks structured tool-calling capability and incurs substantial token and turn overhead (414 tokens and 3.48 rounds on average). It also degrades sharply on POPE (as low as 49.10%). By contrast, SpecEyes lets accepted queries bypass the tool-use chain entirely, avoiding this overhead. The Qwen3-VL-2B (draft only) row establishes a speedup upper bound (4.13×) at notable accuracy cost (78.93%); SpecEyes captures most of this latency saving while preserving full reasoning quality.

To isolate whether the gains come from *routing* or merely from skipping tool calls, we evaluate the backbone ($\mathcal{M}_L$) with tools disabled (*DeepEyes/Thyme w/o tools*). Although this achieves high speedup (3.53× / 2.72× on average), accuracy



(a) $S_{\log}$ **(b)** $S_{\text{sep}}^{\text{mean}}$ **(c)** $S_{\text{sep}}^{\min}$ **(d)** $S_{\text{sep}}^{\text{bottom-r}}$

Fig. 3: KDE of confidence scores for correct vs. incorrect samples on V*. Δ measures gating discriminability via peak distance with Qwen3-VL-2B. Compared to the noticeable overlap in baselines (**a**, **b**, **d**), our **(c)** $S_{\text{sep}}^{\min}$ achieves the largest Δ with sharp bimodal separation, enabling an optimal accuracy-speed trade-off.

collapses on POPE, which dropps from 78.43% to 46.90% on the Adversarial split with DeepEyes, because tool-required queries are forcibly answered without visual inspection. SpecEyes avoids this by routing only genuinely tool-free queries to $\mathcal{M}_S$ while preserving full agentic reasoning for the rest, achieving both strong accuracy and meaningful speedup simultaneously.

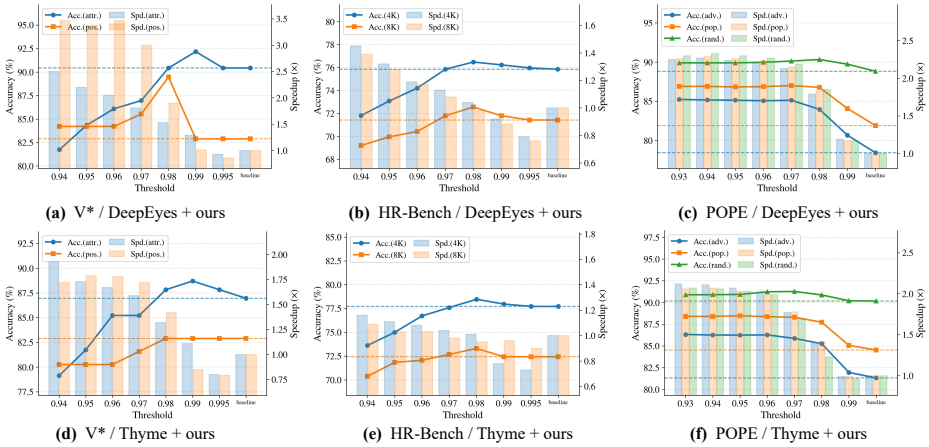
4.3 Analysis of Confidence Calibration

A reliable gating signal must be *discriminative*: confidence scores of correct answers should be stochastically higher than those of incorrect ones. Fig. 3 visualises this property via kernel density estimates (KDE) of each confidence score on correct and incorrect samples from M_S on V* [57], with each subplot annotated by Δ (peak distance between the two distributions) as a direct measure of discriminability. Both $S_{\log}$ (Fig. 3a) and $S_{\text{sep}}^{\text{mean}}$ (Fig. 3b) yield small Δ: the former suffers from softmax overconfidence, and the latter is diluted by averaging over all tokens, leaving the two distributions heavily overlapping. $S_{\text{sep}}^{\text{bottom-r}}$ (Fig. 3d) improves Δ by focusing on the lowest-separability tokens, yet residual overlap remains in the mid-range. $S_{\text{sep}}^{\min}$ (Fig. 3c) achieves the largest Δ: incorrect samples collapse to a low-score peak, while correct samples form a sharp high-score mode, consistent with Proposition 1. Tab. 1 shows that a single threshold to preserve accuracy while maximizing acceptance, explaining why SpecEyes (min) delivers a superior accuracy–speedup trade-off.

4.4 Ablation Study

We study the effects of three key hyperparameters in SpecEyes: the gating threshold, the serving batch size, and the separability computation parameter K.

