BEYOND REASONING: CONSENSUS-GUIDED VISUAL COLLABORATION IN LATENT MULTI-AGENT SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual multi-agent systems (VMAS) leverage complementary reasoning roles for complex vision-centric tasks, but their prevalent reliance on natural-language communication creates an information bottleneck and substantial inference overhead. To address these limitations, we instantiate latent collaboration for VMAS, where agents reuse a shared visual KV cache and propagate intermediate reasoning KV states without generating textual messages. However, reasoning-state collaboration alone does not explicitly coordinate the role-conditioned visual attention patterns over the shared visual input. Our analysis shows that visual regions jointly prioritized across agents exhibit higher ground-truth region hit rates, suggesting cross-agent visual attention as a complementary signal for collaboration. Motivated by this observation, we propose `V-Guide`, a training-free framework that augments latent reasoning-state collaboration with explicit cross-agent visual coordination. Specifically, `V-Guide` first spatially marginalizes role-conditioned visual attention distributions to accommodate local token-level discrepancies, and then aggregates them into a geometric consensus distribution that captures cross-agent support. Finally, the resulting consensus distribution modulates the shared visual Value states, increasing the contribution of consensus-supported visual content for final reasoning. Extensive experiments across five visual reasoning benchmarks, four MLLM backbones, and different multi-agent architectures demonstrate that `V-Guide` outperforms text-mediated and pure-latent multi-agent baselines while retaining the efficiency advantages of latent collaboration.

1 INTRODUCTION

Recent advances in multimodal large language models (MLLMs) [1, 2, 3] have spurred growing interest in visual multi-agent systems (VMAS) [4, 5], where multiple agents with complementary roles collaborate to solve complex vision-centric reasoning tasks [6, 7]. Prevalent VMAS mainly rely on **text-mediated collaboration** [7, 8, 9, 10, 11], requiring each agent to externalize its intermediate reasoning as discrete text before passing it to subsequent agents. Despite its simplicity and flexibility, such an inter-agent communication paradigm introduces an inherent *information bottleneck*: rich multimodal states, particularly fine-grained visual evidence that is difficult to verbalize, must be compressed into language before they can participate in collaboration [12, 13, 14, 15]. Meanwhile, text-mediated collaboration repeatedly converts internal reasoning into discrete tokens and re-processes them as context in subsequent agents, incurring substantial inference overhead [16, 17]. These limitations motivate a more direct form of inter-agent collaboration that operates on internal representations (*e.g.*, KV cache) without repeatedly translating them into natural language [18].

To this end, we explore **pure latent collaboration** for VMAS, inspired by advanced language-only multi-agent systems [18, 19, 20, 21]. Specifically, agents operate on shared visual KV cache of the input image and conduct intermediate reasoning directly in the latent space, with the resulting reasoning KV caches propagated to subsequent agents as part of their context. This avoids repeated image encoding and intermediate text generation, enabling more direct and efficient inter-agent collaboration. However, visual multi-agent reasoning goes beyond propagating reasoning states across agents: each agent also interprets the shared visual representation according to its own role and evolving reasoning context. This produces distinct role-conditioned visual attention distributions over the shared visual tokens across agents, indicating which parts of the image each agent prioritizes

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

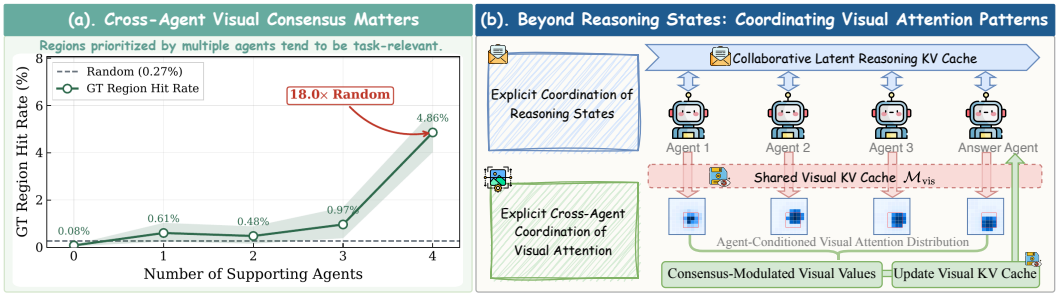


Figure 1: **Beyond reasoning-state collaboration: coordinating visual attention patterns across agents. (a)** The number of supporting agents indicates how many agents rank a visual region within their top 5% of attention score. Visual tokens receiving high attention from all four agents exhibit a substantially higher ground-truth (GT) region hit rate than randomly selected tokens, highlighting the potential of cross-agent visual attention consensus to guide final answer generation. **(b)** Beyond latent reasoning-state propagation, `V-Guide` explicitly consolidates cross-agent visual attention into a consensus prior to modulate visual values for final reasoning.

during reasoning. Under existing latent *reasoning-centric collaboration*, however, these visual attention patterns over the shared visual tokens are not explicitly coordinated across agents.

Behavioral studies on human collaborative problem solving suggest that effective collaboration involves coordinating not only reasoning processes but also visual attention across participants [22, 23, 24]. Such visual attention coordination helps establish a common perceptual basis for collective reasoning by directing participants toward task-relevant information [25, 26]. Motivated by this observation, we hypothesize that cross-agent visual attention may provide complementary information for visual collaboration in VMAS. As shown in Fig. 1(a), visual regions receiving high attention from multiple agents (*i.e.*, exhibiting cross-agent visual attention consensus) are more likely to overlap with ground-truth task-relevant regions than randomly selected regions.

Building on this insight, we propose `V-Guide`, a training-free latent collaboration framework that consolidates heterogeneous visual attention patterns across agents into a consensus prior for visual modulation during final answer generation. First, `V-Guide` collects attention distribution of each agent over the shared visual tokens as its role-conditioned visual attention pattern and applies **Spatial Uncertainty Marginalization**, which redistributes attention mass across neighboring visual tokens using a Gaussian transition kernel. This operation accommodates local discrepancies among agent attention patterns, transforming brittle token-level coincidence into region-level correspondence. Second, `V-Guide` derives a **Geometric Consensus Distribution** over the shared visual tokens as the minimizer of a KL-based agreement objective over the spatially marginalized attention distributions. The resulting distribution places greater mass on visual tokens receiving consistent cross-agent support while downweighting isolated agent-specific responses. Third, **Consensus-Guided Visual Value Modulation** converts the consensus distribution into bounded token-wise weights and modulates the shared visual Values according to these weights, increasing the contribution of visual content with consistent cross-agent support for final reasoning.

We evaluate `V-Guide` on five visual reasoning benchmarks using four MLLM backbones under both sequential and hierarchical multi-agent architectures. Experimental results show that `V-Guide` outperforms text-mediated and pure-latent collaboration baselines, improving average performance by up to 8.36 percentage points over Latent-VMAS while reducing token cost by 32%–66% compared with text-mediated collaboration. Further analyses demonstrate the importance of high-consensus visual tokens for final predictions and the effectiveness of cross-agent visual attention consensus in guiding visual Value modulation (§ 4.3).

Our main contributions are summarized as follows: **(i)** We uncover the value of cross-agent visual attention consensus for visual collaboration in latent multi-agent systems. Our analysis shows that visual regions jointly prioritized across agents exhibit higher ground-truth region hit rates than randomly selected regions, motivating consensus-guided visual coordination beyond reasoning-state collaboration. **(ii)** We introduce `V-Guide`, a training-free latent collaboration framework that consolidates heterogeneous visual attention patterns across agents into a consensus prior for visual modulation during final answer generation. **(iii)** Extensive experiments demonstrate that `V-Guide` improves visual reasoning performance over text-mediated and pure-latent collaboration baselines while preserving the efficiency advantages of latent collaboration.

2 Related Work

LLM-based Multi-Agent Systems. LLM-based multi-agent systems [10, 11, 27, 28, 29, 30] enable collaborative reasoning through role specialization [31, 32, 33, 34, 35, 36] and iterative interaction [37]. However, conventional natural-language communication [37, 38, 39, 40, 41] incurs substantial decoding and re-encoding overhead while limiting the exchange of rich internal representations [12, 13, 14, 15]. To address these limitations, recent studies explore direct communication through internal model representations, including intermediate activations, hidden states, and KV caches [42, 43, 44, 21, 20]. LatentMAS [18] further advances this direction by unifying latent reasoning and inter-agent communication through a shared KV-cache-based working memory. These methods have primarily focused on language-only reasoning, leaving their potential for visual multi-agent collaboration largely unexplored.

Visual Multi-Agent Systems. Recent visual multi-agent systems (VMAS) enhance complex visual reasoning by coordinating multiple vision-language agents with complementary roles [7, 8, 9, 45, 46]. Most existing VMAS adopt **text-mediated collaboration**, where agents convey their intermediate reasoning to subsequent agents through textual messages. Recent studies further enrich inter-agent communication with visual information. EAGLE [47] requires each agent to produce grounding boxes and visual claims alongside its answer, and uses their cross-agent consistency for diagnosis, revision, and final arbitration, while ViF [13] augments inter-agent textual messages with selected visual relay tokens and reallocates attention to counter the progressive decline of visual attention across agent turns, requiring additional training. Beyond purely text-mediated communication, L^2-VMAS [48] augments textual reasoning with external latent memories. It also requires a three-stage reinforcement learning procedure to train its dedicated memory synthesis and orchestration modules. In contrast, our work explores a training-free, **pure-latent collaboration** paradigm for VMAS, where each agent reuses shared visual KV caches and directly propagates reasoning KV caches, without intermediate text generation. Beyond reasoning-state propagation, we further coordinate role-conditioned visual attention patterns across agents to enhance visual collaboration.

3 Methodology

3.1 Overall Framework

We introduce `V-Guide`, a training-free latent collaboration framework that leverages cross-agent visual attention consensus to enhance visual multi-agent reasoning. As shown in Fig. 2 (a), given input image and question, `V-Guide` first encodes the image once and maintains a **shared visual KV cache** accessible to all agents. Each agent performs role-specific **latent reasoning** and propagates the resulting KV caches to subsequent agents according to the multi-agent communication topology, without generating intermediate textual messages (§3.2). To facilitate visual coordination, `V-Guide` consolidates the visual attention patterns across agents into a consensus prior for visual modulation during final answer generation (Fig. 2(b)). Specifically, `V-Guide` first extracts agent-conditioned visual attention distributions over the shared visual tokens from the textual-prefix Query states of individual agents. These distributions are processed by **Spatial Uncertainty Marginalization** to accommodate local spatial variations, followed by **Geometric Consensus** to emphasize visual regions receiving consistent support across agents. (§ 3.3). Finally, **Consensus-Guided Visual Value Modulation** leverages the resulting consensus distribution to modulate the shared visual Values, increasing the contribution of consensus-supported visual content during answer generation (§ 3.4).

3.2 Visual Latent Multi-Agent Collaboration

MLLM-based VMAS Setting. Given an image I, a query Q, and a visual multi-agent system (VMAS) $\mathcal{S}$ consisting of M agents $\mathcal{A} = \{A_i\}_{i=1}^{M}$, we study visual multi-agent collaboration entirely in the latent space. All agents are instantiated from a shared MLLM with different role instructions, allowing the latent states (*i.e.*, KV cache) produced by one agent to be directly used by others. VMAS may adopt different communication topologies depending on the task, and our method does not assume a fixed topology. Following LatentMAS [18], we consider two representative multi-agent architectures, as shown in Fig. 2(c): Sequential MAS, where agents progressively build upon preceding latent reasoning to refine their predictions, and Hierarchical MAS, where specialized agents

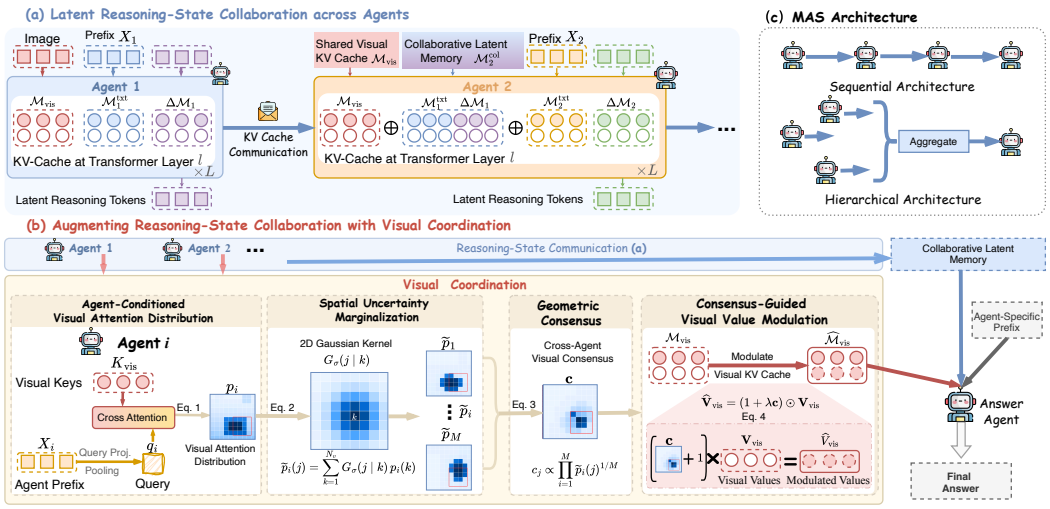


Figure 2: **Overview of `V-Guide`**, a training-free latent collaboration framework with explicit cross-agent visual attention coordination. (a) **Latent Reasoning-State Collaboration.** Agents reuse a shared visual KV cache, perform role-specific *latent reasoning*, and propagate reasoning KV caches across agents without intermediate text generation. (b) **Cross-Agent Visual Coordination.** `V-Guide` first collects role-conditioned visual attention distributions over the shared visual tokens and applies *Spatial Uncertainty Marginalization* to accommodate local spatial variations. The resulting distributions are then aggregated through *Geometric Consensus* to emphasize visual regions receiving consistent support across agents. Finally, *Consensus-Guided Visual Value Modulation* leverages the resulting consensus distribution to modulate the shared visual Values, increasing the contribution of visual tokens receiving consistent cross-agent attention to final answer generation. (c) MAS Architecture.

examine the input from complementary perspectives and their latent states are integrated by a final answering agent. To enable latent collaboration over a shared visual input, **`V-Guide`** maintains a shared visual KV cache accessible to all agents and propagates agent-specific latent reasoning states according to the communication topology $\mathcal{S}$.

Shared Visual KV Cache. Given an input image I, **`V-Guide`** places the visual tokens before all agent-specific textual and latent tokens. The image is encoded only once, yielding a layer-wise visual KV cache $\mathcal{M}_{\text{vis}} = \left\{ \left(K_{\text{vis}}^{(l)}, V_{\text{vis}}^{(l)} \right) \right\}_{l=1}^{L}$, where $K_{\text{vis}}^{(l)}$ and $V_{\text{vis}}^{(l)}$ denote the Key and Value states of the visual tokens at layer l. Under causal attention, these states are unaffected by subsequent agent-specific contexts. By maintaining the same visual-token positions across agents, **`V-Guide`** reuses the visual KV cache throughout collaboration without repeated image encoding.

Latent Reasoning-State Collaboration. For each agent A_i, let X_i denote the textual prefix containing the query and role-specific instruction, with the corresponding KV cache denoted by $\mathcal{M}_i^{\text{txt}}$. The collaborative latent memory $\mathcal{M}_i^{\text{col}}$ preserves the textual-prefix and latent reasoning KV states of preceding agents according to the communication topology $\mathcal{S}$. Each agent operates on the shared visual cache $\mathcal{M}_{\text{vis}}$, followed by the collaborative latent memory $\mathcal{M}_i^{\text{col}}$ and the agent-specific textual prefix cache $\mathcal{M}_i^{\text{txt}}$. Given this context, agent A_i performs T continuous latent reasoning steps without generating intermediate text. At each step, the last-layer hidden state is mapped to the input embedding space through the training-free input–output alignment operator of LatentMAS [18], providing the input for the next latent reasoning step. The newly generated layer-wise KV states are appended to the cache at each step, yielding $\Delta\mathcal{M}_i = \left\{ \left([k_{i,t}^{(l)}]_{t=1}^{T}, [v_{i,t}^{(l)}]_{t=1}^{T} \right) \right\}_{l=1}^{L}$. These reasoning KV states are propagated to subsequent agents according to $\mathcal{S}$, enabling them to build upon preceding latent reasoning without explicit textual communication. Finally, the answering agent generates the natural-language response conditioned on the shared visual cache and the collaborative latent context.

3.3 Cross-Agent Visual Consensus

Although all agents share the same visual KV cache $\mathcal{M}_{\text{vis}}$, their distinct reasoning roles induce heterogeneous attention patterns over the shared visual tokens. Our motivation analysis (§ 1) indicates that visual regions jointly attended to across agents are likely to be task-relevant, motivating the use of

cross-agent visual attention consensus to identify shared visual focus. To exploit this signal, **V-Guide** extracts agent-conditioned visual attention distributions accommodating local spatial variations through Spatial Uncertainty Marginalization, and consolidates the resulting distributions into a Geometric Consensus Distribution that emphasizes consistent cross-agent support.

Agent-Conditioned Visual Attention Distribution. For agent A_i, **V-Guide** mean-pools the Query states of its textual prefix X_i at Transformer layer l and attention head h to obtain an agent-conditioned query $q_i^{(l,h)}$. To characterize where each agent prioritizes its attention within the shared visual representation, we compute a **conditional visual attention distribution** by normalizing the query–key scores exclusively over the shared visual tokens. We then aggregate the resulting distributions across attention heads and intermediate Transformer layers:

$$a_i^{(l,h)} = \mathrm{Softmax}_{\mathcal{V}}\left(\frac{q_i^{(l,h)}\big(K_{\mathrm{vis}}^{(l,h)}\big)^\top}{\sqrt{d_h}}\right), \qquad p_i(j) = \frac{1}{|\mathcal{L}_m|H}\sum_{l\in\mathcal{L}_m}\sum_{h=1}^{H} a_i^{(l,h)}(j). \tag{1}$$

Here, H denotes the number of attention heads, d_h is the head dimension, and $\mathcal{L}_m$ denotes the selected middle decoder layers. Our layer-wise analysis reveals that these layers more reliably localize question-relevant visual regions across agents (Appendix D), consistent with similar observations in prior studies [49, 50, 51]. The operator $\mathrm{Softmax}_{\mathcal{V}}$ normalizes the scores over the N_v shared visual tokens. This conditional normalization is deliberate: agents may operate under different reasoning contexts and allocate different proportions of their overall attention to visual and non-visual tokens. Rather than measuring the absolute attention allocated to visual content, our objective is to identify the relative visual attention patterns of different agents within the shared image. Accordingly, p_i characterizes the visual locations prioritized by agent A_i, providing a common basis for identifying spatial agreement across agents. As detailed in Appendix E.1, the pooled-query readout admits an interpretation as a normalized geometric aggregation of token-level conditional visual attention distributions. This readout reuses Query states from the existing forward pass and requires only query–visual-key products, without materializing full sequence-level attention matrices.