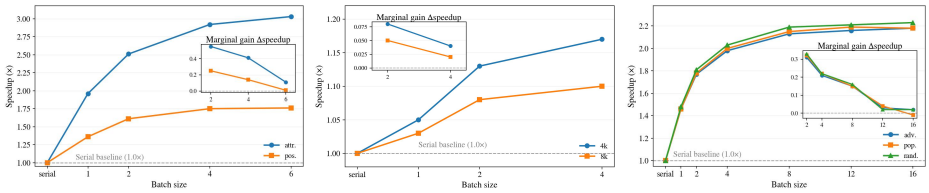
Ablation on Gating Threshold. Fig. 4 visualizes the accuracy–speedup trade-off as the gating threshold varies, using $S_{\text{sep}}^{\min}$ across three benchmarks with both agentic backbones. Lower thresholds increase the acceptance ratio and speedup, while accuracy degrades gracefully. On V* and POPE, accuracy remains above or close to the agentic baseline over a wide range, indicating that many queries can be safely bypassed. HR-Bench is more sensitive, with limited speedup gains and accuracy drops below 0.97 due to a higher fraction of tool-required queries. Overall, SpecEyes achieves a broad operating region that improves both



(a) V* / DeepEyes + ours (b) HR-Bench / DeepEyes + ours (c) POPE / DeepEyes + ours

(d) V* / Thyme + ours (e) HR-Bench / Thyme + ours (f) POPE / Thyme + ours

Fig. 4: Ablation on the gating threshold of SpecEyes. Lowering the threshold increases speedup at cost of accuracy. Dashed horizontal lines indicate baseline accuracy.



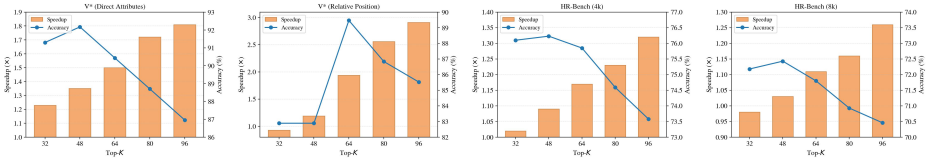
(a) V* Bench (b) HR-Bench (c) POPE

Fig. 5: Ablation on serving batch size. Larger batches amortize the stateless speculative stage, improving speedup with diminishing marginal gains as the stateful fallback bottlenecks. Curves report end-to-end speedup over the serial agentic baseline.

accuracy and efficiency, showing that the threshold serves as a smooth control knob for navigating the accuracy–efficiency Pareto front.

Ablation on Batch Size. Fig. 5 studies the impact of serving batch size with the same gating threshold as the main results. Increasing batch size consistently improves end-to-end speedup without affecting accuracy, as batching only changes system execution. This follows from our heterogeneous funnel design (Sec. 3.4): the stateless speculative stage is highly batchable, while the agentic fallback stage remains constrained by per-query tool dependencies, leading to diminishing gains at larger batches. Benchmarks with higher bypass rates benefit more from batching, whereas HR-Bench saturates earlier due to more tool-required queries.

Ablation on Top-K in Separability Computation. As shown in Fig. 6, K acts as a *control knob*: increasing K monotonically improves speedup but degrades accuracy, mirroring the effect of lowering the gating threshold, as larger K includes tokens with weaker contrastive signal and thereby inflates confidence estimates. We set K=64 as a balanced default, which matches baseline accuracy on Direct Attributes subsets and achieves a strong speedup on Relative Position (1.94×), while overly large K over-optimizes for speed at the cost of accuracy.



(a) V* (Attr.) **(b)** V* (Pos.) **(c)** HR-Bench (4k) **(d)** HR-Bench (8k)

Fig. 6: Ablation on Top-K in separability-based gating. Increasing K boosts speed but degrades accuracy, tuning the model's speculative aggressiveness.

Ablation on Draft Model. As shown in Tab. 2, replacing Qwen3-VL-2B with a larger non-agentic model demonstrates that SpecEyes is model-agnostic. With Qwen3-VL-8B as $\mathcal{M}_S$, SpecEyes achieves higher accuracy on V* (92.17% Attr.) and HR-Bench 4K (78.49%) under DeepEyes, showing that stronger speculators improve acceptance quality. However, the larger model reduces speedup to 1.67× (vs. 1.73× for 2B), making the 2B variant the better accuracy–efficiency trade-off. Qwen2.5-VL-7B follows a similar trend: despite comparable accuracy to the 2B variant, it achieves lower speedup (1.18× with DeepEyes and 1.04× with Thyme), indicating that c_S becomes the main bottleneck for larger speculators.