Spatial Uncertainty Marginalization. Different agents may attend to the same visual region but exhibit attention peaks at neighboring visual tokens, making direct token-wise comparison sensitive to small spatial shifts. To accommodate such variations, **V-Guide models uncertainty in attention localization through a Gaussian transition kernel** over the discrete visual token grid:

$$\widetilde{p}_i(j) = \sum_{k=1}^{N_v} G_\sigma(j \mid k)\, p_i(k), \qquad G_\sigma(j \mid k) = \frac{\exp\left(-\frac{\|r_j - r_k\|_2^2}{2\sigma^2}\right)}{\sum_{m=1}^{N_v}\exp\left(-\frac{\|r_m - r_k\|_2^2}{2\sigma^2}\right)}. \tag{2}$$

Here, r_j denotes the two-dimensional patch-grid coordinate of visual token j, and σ controls the spatial extent of the uncertainty. The kernel $G_\sigma(j \mid k)$ represents the conditional probability of assigning attention to token j given an original attention location k. This operation redistributes attention to spatially neighboring tokens, allowing nearby attention responses to contribute to cross-agent agreement. We use the spatial grid corresponding to the visual tokens entering the language model and set $\sigma = 1$ in patch-grid units. Since G_σ is a probability-preserving Markov operator shared across agents, it satisfies $\|G_\sigma p - G_\sigma q\|_1 \le \|p - q\|_1, D_{\mathrm{KL}}(G_\sigma p \| G_\sigma q) \le D_{\mathrm{KL}}(p\|q)$. Thus, spatial marginalization introduces local spatial tolerance without increasing the discrepancy between agent-conditioned attention distributions. The corresponding properties are established in Appendix E.2.

Geometric Consensus Distribution. Given the uncertainty-marginalized attention distributions $\{\widetilde{p}_i\}_{i=1}^M$, we seek a consensus distribution that favors visual regions receiving consistent support across agents while downweighting isolated agent-specific responses. We define the **consensus distribution by minimizing the average reverse KL divergence to the agent-conditioned distributions**:

$$\mathbf{c} = \arg\min_{\boldsymbol{\pi}\in\Delta} \frac{1}{M}\sum_{i=1}^{M} D_{\mathrm{KL}}(\boldsymbol{\pi}\|\widetilde{\mathbf{p}}_i), \qquad c_j = \frac{\prod_{i=1}^{M}\widetilde{p}_i(j)^{1/M}}{\sum_{k=1}^{N_v}\prod_{i=1}^{M}\widetilde{p}_i(k)^{1/M}}. \tag{3}$$

Here, Δ denotes the probability simplex over visual tokens. The objective admits the normalized geometric mean as its closed-form solution. Unlike arithmetic averaging, geometric consensus favors visual tokens receiving high attention across agents while downweighting isolated agent-specific responses. The resulting consensus distribution $\mathbf{c}$ captures shared visual focus while accommodating local spatial variations. The derivation is provided in Appendix E.3. We subsequently use this consensus to modulate the shared visual Values (*i.e.*, the Value states in $\mathcal{M}_{\mathrm{vis}}$) during final answering.

Table 1: We compare **V-Guide** with the Single baseline, Debate [38], Reconcile [37], and Text-VMAS and Latent-VMAS under Sequential and Hierarchical collaboration settings.

Methods	MMStar	RealWorldQA	VStar	MMBench	VisuLogic	AVG
Qwen3-VL-4B-Instruct						
Single	63.27	70.46	79.06	69.20	19.00	60.20
Debate	66.60	70.72	78.53	77.93	22.10	63.18
Reconcile	66.07	71.37	80.10	77.85	23.40	63.76
Sequential VMAS						
Text-VMAS	67.13 ↑3.86	69.28 ↓1.18	73.29 ↓5.77	79.02 ↑9.82	**25.90** ↑6.90	62.92 ↑2.72
Latent-VMAS	64.47 ↑1.20	66.14 ↓4.32	75.39 ↓3.67	80.73 ↑11.53	23.50 ↑4.50	62.05 ↑1.85
V-Guide	**67.93** ↑4.66	**72.55** ↑2.09	**81.68** ↑2.62	**81.97** ↑12.77	**25.90** ↑6.90	**66.01** ↑5.81
Hierarchical VMAS						
Text-VMAS	64.87 ↑1.60	71.63 ↑1.17	76.96 ↓2.10	75.00 ↑5.80	24.30 ↑5.30	62.55 ↑2.35
Latent-VMAS	63.07 ↓0.20	66.80 ↓3.66	77.49 ↓1.57	74.54 ↑5.34	24.00 ↑5.00	61.18 ↑0.98
V-Guide	**70.67** ↑7.40	**72.42** ↑1.96	**80.63** ↑1.57	**79.74** ↑10.54	**25.60** ↑6.60	**65.81** ↑5.61
Qwen3-VL-8B-Instruct						
Single	67.07	71.11	83.77	70.59	22.50	63.01
Debate	68.40	72.55	82.72	80.23	24.60	65.70
Reconcile	68.33	71.90	83.77	80.80	24.80	65.92
Sequential VMAS						
Text-VMAS	69.53 ↑2.46	69.80 ↓1.31	80.10 ↓3.67	80.88 ↑10.29	26.10 ↑3.60	65.28 ↑2.27
Latent-VMAS	68.00 ↑0.93	70.59 ↓0.52	79.58 ↓4.19	81.65 ↑11.06	24.80 ↑2.30	64.92 ↑1.91
V-Guide	**72.07** ↑5.00	**73.86** ↑2.75	**85.34** ↑1.57	**84.29** ↑13.70	**26.20** ↑3.70	**68.35** ↑5.34
Hierarchical VMAS						
Text-VMAS	69.73 ↑2.66	72.54 ↑1.43	78.53 ↓5.24	80.80 ↑10.21	25.60 ↑3.10	65.44 ↑2.43
Latent-VMAS	69.47 ↑2.40	70.46 ↓0.65	79.58 ↓4.19	80.96 ↑10.37	24.60 ↑2.10	65.01 ↑2.00
V-Guide	**72.00** ↑4.93	**72.94** ↑1.83	**85.86** ↑2.09	**82.12** ↑11.53	**26.70** ↑4.20	**67.92** ↑4.91

3.4 Consensus-Guided Visual Value Modulation

Although the visual KV cache remains available throughout latent collaboration, the cross-agent visual attention consensus provides an additional signal for identifying visual content consistently attended to across reasoning roles. To exploit this signal during final answer generation, **V-Guide** applies the consensus distribution $\mathbf{c}$ only when the final answering agent operates on the accumulated latent context. Rather than directly modifying the visual Keys, V-Guide modulates only the visual Values in $\mathcal{M}_{\text{vis}}$, reinforcing the content contributed by consensus-supported visual tokens. The **consensus scores are converted into bounded token-wise weights and applied to the corresponding visual Values as**:

$$\widehat{V}^{(l)}_{\text{vis},j} = (1 + \lambda w_j) V^{(l)}_{\text{vis},j}, \qquad w_j = \frac{c_j - c_{\min}}{c_{\max} - c_{\min} + \epsilon}, \qquad \widehat{K}^{(l)}_{\text{vis},j} = K^{(l)}_{\text{vis},j}. \tag{4}$$

where $c_{\min} = \min_k c_k$, $c_{\max} = \max_k c_k$, and $\lambda \geq 0$ controls the modulation strength. Since $w_j \in [0, 1]$, the scaling factor satisfies $1 \leq 1 + \lambda w_j \leq 1 + \lambda$, assigning larger Value contributions to visual tokens with stronger cross-agent consensus while leaving the corresponding Keys unchanged.

We further characterize the local effect of Value-only modulation on an attention computation. For a fixed incoming query, let α_j denote the attention weight of token j and $o = \sum_{j \in \mathcal{T}} \alpha_j V_j$ the original attention output. Conditioned on the same incoming query, leaving the visual Keys unchanged preserves the query–key logits and attention weights of that attention computation, while the modulated output becomes

$$\widehat{o} = o + \lambda \sum_{j \in \mathcal{V}} \alpha_j w_j V_j, \qquad \|\widehat{o} - o\| \leq \lambda B \sum_{j \in \mathcal{V}} \alpha_j \quad \text{if } \|V_j\| \leq B. \tag{5}$$

Hence, for a fixed incoming query, the calibration changes the contribution of visual tokens according to their cross-agent consensus without directly modifying the query–key scores of that attention computation. The induced change is further bounded by both the calibration strength and the total attention weight assigned to visual tokens. Overall, the proposed calibration requires no additional training, introduces no extra visual tokens, and avoids repeated image encoding.

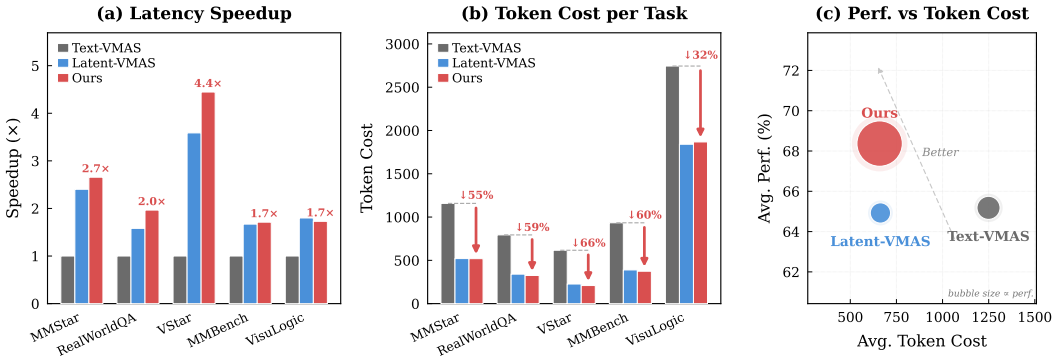


Figure 3: (a) End-to-end speedup over Text-VMAS. (b) Collaboration token cost. (c) Performance–token cost trade-off. `V-Guide` preserves latent collaboration efficiency while achieving the highest overall performance.