5 Conclusion and Future Work

In this paper, we present SpecEyes, an agentic-level speculative acceleration framework that lifts the speculation paradigm from individual tokens to the entire agentic pipeline. A lightweight, tool-free model speculatively answers queries that do not require multi-step tool use, governed by a *cognitive gating* mechanism based on answer separability and served through a *heterogeneous parallel funnel* that converts per-query latency savings into system-level throughput gains. Across three diverse image understanding benchmarks, SpecEyes reduces end-to-end latency by up to 3.35×, while it is comparable with the agentic baseline in accuracy and delivers consistent throughput improvements under concurrent serving. However, our speculative model currently operates at agentic depth D=0 (fully tool-free), limiting speedups on benchmarks (*e.g.* HR-Bench) where most queries genuinely require tool assistance. A natural extension in future work is *multi-depth speculation* (D=1, 2, . . . , n), allowing the speculative model a bounded number of lightweight tool calls before gating, thereby intercepting queries at the earliest sufficient depth and further reducing unnecessary fallbacks.

Acknowledgments

This work was supported by the New Generation Artificial Intelligence-National Science and Technology Major Project (No. 2025ZD0122702), the National Natural Science Foundation of China (No. 62576299), the Fundamental Research Funds for the Central Universities, and the Xiaomi Young Talents Program.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
2. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
3. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. arXiv preprint arXiv:2210.09461 (2022)
4. Cai, T., Li, Y., Geng, Z., Peng, H., Lee, J.D., Chen, D., Dao, T.: Medusa: Simple llm inference acceleration framework with multiple decoding heads. arXiv preprint arXiv:2401.10774 (2024)
5. Chen, C., Borgeaud, S., Irving, G., Lespiau, J.B., Sifre, L., Jumper, J.: Accelerating large language model decoding with speculative sampling. arXiv preprint arXiv:2302.01318 (2023)
6. Chen, W., Zeng, Y., Luo, Y., Xie, T., Lin, L., Ji, J., Zhang, Y., Zheng, X.: Wavelet-based frame selection by detecting semantic boundary for long video understanding (2026), https://arxiv.org/abs/2603.00512
7. Chen, Y., Pan, X., Li, Y., Ding, B., Zhou, J.: Ee-llm: Large-scale training and inference of early-exit large language models with 3d parallelism. arXiv preprint arXiv:2312.04916 (2023)
8. Chng, Y.X., Hu, T., Tong, W., Li, X., Chen, J., Yu, H., Lu, J., Guo, H., Deng, H., Xie, C., Huang, G., Lin, D., Lu, L.: Sensenova-mars: Empowering multimodal agentic reasoning and search via reinforcement learning. arXiv preprint arXiv:2512.24330 (2025)
9. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems **36**, 49250–49267 (2023)
10. Doshi, R.: Introducing Agentic Vision in Gemini 3 Flash. https://blog.google/innovation-and-ai/technology/developers-tools/agentic-vision-gemini-3-flash/ (January 2026), accessed: 2026-02-24
11. Endo, M., Wang, X., Yeung-Levy, S.: Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. ICCV (2025)
12. Fan, S., Jiang, X., Li, X., Meng, X., Han, P., Shang, S., Sun, A., Wang, Y., Wang, Z.: Not all layers of llms are necessary during inference. arXiv preprint arXiv:2403.02181 (2024)
13. Fu, T., Liu, T., Han, Q., Dai, G., Yan, S., Yang, H., Ning, X., Wang, Y.: Framefusion: Combining similarity and importance for video token reduction on large vision language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22654–22663 (2025)
14. Fu, Y., Wang, X., Tian, Y., Zhao, J.: Deep think with confidence. arXiv preprint arXiv:2508.15260 (2025)
15. Guan, Y., Lan, Q., Fei, S., Ding, D., Acharya, D., Wang, C., Wang, W.Y., Hua, W.: Dynamic speculative agent planning. arXiv preprint arXiv:2509.01920 (2025)
16. Guo, Z., Hong, M., Zhang, F., Jia, K., Jin, T.: Thinking with programming vision: Towards a unified view for thinking with images. arXiv preprint arXiv:2512.03746 (2025)
17. He, Y., Chen, F., Liu, J., Shao, W., Zhou, H., Zhang, K., Zhuang, B.: Zipvl: Efficient large vision-language models with dynamic token sparsification (2024), https://arxiv.org/abs/2410.08584