4 Experiments

4.1 Experimental Setup

Models and Baselines. We conduct experiments on four MLLM backbones spanning different model families and scales, including Qwen3-VL-4B-Instruct, Qwen3-VL-8B-Instruct, Qwen3-VL-32B-Instruct [1], and InternVL3-8B [2]. We compare `V-Guide` against several representative baselines: (i) **Single**, where a single MLLM directly produces the final response without multi-agent collaboration; (ii) **Debate** [38], where four agents engage in up to two rounds of debate, followed by majority voting; (iii) **Reconcile** [37], where four agents conduct confidence-aware discussions and determine the final answer through confidence-weighted voting; (iv) **Text-VMAS**, where each agent independently encodes the visual input and communicates intermediate reasoning through explicit text; and (v) **Latent-VMAS**, which we instantiate by adapting the latent collaboration paradigm of LatentMAS [18] to visual multi-agent reasoning. The visual input is encoded only once, with the resulting visual KV-Cache shared across agents, while intermediate reasoning is propagated directly through latent states without textual decoding. For both Text-VMAS and Latent-VMAS, we consider two representative agent coordination settings: a **Sequential** structure following the chain-of-agents paradigm [52], and a **Hierarchical** structure in which specialized agents are organized through hierarchical aggregation [53]. Agent prompts are detailed in Appendix F.

Benchmarks. We evaluate on five multimodal benchmarks that cover complementary aspects of visual understanding and reasoning. Specifically, we use MMStar [54] to assess vision-dependent multimodal reasoning, RealWorldQA [55] for real-world spatial reasoning, VStar [56] for fine-grained visual search in high-resolution scenes, VisuLogic [57] for vision-centric logical reasoning, and MMBench [6] for comprehensive multimodal understanding. Together, these benchmarks enable us to evaluate whether multi-agent collaboration can effectively exploit visual evidence across diverse perceptual and reasoning demands.

Evaluation settings. We use VLMEvalKit [58] for benchmark evaluation. For MMBench [6], VLMEvalKit adopts the standard Circular Evaluation protocol [6], which evaluates each question under multiple circular permutations of the answer choices. We use the same decoding configuration for all baselines and our method, adopting greedy decoding with a maximum generation length of 2048 tokens. See Appendix A for more details.

4.2 Main Results

Table 1 presents the main results across MLLM backbones and MAS architectures. Multi-agent collaboration does not consistently outperform single-agent inference, with both Text-VMAS and Latent-VMAS underperforming the Single baseline on VStar and RealWorldQA, two benchmarks requiring fine-grained visual grounding and spatial understanding. While latent collaboration avoids the information bottleneck of text-mediated communication by directly propagating reasoning states and reusing the shared visual KV cache, it does not explicitly exploit cross-agent visual attention consensus. In contrast, `V-Guide` consistently improves upon Latent-VMAS across all five benchmarks and achieves higher average performance than the Single, Debate, Reconcile, and Text-VMAS baselines under both MAS architectures. Specifically, `V-Guide` improves the average score over Latent-VMAS

Table 2: Effect of consensus estimation designs.

Consensus Formulation		Spatial Uncertainty Marginalization	MMStar	RealWorldQA
Arithmetic	Geometric			
✓		✓	71.07	72.94
	✓	✓	**72.07**	**73.86**
	✓	✗	70.53	69.80

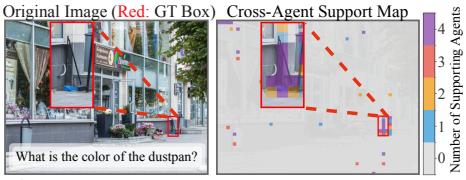


Figure 4: Cross-agent visual attention support.

Table 3: Consensus-Based Visual Token Removal.

Removed Tokens	MMStar	RealWorldQA	VStar
None (Latent-VMAS)	**68.00**	**70.59**	**79.58**
Top 25%	59.87	65.10	68.06
Bottom 25%	66.67	68.36	77.49
Random 25%	65.93	67.71	72.80

Table 4: Effect of visual Value modulation strategies.

Modulation Strategy	MMStar	RealWorldQA	VStar
None (Latent-VMAS)	68.00	70.59	79.58
Uniform Weighting	70.13	70.33	79.58
Final-Agent Attention	70.33	70.33	83.77
Cross-Agent Consensus	**72.07**	**73.86**	**85.34**

by 3.96 and 4.63 percentage points on Qwen3-VL-4B under Sequential and Hierarchical collaboration, respectively. These results demonstrate the effectiveness of leveraging cross-agent visual attention consensus to enhance latent multi-agent collaboration beyond reasoning-state propagation.

Efficiency Analysis. We evaluate the inference efficiency of different collaboration paradigms on Qwen3-VL-8B-Instruct [1] using two complementary metrics: **inference speedup** relative to Text-VMAS, calculated as the ratio of cumulative inference latency over all samples in each benchmark, and **token cost**, measured by the average number of generated tokens per sample, including intermediate textual or latent reasoning tokens and final answer tokens. As shown in Fig. 3(a), `V-Guide` achieves speedups of 1.7×–4.4× over Text-VMAS across the five benchmarks, while maintaining comparable inference efficiency to Latent-VMAS. Fig. 3(b) further shows that `V-Guide` reduces token cost by 32%–66% relative to Text-VMAS, while requiring a similar number of tokens to Latent-VMAS. Finally, Fig. 3(c) demonstrates that `V-Guide` achieves the highest average task performance among the evaluated methods while retaining the efficiency advantages of latent collaboration.

4.3 Why Cross-Agent Visual Attention Consensus for Visual Modulation?

Visual Regions Highly Attended by Multiple Agents Align with Task-Relevant Regions. Fig. 4 visualizes cross-agent visual attention support with the ground-truth (GT) target region highlighted by a red box. The number of supporting agents indicates how many agents rank a given visual token among their top 5% most-attended regions. Visual regions receiving high attention from multiple agents tend to overlap with the GT task-relevant region, motivating consensus-guided modulation to emphasize such visual content during final answer generation. See Fig. 7 for more visualizations.

Do High-Consensus Visual Tokens Matter for Final Predictions? To examine whether visual tokens with higher cross-agent consensus scores play a greater role in final predictions, we conduct a token removal experiment using Qwen3-VL-8B-Instruct [1]. Specifically, we rank visual tokens by their consensus scores c_j from Eq. 3 and separately remove the top 25%, bottom 25%, or a randomly selected 25% of tokens. As shown in Table 3, removing the top-25% consensus tokens leads to substantially larger performance drops across all three benchmarks than removing the bottom-25% or randomly selected tokens. These results suggest that visual tokens with higher consensus scores contain information on which the model relies more heavily for final predictions.

Does Cross-Agent Consensus Better Guide Visual Value Modulation? To assess the role of cross-agent consensus in visual Value modulation, we compare different strategies for assigning the token-wise modulation weights w_j in Eq. 4, while keeping modulation strength λ fixed. **Uniform Weighting** assigns identical weights to all visual tokens, while **Final-Agent Attention** derives w_j solely from the final agent's visual attention, instead of the cross-agent consensus distribution $\mathbf{c}$ used by `V-Guide`. As shown in Table 4, Uniform Weighting fails to reproduce the gains of `V-Guide`, indicating that uniform amplification alone is insufficient. Moreover, consensus-guided modulation consistently outperforms Final-Agent Attention across all three benchmarks, demonstrating the additional benefit of leveraging cross-agent visual attention to guide visual Value modulation.

4.4 Ablation Study

Generalization across MLLM Backbones. We further evaluate `V-Guide` on Qwen3-VL-32B-Instruct [1] and InternVL3-8B [2] under the Sequential VMAS setting. As shown in Table 5, `V-Guide`

Table 5: Performance comparison across different MLLM backbones.

Methods	MMStar	RealWorldQA	VStar	MMBench	VisuLogic	AVG
Qwen3-VL-32B-Instruct						
Single	68.33	74.64	85.86	72.58	24.80	65.24
Text-VMAS	72.87	75.69	84.82	87.31	24.80	69.10
Latent-VMAS	72.33	76.47	84.82	87.54	22.00	68.63
`V-Guide`	**73.27**	**78.43**	**87.96**	**88.19**	**27.10**	**70.99**
InternVL3-8B						
Single	63.80	64.18	63.87	51.01	27.10	53.99
Text-VMAS	68.80	64.97	71.20	80.73	24.40	62.02
Latent-VMAS	62.20	67.32	62.83	63.85	23.00	55.84
`V-Guide`	**69.20**	**69.28**	**72.25**	**81.65**	**28.60**	**64.20**

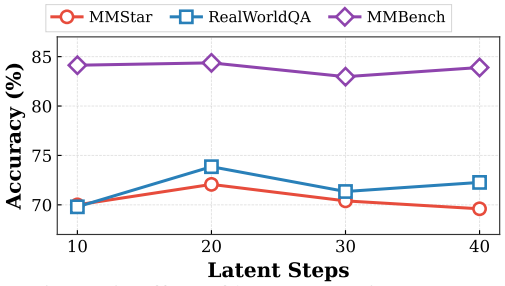


Figure 5: Effect of latent reasoning steps.

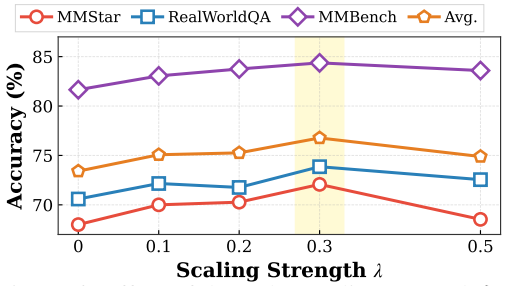


Figure 6: Effect of the Value scaling strength λ.

achieves the highest average performance on both backbones, demonstrating the effectiveness of `V-Guide` across different MLLM architectures and scales.

Sensitivity to Scaling Strength. We examine the effect of the visual Value scaling strength λ in Eq. 4, where $\lambda = 0$ corresponds to no modulation. As shown in Fig. 6, performance improves as λ increases to 0.3, but declines at 0.5. This suggests that moderate consensus-guided modulation enhances the contribution of visual information, whereas excessive scaling can degrade performance. We therefore set $\lambda = 0.3$ throughout our experiments.

Effect of Latent Steps. We examine the effect of varying the number of latent reasoning steps from 10 to 40. As shown in Fig. 5, performance improves from 10 to 20 steps, peaking at 20 across all three benchmarks, but declines moderately with further increases. This suggests that additional latent computation does not necessarily improve performance and may introduce redundant intermediate states. We therefore use 20 latent steps throughout our experiments.