18. Hong, J., Zhao, C., Zhu, C., Lu, W., Xu, G., Yu, X.: Deepeyesv2: Toward agentic multimodal model. arXiv preprint arXiv:2511.05271 (2025)
19. Hou, X., Xu, S., Biyani, M., Li, M., Liu, J., Hollon, T.C., Wang, B.: Codev: Code with images for faithful visual reasoning via tool-aware policy optimization (2026), https://arxiv.org/abs/2511.19661
20. Hu, P., Cao, M., Wang, Y., Wang, Y., Dong, J., Song, J., Cheng, Y., Zheng, B., Liang, X.: Thinking with drafts: Speculative temporal reasoning for efficient long video understanding. ArXiv **abs/2512.00805** (2025), https://api.semanticscholar.org/CorpusID:283449335
21. Huang, C., Zheng, T., Huang, L., Li, J., Liu, H., Huang, J.: Relayllm: Efficient reasoning via collaborative decoding. ArXiv **abs/2601.05167** (2026), https://api.semanticscholar.org/CorpusID:284544142
22. Kim, M., Gao, S., Hsu, Y.C., Shen, Y., Jin, H.: Token fusion: Bridging the gap between token pruning and token merging. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1383–1392 (2024)
23. Kumar, A., Nag, S., Clemons, J., John, L., Das, P.: Helios: Adaptive model and early-exit selection for efficient llm inference serving. arXiv preprint arXiv:2504.10724 (2025)
24. Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C.H., Gonzalez, J., Zhang, H., Stoica, I.: Efficient memory management for large language model serving with pagedattention. In: Proceedings of the 29th symposium on operating systems principles. pp. 611–626 (2023)
25. Lai, X., Li, J., Li, W., Liu, T., Li, T., Zhao, H.: Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. ArXiv **abs/2509.07969** (2025), https://api.semanticscholar.org/CorpusID:281217747
26. Leviathan, Y., Kalman, M., Matias, Y.: Fast inference from transformers via speculative decoding. In: International Conference on Machine Learning. pp. 19274–19286. PMLR (2023)
27. Li, J., Li, D., Savarese, S., Hoi, S.: 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 (2023)
28. Li, X., Liang, Y., Chen, X., Zheng, Y., Chen, H., Li, B., Xue, X.: Hero: Rethinking visual token early dropping in high-resolution large vision-language models (2025), https://arxiv.org/abs/2509.13067
29. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 292–305. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.20, https://aclanthology.org/2023.emnlp-main.20/
30. Li, Y., Wei, F., Zhang, C., Zhang, H.: Eagle-2: Faster inference of language models with dynamic draft trees. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 7421–7432 (2024)
31. Li, Y., Wei, F., Zhang, C., Zhang, H.: Eagle: Speculative sampling requires rethinking feature uncertainty. arXiv preprint arXiv:2401.15077 (2024)
32. Li, Y., Wei, F., Zhang, C., Zhang, H.: Eagle-3: Scaling up inference acceleration of large language models via training-time test. arXiv preprint arXiv:2503.01840 (2025)
33. Lian, S., Wu, Y., Ma, J., Ding, Y., Song, Z., Chen, B., Zheng, X., Li, H.: Ui-agile: Advancing gui agents with effective reinforcement learning and precise inference-time grounding. arXiv preprint arXiv:2507.22025 (2025)