Effect of Cross-Agent Visual Consensus Estimation Designs. To derive a consensus distribution that accommodates local spatial variations while emphasizing consistent cross-agent visual attention support, `V-Guide` introduces Spatial Uncertainty Marginalization and geometric consensus. We evaluate both designs in Table 2. First, we replace geometric consensus with a simple arithmetic average of the spatially marginalized visual attention distributions. Geometric consensus outperforms arithmetic averaging, suggesting that emphasizing visual tokens receiving high attention from multiple agents is more effective than simply averaging individual attention distributions. Second, removing Spatial Uncertainty Marginalization leads to performance degradation on both benchmarks. This highlights the importance of accommodating local spatial variations, as different agents may attend to the same visual region while assigning high attention to neighboring tokens. Together, these results support the effectiveness of both designs in cross-role consensus estimation.

5 Conclusion

In this work, we introduced `V-Guide`, a training-free latent collaboration framework that extends visual multi-agent reasoning beyond reasoning-state propagation through explicit cross-agent visual coordination. By consolidating role-conditioned visual attention distributions into a spatially tolerant geometric consensus, `V-Guide` guides the modulation of shared visual Values to enhance the contribution of consensus-supported visual content during final answer generation. Extensive experiments demonstrate that `V-Guide` consistently improves visual reasoning performance while retaining the efficiency advantages of latent collaboration. Our findings highlight the importance of coordinating visual attention alongside reasoning states, motivating further exploration of richer visual communication mechanisms in latent multi-agent systems.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language proofreading, improve the clarity of the manuscript, and support the retrieval and discovery of related work. Generative AI tools were also used to assist with parts of the coding process. We did not use generative AI tools to generate experimental results, perform data analysis, or draw scientific conclusions. All AI-assisted content was carefully reviewed and verified by the authors. We take full responsibility for the final content of this manuscript.

REFERENCES

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 1, 7, 8

[2] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 1, 7, 8

[3] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. On scaling up a multilingual vision and language model. In *CVPR*, pages 14432–14444, 2024. 1

[4] Xinlei Yu, Zhangquan Chen, Yongbo He, Tianyu Fu, Cheng Yang, Chengming Xu, Yue Ma, Xiaobin Hu, Zhe Cao, Jie Xu, et al. The latent space: Foundation, evolution, mechanism, ability, and outlook. *arXiv preprint arXiv:2604.02029*, 2026. 1

[5] Zheng Nie, Ruolin Shen, Xinlei Yu, Bo Yin, Jiangning Zhang, and Xiaobin Hu. Skillgraph: Self-evolving multi-agent collaboration with multimodal graph topology. *arXiv preprint arXiv:2604.17503*, 2026. 1

[6] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 1, 7

[7] Zeqing Wang, Wentao Wan, Qiqing Lao, Runmeng Chen, Minjie Lang, Xiao Wang, Keze Wang, and Liang Lin. Towards top-down reasoning: An explainable multi-agent approach for visual question answering. *arXiv preprint arXiv:2311.17331*, 2023. 1, 3

[8] Mohammed Elhenawy, Ahmad Abutahoun, Taqwa I Alhadidi, Ahmed Jaber, Huthaifa I Ashqar, Shadi Jaradat, Ahmed Abdelhay, Sebastien Glaser, and Andry Rakotonirainy. Visual reasoning and multi-agent approach in multimodal large language models (mllms): Solving tsp and mtsp combinatorial challenges. *arXiv preprint arXiv:2407.00092*, 2024. 1, 3

[9] James Y Huang, Sheng Zhang, Qianchu Liu, Guanghui Qin, Tinghui Zhu, Tristan Naumann, Muhao Chen, and Hoifung Poon. Be my eyes: Extending large language models to new modalities through multi-agent collaboration. *arXiv preprint arXiv:2511.19417*, 2025. 1, 3

[10] Lixiang Zhang, Jingchen Li, Yi'an Zhu, Haobin Shi, and Kao-Shing Hwang. Multi-agent reinforcement learning by the actor-critic model with an attention interface. *Neurocomputing*, 471:275–284, 2022. 1, 3

[11] Jusheng Zhang, Kaitong Cai, Yijia Fan, Ningyuan Liu, and Keze Wang. MAT-agent: Adaptive multi-agent training optimization. In *The Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025. 1, 3

[12] Yiran Wang and Masao Utiyama. To be continuous, or to be discrete, those are bits of questions. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8036–8049, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.436. URL https://aclanthology.org/2024.acl-long.436/. 1, 3

[13] Xinlei Yu, Chengming Xu, Guibin Zhang, Yongbo He, Zhangquan Chen, Zhucun Xue, Jiangning Zhang, Yue Liao, Xiaobin Hu, Yu-Gang Jiang, et al. Visual multi-agent system: Mitigating hallucination snowballing via visual flow. In *International Conference on Learning Representations*, volume 2026, pages 43581–43606, 2026. 1, 3

[14] Peter West, Ari Holtzman, Jan Buys, and Yejin Choi. BottleSum: Unsupervised and self-supervised sentence summarization using the information bottleneck principle. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3752–3761, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1389. URL https://aclanthology.org/D19-1389/. 1, 3

[15] Zihan Huang, Junda Wu, Rohan Surana, Raghav Jain, Tong Yu, Raghavendra Addanki, David Arbour, Sungchul Kim, and Julian McAuley. Traceable and explainable multimodal large language models: An information-theoretic view. In *Second Conference on Language Modeling (CoLM)*, 2025. 1, 3

[16] Xiaoze Liu, Ruowang Zhang, Weichen Yu, Siheng Xiong, Liu He, Feijie Wu, Hoin Jung, Matt Fredrikson, Xiaoqian Wang, and Jing Gao. The vision wormhole: Latent-space communication in heterogeneous multi-agent systems. *arXiv preprint arXiv:2602.15382*, 2026. 1

[17] Haibo Jin, Peng Kuang, Ye Yu, Xiaopeng Yuan, and Haohan Wang. Agent primitives: Reusable latent building blocks for multi-agent systems. *arXiv preprint arXiv:2602.03695*, 2026. 1

[18] Jiaru Zou, Ruizhong Qiu, Gaotang Li, Xiyuan Yang, Katherine Tieu, Pan Lu, Ke Shen, Hanghang Tong, Yejin Choi, Jingrui He, James Zou, Mengdi Wang, and Ling Yang. Latent collaboration in multi-agent systems. In *International Conference on Machine Learning*, 2026. 1, 3, 4, 7

[19] Luyang Liu, Jonas Pfeiffer, Jiaxing Wu, Jun Xie, and Arthur Szlam. Deliberation in latent space via differentiable cache augmentation. *arXiv preprint arXiv:2412.17747*, 2024. 1

[20] Yujia Zheng, Zhuokai Zhao, Zijian Li, Yaqi Xie, Mingze Gao, Lizhu Zhang, and Kun Zhang. Thought communication in multiagent collaboration. *arXiv preprint arXiv:2510.20733*, 2025. 1, 3

[21] Tianyu Fu, Zihan Min, Hanling Zhang, Jichao Yan, Guohao Dai, Wanli Ouyang, and Yu Wang. Cache-to-cache: Direct semantic communication between large language models. *arXiv preprint arXiv:2510.03215*, 2025. 1, 3

[22] Kshitij Sharma, Jennifer K Olsen, Himanshu Verma, Daniela Caballero, and Patrick Jermann. Challenging joint visual attention as a proxy for collaborative performance. In *Proceedings of the 14th International Conference on Computer-Supported Collaborative Learning-CSCL 2021*. International Society of the Learning Sciences, 2021. 2

[23] Kshitij Sharma and Jennifer K Olsen. What brings students together?: Investigating the causal relationship between joint mental effort and joint visual attention. In *Proceedings of the 15th International Conference on Computer-Supported Collaborative Learning-CSCL 2022, pp. 131-138*. International Society of the Learning Sciences, 2022. 2

[24] Jessica FA Salminen-Saari, Enrique Garcia Moreno-Esteva, Eeva Haataja, Miika Toivanen, Markku S Hannula, and Anu Laine. Phases of collaborative mathematical problem solving and joint attention: a case study utilizing mobile gaze tracking. *ZDM–Mathematics Education*, 53 (4):771–784, 2021. 2

[25] Daniel C Richardson, Rick Dale, and Natasha Z Kirkham. The art of conversation is coordination. *Psychological science*, 18(5):407–413, 2007. 2

[26] Lucas Haraped, D Jacob Gerlofs, Olive Chung-Hui Huang, Cam Hickling, Walter F Bischof, Pierre Sachse, and Alan Kingstone. Coordinating attention in face-to-face collaboration: The dynamics of gaze, pointing, and verbal reference. *Cognitive Science*, 49(10):e70123, 2025. 2

[27] Wanjia Zhao, Mert Yuksekgonul, Shirley Wu, and James Zou. Sirius: Self-improving multi-agent systems via bootstrapped reasoning. *Advances in Neural Information Processing Systems*, 38:124475–124504, 2026. 3

[28] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022. 3

[29] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024. 3

[30] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: communicative agents for "mind" exploration of large language model society. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc. 3

[31] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Steven Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, volume 2024, pages 23247–23275, 2024. 3

[32] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023. 3

[33] Yingxuan Yang, Qiuying Peng, Jun Wang, Ying Wen, and Weinan Zhang. Llm-based multi-agent systems: Techniques and business perspectives. *arXiv preprint arXiv:2411.14033*, 2024. 3

[34] Wei Tao, Yucheng Zhou, Yanlin Wang, Wenqiang Zhang, Hongyu Zhang, and Yu Cheng. Magis: Llm-based multi-agent framework for github issue resolution. *Advances in Neural Information Processing Systems*, 37:51963–51993, 2024. 3

[35] Zhao Wang, Sota Moriyama, Wei-Yao Wang, Briti Gangopadhyay, and Shingo Takamatsu. Talk structurally, act hierarchically: A collaborative framework for llm multi-agent systems. *arXiv preprint arXiv:2502.11098*, 2025. 3