34. Lin, B., Tang, Z., Ye, Y., Huang, J., Zhang, J., Pang, Y., Jin, P., Ning, M., Luo, J., Yuan, L.: Moe-llava: Mixture of experts for large vision-language models. IEEE Transactions on Multimedia (2026)
35. Lin, L., Lin, Z., Zeng, Z., Ji, R.: Speculative decoding reimagined for multimodal large language models. arXiv preprint arXiv:2505.14260 (2025)
36. Lin, Z., Sun, Z., Chen, S., Chen, X., Zhao, R.: Accelerating multi-modal llm gaming performance via input prediction and mishit correction. arXiv preprint arXiv:2512.17250 (2025)
37. Liu, Z., Liu, B., Wang, J., Dong, Y., Chen, G., Rao, Y., Krishna, R., Lu, J.: Efficient inference of vision instruction-following models with elastic cache. In: European Conference on Computer Vision. pp. 54–69. Springer (2024)
38. Luo, Y., Zheng, X., Li, G., Yin, S., Lin, H., Fu, C., Huang, J., Ji, J., Chao, F., Luo, J., et al.: Video-rag: Visually-aligned retrieval-augmented long video comprehension. Advances in Neural Information Processing Systems **38**, 168008–168033 (2026)
39. Ma, Q., Ma, Y., Xie, Y., Xie, T., Zheng, X., Ji, R.: A2rbench: An automatic paradigm for formally verifiable abstract reasoning benchmark generation (2026), https://arxiv.org/abs/2605.17278
40. Ma, Q., Wu, Y., Zheng, X., Ji, R.: Benchmarking abstract and reasoning abilities through a theoretical perspective (2025), https://arxiv.org/abs/2505.23833
41. OpenAI: Introducing OpenAI o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/ (April 2025), accessed: 2026-02-24
42. Pan, R., Dai, Y., Zhang, Z., Oliaro, G., Jia, Z., Netravali, R.: Specreason: Fast and accurate inference-time compute via speculative reasoning. arXiv preprint arXiv:2504.07891 (2025)
43. Peng, Y., Wang, P., Wang, X., Wei, Y., Pei, J., Qiu, W., Jian, A., Hao, Y., Pan, J., Xie, T., et al.: Skywork r1v: Pioneering multimodal reasoning with chain-of-thought. arXiv preprint arXiv:2504.05599 (2025)
44. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., Scialom, T.: Toolformer: Language models can teach themselves to use tools. Advances in neural information processing systems **36**, 68539–68551 (2023)
45. Shao, Z., Wang, M., Yu, Z., Pan, W., Yang, Y., Wei, T., Zhang, H., Mao, N., Chen, W., Yu, J.: Growing a twig to accelerate large vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20064–20074 (2025)
46. Shen, Y., Song, K., Tan, X., Li, D., Lu, W., Zhuang, Y.: Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. Advances in Neural Information Processing Systems **36**, 38154–38180 (2023)
47. Song, Q., Li, H., Yu, Y., Zhou, H., Yang, L., Bai, S., She, Q., Huang, Z., Zhao, Y.: Codedance: A dynamic tool-integrated mllm for executable visual reasoning. ArXiv **abs/2512.17312** (2025), https://api.semanticscholar.org/CorpusID:284058227
48. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
49. Team, K., Bai, T., Bai, Y., Bao, Y., Cai, S., Cao, Y., Charles, Y., Che, H., Chen, C., Chen, G., et al.: Kimi k2. 5: Visual agentic intelligence. arXiv preprint arXiv:2602.02276 (2026)
50. Team, Q.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388