[36] Jiaxing Zhao, Hongbin Xie, Yuzhen Lei, Xuan Song, Zhuoran Shi, Lianxin Li, Shuangxue Liu, Linguo Xie, and Haoran Zhang. Cochain: Balancing insufficient and excessive collaboration in llm agent workflows, 2026. URL https://arxiv.org/abs/2505.10936. 3

[37] Justin Chen, Swarnadeep Saha, and Mohit Bansal. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085, 2024. 3, 6, 7

[38] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023. 3, 6, 7

[39] Zhenglun Kong, Yize Li, Fanhu Zeng, Lei Xin, Shvat Messica, Xue Lin, Pu Zhao, Manolis Kellis, Hao Tang, and Marinka Zitnik. Token reduction should go beyond efficiency in generative models – from vision, language to multimodality, 2026. URL https://arxiv.org/abs/2505.18227. 3

[40] Elizabeth Mieczkowski, Ruaridh Mon-Williams, Neil Bramley, Christopher G Lucas, Natalia Velez, and Thomas L Griffiths. Predicting multi-agent specialization via task parallelizability. *arXiv preprint arXiv:2503.15703*, 2025. 3

[41] Qihe Huang, Zhengyang Zhou, Yangze Li, Kuo Yang, Binwu Wang, and Yang Wang. Many minds, one goal: Time series forecasting via sub-task specialization and inter-agent cooperation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=Uon41HfqR3. 3

[42] Hancheng Ye, Zhengqi Gao, Mingyuan Ma, Qinsi Wang, Yuzhe Fu, Ming-Yu Chung, Yueqian Lin, Zhijian Liu, Jianyi Zhang, Danyang Zhuo, et al. Kvcomm: Online cross-context kv-cache communication for efficient llm-based multi-agent systems. *arXiv preprint arXiv:2510.12872*, 2025. 3

[43] Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*, 2025. 3

[44] Julian Coda-Forno, Zhuokai Zhao, Qiang Zhang, Dipesh Tamboli, Weiwei Li, Xiangjun Fan, Lizhu Zhang, Eric Schulz, and Hsiao-Ping Tseng. Exploring system 1 and 2 communication for latent reasoning in llms. *arXiv preprint arXiv:2510.00494*, 2025. 3

[45] Zhengyang Li, Qijin Ji, Xinghong Ling, and Quan Liu. A comprehensive review of multi-agent reinforcement learning in video games. *IEEE Transactions on Games*, 2025. 3

[46] Longyan Tan, Fanping Wei, Xiaobing Ma, Rui Peng, Hui Xiao, and Li Yang. Systemic condition-based maintenance optimization under inspection uncertainties: A customized multiagent reinforcement learning approach. *IEEE Transactions on Reliability*, 2025. 3

[47] Yuhan Wang, Shuochen Chang, Yalin Feng, Dongsheng Ma, Yuanzi Li, Zhengren Wang, Yinglong Yang, Yufei Chen, Yikang Wang, Shaoxu Sun, et al. Seeing before agreeing: Aligning multi-agent consensus with visual evidence. *arXiv preprint arXiv:2605.30698*, 2026. 3

[48] Xinlei Yu, Chengming Xu, Zhangquan Chen, Bo Yin, Cheng Yang, Yongbo He, Yihao Hu, Jiangning Zhang, Cheng Tan, Xiaobin Hu, et al. Dual latent memory for visual multi-agent system. *arXiv preprint arXiv:2602.00471*, 2026. 3

[49] Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14558. IEEE, 2025. 5, 20

[50] Rinyoichi Takezoe, Yaqian Li, Zi-Hao Bo, Anzhou Hou, Mo Guang, and Kaiwen Long. Learn-pruner: Rethinking attention-based token pruning in vision language models. In *International Conference on Learning Representations*, volume 2026, pages 66381–66400, 2026. 5, 20

[51] Yuyao Sun, Tao Deng, Shuang Li, Deqing Wang, Hao Geng, and Minjun Yu. Learning to predict middle-layer attention in mllms for visual token pruning. *arXiv preprint arXiv:2608.06411*, 2026. 5, 20

[52] Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Ö Arık. Chain of agents: Large language models collaborating on long-context tasks. *Advances in Neural Information Processing Systems*, 37:132208–132237, 2024. 7

[53] Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. Language agents as optimizable graphs. *arXiv preprint arXiv:2402.16823*, 2024. 7

[54] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024. 7

[55] xAI. Realworldqa: A benchmark for real-world spatial understanding. https://huggingface.co/datasets/xai-org/RealworldQA, 2024. Accessed: 2025-04-26. 7

[56] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13084–13094. IEEE, 2024. 7, 16, 18

[57] Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. In *International Conference on Learning Representations*, volume 2026, pages 25966–26003, 2026. 7

[58] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201, 2024. 7

APPENDIX

The document is organized as follows:

- §A Experimental Details.
- §B Analysis of Cross-Agent Visual Attention Support.
- §C Visualizations of Cross-Agent Visual Attention Support.
- §D Why Extract Visual Attention from Middle Layers?
- §E Derivation and Properties of Cross-Agent Visual Consensus.
- §F Multi-Agent Role Prompts.
- §G Limitations

A EXPERIMENTAL DETAILS

Unless otherwise specified, we set the number of latent reasoning steps to $T = 20$, the visual modulation strength to $\lambda = 0.3$, and the spatial kernel bandwidth to $\sigma = 1$. We extract visual attention from middle decoder layers, selecting approximately 40%–70% of the total layer depth for each backbone. Specifically, we use layers 15–25 for Qwen3-VL-4B-Instruct and Qwen3-VL-8B-Instruct (36 layers), layers 27–44 for Qwen3-VL-32B-Instruct (64 layers), and layers 12–19 for InternVL3-8B (28 layers). Unless otherwise specified, we apply Consensus-Guided Visual Value Modulation to the same selected layers and adopt Sequential MAS as the default collaboration architecture.

Uniform Weighting Baseline. To isolate the effect of token-wise consensus guidance from the overall magnitude of Value modulation, we match the average modulation strength of `V-Guide`. Specifically, given the `V-Guide` weights $\{w_j\}_{j=1}^{N_v}$ for an input sample, we compute

$$\bar{w} = \frac{1}{N_v} \sum_{j=1}^{N_v} w_j, \tag{6}$$

and assign $w_j^{\text{uni}} = \bar{w}$ to every visual token. The uniformly modulated Values are therefore

$$\widehat{V}_{\text{vis},j}^{(l)} = (1 + \lambda \bar{w}) V_{\text{vis},j}^{(l)}. \tag{7}$$

This baseline has the same average Value scaling factor as `V-Guide` for each sample, but discards the token-wise variation induced by cross-agent visual consensus.

Implementation of Visual Token Removal. For the token removal experiment in § 4.3, we implement token removal through attention masking rather than physically deleting the corresponding KV-cache entries. Specifically, the selected visual tokens are masked during final answer generation, preventing the answering agent from attending to them. This preserves the original token positions and KV-cache structure while isolating the effect of removing the selected visual tokens from the attention context.

Compatibility of the Shared Visual KV Cache. Our shared visual KV cache is reused across agents under a homogeneous-agent setting, where all agents use the same VLM backbone. In particular, the visual tokens are always placed at the beginning of the input sequence and occupy the same positions for every agent. The corresponding visual Keys and Values are therefore computed once with identical model parameters and position indices, and are subsequently reused by all agents. Agent-specific role instructions and reasoning states are appended after this fixed visual prefix and hence do not alter the cached visual states. This design also ensures compatibility with positional encoding such as RoPE, since the cached visual Keys are associated with the same position IDs across agents.

We intentionally adopt homogeneous agents because our focus is on improving latent visual collaboration rather than aligning internal representations across heterogeneous models. Heterogeneous agents may differ in hidden dimensions, layer structures, visual tokenization, multimodal templates,

and positional encodings, making direct KV sharing generally incompatible and potentially requiring additional learned projections or alignment modules. Such mechanisms would introduce a separate research problem and depart from the training-free nature of our current framework. We therefore leave heterogeneous latent collaboration as an interesting direction for future work. Our method does not assume that KV states can be directly shared across different model architectures.

B Analysis of Cross-Agent Visual Attention Support

We provide additional quantitative analyses to further examine the relationship between cross-agent visual attention support and ground-truth (GT) target regions on VStar [56]. All analyses in this section are conducted on the complete VStar evaluation set using Qwen3-VL-8B-Instruct. Importantly, the analysis directly uses the agent-conditioned visual attention distributions defined in Eq. 1, *i.e.*, the same visual attention readout employed by **`V-Guide`**. Therefore, Fig. 1(a) directly examines whether the cross-agent visual agreement signal exploited by **`V-Guide`** is associated with task-relevant visual regions.

B.1 Analysis Protocol

For each agent A_i, Eq. 1 constructs an agent-conditioned visual attention distribution p_i over the shared visual tokens. Specifically, **`V-Guide`** mean-pools the Query states of the textual prefix, computes a conditional visual attention distribution over the shared visual Keys, and aggregates the resulting distributions across attention heads and the selected middle decoder layers. Consistent with the default setting of **`V-Guide`**, we use decoder layers 15–25.

As derived in Appendix E.1, for an individual query position, normalizing its Query–Key scores only over the visual-token subset is exactly equivalent to conditioning its native full-sequence attention distribution on the visual tokens. Mean-pooling the textual-prefix Query states further yields the normalized geometric aggregation of these token-level conditional visual attention distributions. Therefore, p_i provides a compact agent-level characterization of the relative visual locations prioritized by agent A_i within the shared visual representation.

For each image x and agent A_i, we identify the top-5% visual tokens according to $p_{x,i}$. For visual token j, we define its cross-agent support count as

$$n_{x,j} = \sum_{i=1}^{M} \mathbb{I}\left[j \in \mathrm{Top}_{5\%}(p_{x,i})\right], \qquad n_{x,j} \in \{0, \ldots, M\}, \tag{8}$$

where $M = 4$. Thus, $n_{x,j}$ measures how many agents jointly prioritize visual token j under the same visual attention readout used by **`V-Guide`**.