51. Teerapittayanon, S., McDanel, B., Kung, H.T.: Branchynet: Fast inference via early exiting from deep neural networks. In: 2016 23rd international conference on pattern recognition (ICPR). pp. 2464–2469. IEEE (2016)
52. Wan, Z., Shen, H., Wang, X., Liu, C., Mai, Z., Zhang, M.: Meda: Dynamic kv cache allocation for efficient multimodal long-context inference. 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. 2485–2497 (2025)
53. Wan, Z., Wu, Z., Liu, C., Huang, J., Zhu, Z., Jin, P., Wang, L., Yuan, L.: Look-m: Look-once optimization in kv cache for efficient multimodal long-context inference. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 4065–4078 (2024)
54. Wang, H., Kai, J., Bai, H., Hou, L., Jiang, B., He, Z., Lin, Z.: Fourier-vlm: Compressing vision tokens in the frequency domain for large vision-language models (2025), https://arxiv.org/abs/2508.06038
55. Wang, W., Ding, L., Zeng, M., Zhou, X., Shen, L., Luo, Y., Yu, W., Tao, D.: Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 7907–7915 (2025)
56. Wang, Y., Yang, Y.: Efficient visual transformer by learnable token merging. IEEE transactions on pattern analysis and machine intelligence (2025)
57. Wu, P., Xie, S.: V*: Guided visual search as a core mechanism in multimodal llms. arXiv preprint arXiv:2312.14135 (2023)
58. Xia, H., Ge, T., Wang, P., Chen, S.Q., Wei, F., Sui, Z.: Speculative decoding: Exploiting speculative execution for accelerating seq2seq generation. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 3909–3925. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.findings-emnlp.257, https://aclanthology.org/2023.findings-emnlp.257/
59. Xia, H., Li, Y., Zhang, J., Du, C., Li, W.: Swift: On-the-fly self-speculative decoding for llm inference acceleration. arXiv preprint arXiv:2410.06916 (2024)
60. Xie, T., Huang, J., Ma, Y., Luo, R., Yang, Y., Chen, W., Zeng, Y., Fang, R., Zou, Y., Zheng, X., et al.: Socialomni: Benchmarking audio-visual social interactivity in omni models. arXiv preprint arXiv:2603.16859 (2026)
61. Xie, T., Ma, Y., Wu, Y., Chen, W., Ji, J., Chua, T.S., Zheng, X., Ji, R.: Training-free multimodal large language model orchestration. arXiv preprint arXiv:2508.10016 (2025)
62. Xing, L., Huang, Q., Dong, X., Lu, J., Zhang, P., Zang, Y., Cao, Y., He, C., Wang, J., Wu, F., et al.: Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. arXiv preprint arXiv:2410.17247 (2024)
63. Xiong, J., Chen, Q., Ye, F., Wan, Z., Zheng, C., Zhao, C., Shen, H., Li, H., Tao, C., Tan, H., Bai, H., Shang, L., Kong, L., Wong, N.: Atts: Asynchronous test-time scaling via conformal prediction (2026), https://arxiv.org/abs/2509.15148
64. Xu, J., Pan, J., Zhou, Y., Chen, S., Li, J., Lian, Y., Wu, J., Dai, G.: Specee: Accelerating large language model inference with speculative early exiting. In: Proceedings of the 52nd Annual International Symposium on Computer Architecture. pp. 467–481 (2025)
65. Yang, P., Du, C., Zhang, F., Wang, H., Pang, T., Du, C., An, B.: Longspec: Long-context speculative decoding with efficient drafting and verification. arXiv e-prints pp. arXiv–2502 (2025)

66. Yang, S., Chen, Y., Tian, Z., Wang, C., Li, J., Yu, B., Jia, J.: Visionzip: Longer is better but not necessary in vision language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19792–19802 (2025)
67. Yang, W., Xia, Y., Huang, J., Lu, S., Chen, Q.G., Xu, Z., Luo, W., Zhang, K., Wan, Y., Zhang, L.: Deep but reliable: Advancing multi-turn reasoning for thinking with images (2026), https://arxiv.org/abs/2512.17306
68. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K.R., Cao, Y.: React: Synergizing reasoning and acting in language models. In: The eleventh international conference on learning representations (2022)
69. Yu, Z., Zhang, J., Su, H., Zhao, Y., Wu, Y., Deng, M., Xiang, J., Lin, Y., Tang, L., Luo, Y., et al.: Recode: Unify plan and action for universal granularity control. arXiv preprint arXiv:2510.23564 (2025)
70. Zhang, J., Wang, J., Li, H., Shou, L., Chen, K., Chen, G., Mehrotra, S.: Draft& verify: Lossless large language model acceleration via self-speculative decoding. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 11263–11282 (2024)
71. Zhang, Y.F., Lu, X., Yin, S., Fu, C., Chen, W., Hu, X., Wen, B., Jiang, K., Liu, C., Zhang, T., et al.: Thyme: Think beyond images. arXiv preprint arXiv:2508.11630 (2025)
72. Zhang, Y., Hu, L., Sun, H., Wang, P., Wei, Y., Yin, S., Pei, J., Shen, W., Xia, P., Peng, Y., et al.: Skywork-r1v4: Toward agentic multimodal intelligence through interleaved thinking with images and deepresearch. arXiv preprint arXiv:2512.02395 (2025)
73. Zhao, S., Lin, S., Li, M., Zhang, H., Peng, W., Zhang, K., Wei, C.: Pyvision-rl: Forging open agentic vision models via rl. arXiv preprint arXiv:2602.20739 (2026)
74. Zhao, W., Han, Y., Tang, J., Li, Z., Song, Y., Wang, K., Wang, Z., You, Y.: A stitch in time saves nine: Small vlm is a precise guidance for accelerating large vlms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19814–19824 (2025)
75. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing" thinking with images" via reinforcement learning. arXiv preprint arXiv:2505.14362 (2025)
76. Zhu, Y., Yang, X., Wu, Y., Zhang, W.: Hierarchical skip decoding for efficient autoregressive text generation. arXiv preprint arXiv:2403.14919 (2024)