The GT target bounding box is mapped to the visual-token grid according to the image preprocessing used by the model. A visual token is considered a GT-region hit if its corresponding spatial location falls within the mapped GT target region. Let $h_{x,j} \in \{0, 1\}$ denote the resulting GT-hit indicator.

B.2 Exact Cross-Agent Attention Support

We first examine visual tokens associated with each exact support count. For image x, the GT-region hit rate for visual tokens supported by exactly k agents is defined as

$$H_x(k) = \frac{\sum_{j \in \mathcal{V}_x} \mathbb{I}[n_{x,j} = k]\, h_{x,j}}{\sum_{j \in \mathcal{V}_x} \mathbb{I}[n_{x,j} = k]}, \tag{9}$$

where $\mathcal{V}_x$ denotes the visual-token set of image x.

To handle the case where an image contains no visual token with a particular exact support count, we define

$$\mathcal{X}_k = \left\{ x : \sum_{j \in \mathcal{V}_x} \mathbb{I}[n_{x,j} = k] > 0 \right\}. \tag{10}$$

We then macro-average the image-level hit rates over the valid images:

$$H(k) = \frac{1}{|\mathcal{X}_k|} \sum_{x \in \mathcal{X}_k} H_x(k). \tag{11}$$

This image-level macro averaging assigns equal weight to each evaluated example, independently of the number of visual tokens in the image, and corresponds to the quantity reported in Fig. 1(a).

Table 6: Exact cross-agent visual attention support under the top-5% criterion, corresponding to Fig. 1(a). GT Hit Rate is first computed within each image and then macro-averaged across the complete VStar evaluation set. We report 95% confidence intervals obtained by image-level bootstrap.

Support k	GT Hit Rate (%)	95% CI (%)
0	0.080	[0.046, 0.122]
1	0.606	[0.246, 1.032]
2	0.479	[0.156, 0.886]
3	0.966	[0.387, 1.692]
4	4.859	[4.031, 5.759]

As shown in Table 6, visual tokens receiving cross-agent attention support generally exhibit higher GT-region hit rates than unsupported tokens. Most notably, visual tokens jointly prioritized by all four agents achieve the highest GT-region hit rate, reaching 4.86%, compared with 0.08% for unsupported tokens. For the all-agent support group, the size-constrained oracle, defined as the maximum attainable GT-region hit rate under the same support-set size and GT-region size, is 7.43%. The intermediate exact-support groups are not strictly monotonic; therefore, we do not interpret the exact support count as inducing a monotonic ordering of GT-region alignment.

B.3 Cumulative Cross-Agent Attention Support

The exact-support analysis above characterizes visual tokens associated with each individual support count. We further consider a cumulative criterion that retains all visual tokens supported by at least k agents. This analysis examines whether imposing an increasingly stringent minimum level of cross-agent agreement progressively concentrates the selected visual tokens within GT target regions.

For image x, we define the cumulative GT-region hit rate at minimum support level k as

$$H_x(\geq k) = \frac{\sum_{j \in \mathcal{V}_x} \mathbb{I}[n_{x,j} \geq k]\, h_{x,j}}{\sum_{j \in \mathcal{V}_x} \mathbb{I}[n_{x,j} \geq k]}. \tag{12}$$

Similarly, we define the valid image set

$$\mathcal{X}_{\geq k} = \left\{ x : \sum_{j \in \mathcal{V}_x} \mathbb{I}[n_{x,j} \geq k] > 0 \right\}, \tag{13}$$

and compute the image-level macro average as

$$H(\geq k) = \frac{1}{|\mathcal{X}_{\geq k}|} \sum_{x \in \mathcal{X}_{\geq k}} H_x(\geq k). \tag{14}$$

As shown in Table 7, the GT-region hit rate increases consistently as the minimum required number of supporting agents increases. Specifically, it increases from 2.99% for visual tokens supported by at least one agent to 3.50%, 4.08%, and 4.86% when requiring support from at least two, three, and four agents, respectively. Thus, although the intermediate exact-support groups are not strictly monotonic, progressively stronger minimum cross-agent support results in a consistent increase in GT-region concentration.

Table 7: Cumulative cross-agent visual attention support under the top-5% criterion. GT Hit Rate is first computed within each image and then macro-averaged across the complete VStar evaluation set.

Minimum Support	GT Hit Rate (%)
≥ 1	2.994
≥ 2	3.502
≥ 3	4.076
≥ 4	4.859

B.4 Robustness to the Attention-Support Threshold

The top-5% criterion used in Fig. 1(a) provides a fixed-budget definition of strongly attended visual tokens. To examine whether the observed relationship depends on this particular choice, we repeat the same image-level macro analysis using the top-10%, top-15%, and top-20% visual tokens according to each agent's distribution p_i.

Table 8 reports the exact-support GT-region hit rates under the four attention thresholds. Across all settings, visual tokens jointly supported by all four agents consistently achieve the highest GT-region hit rate, whereas unsupported tokens exhibit the lowest rate. Although the intermediate support levels are not always monotonically ordered, the substantially stronger GT-region alignment of the all-agent support group persists across all tested thresholds. This indicates that the observation in Fig. 1(a) is not specific to the particular top-5% criterion.

Table 8: Robustness of exact cross-agent visual attention support to different attention thresholds. Each entry reports the image-level macro-averaged GT-region hit rate $H(k)$ in percentage. All attention scores are obtained from the agent-conditioned visual attention distributions in Eq. 1.

Attention Criterion	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$
Top-5%	0.080	0.606	0.479	0.966	4.859
Top-10%	0.059	0.384	0.175	0.247	2.743
Top-15%	0.048	0.162	0.300	0.192	1.912
Top-20%	0.042	0.103	0.143	0.181	1.475

C Visualizations of Cross-Agent Visual Attention Support

We provide additional visualizations from VStar [56] to illustrate the correspondence between cross-agent visual attention support and ground-truth (GT) target regions. The quantitative analysis in Fig. 1(a) is conducted on VStar using Qwen3-VL-8B-Instruct. VStar provides GT bounding boxes for target objects, which are highlighted in red in the original images. Given the high resolution of VStar images, we also present enlarged views of the GT regions for clearer visualization of local attention patterns. For each example, we visualize a cross-agent support map, where the number of supporting agents indicates how many agents rank a given visual region among their top 5% most-attended regions. As illustrated in Fig. 7, visual regions jointly attended to across agents frequently overlap with the GT regions, providing further qualitative support for leveraging cross-agent visual attention consensus to guide visual modulation during final answer generation.

D Why Extract Visual Attention from Middle Layers?

D.1 Question-Relevant Visual Localization across Decoder Layers

To investigate how visual attention varies across decoder layers, we analyze the localization of question-relevant visual regions for each agent on VStar [56]. VStar provides ground-truth bounding boxes for target objects, enabling us to examine the spatial alignment between visual attention and question-relevant regions. Since these targets are often small, we adopt a region-level hit criterion rather than requiring precise overlap between attention regions and ground-truth bounding boxes.

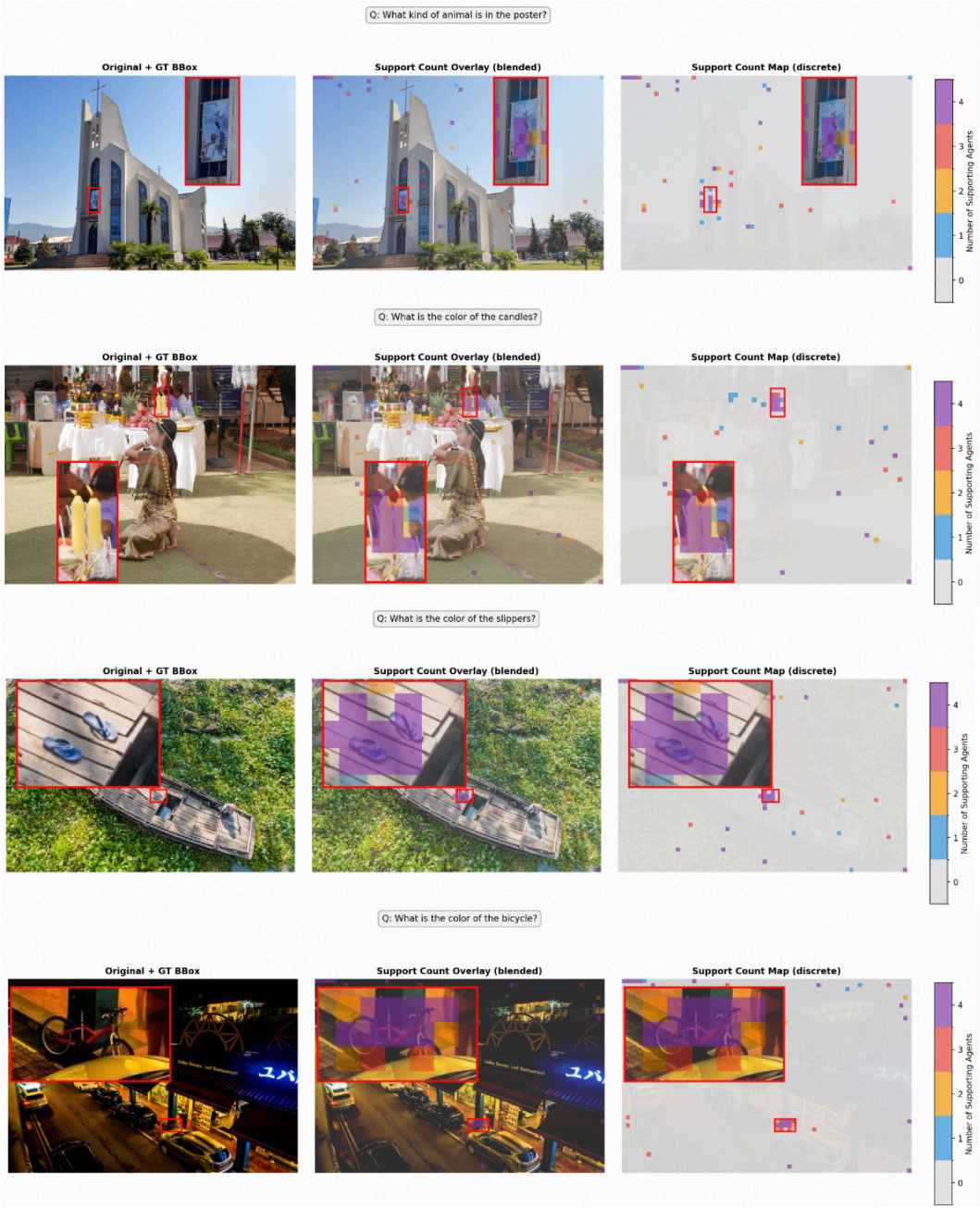


Figure 7: **Additional visualizations of cross-agent visual attention support on VStar.** Red bounding boxes indicate the ground-truth (GT) regions of target objects. Given the high resolution of VStar images, we provide enlarged views of the GT regions for clearer visualization. The cross-agent support maps show the number of agents ranking each visual region among their top 5% most-attended regions. Visual regions jointly attended to across agents frequently overlap with the GT regions.

Specifically, we evaluate ROI Recall@K for $K \in \{1, 3, 5\}$. For each decoder layer, we identify connected components in the visual attention map and select the top-K regions according to their attention scores. A sample is considered successfully localized if at least one selected region intersects the ground-truth bounding box. Formally, ROI Recall@K at layer l is defined as

$$\text{ROIRecall@K}^{(l)} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{1} \left\{ \exists\, C \in \mathcal{C}_{n,K}^{(l)} \text{ s.t. } C \cap B_n^{\text{GT}} \neq \emptyset \right\}. \tag{15}$$

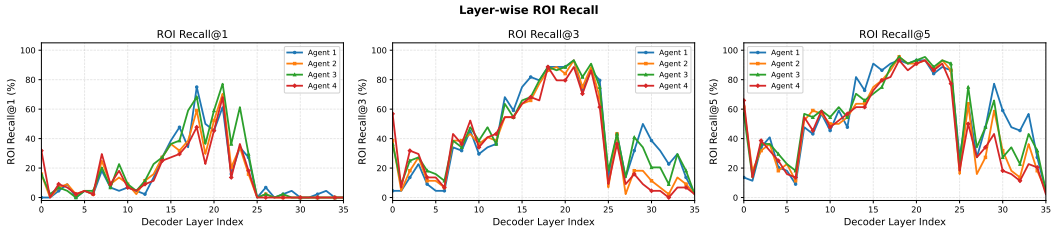


Figure 8: **Layer-wise ROI Recall@K across four agents on VStar.** We report ROI Recall@K for $K \in \{1, 3, 5\}$ at different decoder layers.

Table 9: **Effect of layer selection for visual attention extraction.** We compare different layer ranges for extracting role-conditioned visual attention patterns on Qwen3-VL-8B-Instruct under Sequential MAS. None (Latent-VMAS) denotes the original latent collaboration framework without visual attention extraction and coordination.

Layer Selection	MMStar	RealWorldQA	VisuLogic	VStar	MMBench	AVG
None (Latent-VMAS)	68.00	70.59	24.80	79.58	81.65	64.92
Early Layers (0–10)	68.00	70.31	23.10	83.77	83.36	65.71
Mid Layers (15–25)	**72.07**	**73.86**	**26.20**	**85.34**	**84.29**	**68.35**
Late Layers (25–35)	71.00	71.90	23.50	84.29	83.36	66.81
All Layers (0–35)	68.80	71.63	23.80	84.29	83.05	66.31

where N denotes the number of evaluated samples, $\mathcal{C}_{n,K}^{(l)}$ contains the top-K connected regions extracted from the attention map of sample n at layer l, and B_n^{GT} denotes its ground-truth bounding box. The metric measures whether the highly attended regions intersect the target object, without requiring precise boundary alignment.

As shown in Fig. 8, middle decoder layers consistently exhibit higher ROI Recall@K than early and late layers across all four agents and different values of K. This suggests that middle-layer attention more reliably localizes question-relevant visual regions, despite the different reasoning roles of individual agents. The consistent trend across different values of K further indicates that this observation is not specific to a particular number of selected regions. One possible explanation is that middle layers play an important role in integrating question-specific visual information into language representations, whereas deeper layers increasingly propagate the integrated representations toward answer generation [49]. **Similar layer-wise patterns have been observed in prior studies** [50, 51]. Our observations provide further empirical support for extracting role-conditioned visual attention from middle decoder layers for subsequent cross-agent visual coordination.

D.2 Effect of Layer Selection for Visual Attention Extraction

To investigate how the choice of layers affects visual attention coordination, we vary the layer ranges used for extracting role-conditioned visual attention patterns while keeping the Value modulation layers fixed at the middle layers (15–25) and all other components unchanged. As shown in Table 9, extracting visual attention from middle layers (15–25) achieves the highest performance across all five benchmarks, with an average score of 68.35. This corresponds to gains of 3.43 and 2.04 percentage points over Latent-VMAS and all-layer extraction, respectively. In contrast, extracting attention from early or late layers yields lower average scores of 65.71 and 66.81, respectively. These results suggest that middle-layer attention provides a more effective signal for cross-agent visual coordination than attention extracted from early, late, or all layers.

E Derivation and Properties of Cross-Agent Visual Consensus

We provide the mathematical derivations and properties underlying the three components of cross-agent visual consensus: agent-conditioned visual attention aggregation, spatial uncertainty marginalization, and geometric consensus.

Throughout this section, let $\mathcal{V} = \{1, \dots, N_v\}$ denote the visual token set and

$$\Delta = \left\{ p \in \mathbb{R}^{N_v}_{\geq 0} \;\middle|\; \sum_{j=1}^{N_v} p(j) = 1 \right\} \tag{16}$$

denote the probability simplex over visual tokens.

E.1 Mathematical Interpretation of Conditional Visual Attention Readout

We provide a mathematical interpretation of the conditional visual attention readout in Eq. 1. Our objective is to characterize the relative attention patterns of different agents within the shared visual representation, rather than their absolute attention allocation to visual content. We first establish the relationship between visual-only normalization and full-sequence attention, and then show that mean-pooling Query states induces a geometric aggregation of token-level conditional visual attention distributions.

Conditional visual attention. Consider an attention head at a fixed Transformer layer. Let $\mathcal{T}_t$ denote the set of tokens accessible to a query at position t, and let $\mathcal{V} \subseteq \mathcal{T}_t$ denote the shared visual tokens. For simplicity, we omit the layer and head indices. The attention score between query q_t and key k_j is

$$s_{t,j} = \frac{q_t k_j^\top}{\sqrt{d_h}}. \tag{17}$$

The full-sequence attention distribution is given by

$$A_t(j) = \frac{\exp(s_{t,j})}{\sum_{k \in \mathcal{T}_t} \exp(s_{t,k})}, \qquad j \in \mathcal{T}_t. \tag{18}$$

Restricting the normalization to the visual tokens yields

$$P_t^{\text{vis}}(j) = \frac{\exp(s_{t,j})}{\sum_{k \in \mathcal{V}} \exp(s_{t,k})}, \qquad j \in \mathcal{V}. \tag{19}$$

By substituting the full-sequence attention weights, we obtain

$$\begin{aligned} \frac{A_t(j)}{\sum_{k \in \mathcal{V}} A_t(k)} &= \frac{\frac{\exp(s_{t,j})}{\sum_{u \in \mathcal{T}_t} \exp(s_{t,u})}}{\frac{\sum_{k \in \mathcal{V}} \exp(s_{t,k})}{\sum_{u \in \mathcal{T}_t} \exp(s_{t,u})}} \\ &= \frac{\exp(s_{t,j})}{\sum_{k \in \mathcal{V}} \exp(s_{t,k})} \\ &= P_t^{\text{vis}}(j). \end{aligned} \tag{20}$$

Thus, visual-only softmax corresponds exactly to the full-sequence attention distribution conditioned on the visual-token subset, provided that the same attention scores and accessibility constraints are used. Importantly, this conditional distribution does not measure the absolute proportion of attention allocated to visual tokens. Instead, it characterizes how a query distributes its attention among the shared visual tokens, independently of the total attention mass assigned to non-visual tokens.

Geometric aggregation induced by Query pooling. In Eq. 1, we construct an agent-conditioned query by mean-pooling the Query states of the textual prefix. Let $\mathcal{X}_i$ denote the selected query positions of agent A_i, with $T_i = |\mathcal{X}_i|$. For a fixed layer and attention head, the pooled query is

$$\bar{q}_i = \frac{1}{T_i} \sum_{t \in \mathcal{X}_i} q_{i,t}. \tag{21}$$

The conditional visual attention distribution computed from this pooled query is

$$\begin{aligned}\bar{P}_i^{\text{vis}}(j) &= \frac{\exp\left(\bar{q}_i k_j^\top / \sqrt{d_h}\right)}{\sum_{u\in\mathcal{V}} \exp\left(\bar{q}_i k_u^\top / \sqrt{d_h}\right)} \\ &= \frac{\exp\left(\frac{1}{T_i}\sum_{t\in\mathcal{X}_i} s_{i,t,j}\right)}{\sum_{u\in\mathcal{V}} \exp\left(\frac{1}{T_i}\sum_{t\in\mathcal{X}_i} s_{i,t,u}\right)}.\end{aligned} \tag{22}$$

For each query position t, define its token-level conditional visual attention distribution as

$$P_{i,t}^{\text{vis}}(j) = \frac{\exp(s_{i,t,j})}{Z_{i,t}^{\text{vis}}}, \qquad Z_{i,t}^{\text{vis}} = \sum_{u\in\mathcal{V}} \exp(s_{i,t,u}). \tag{23}$$

Since

$$\begin{aligned}\exp\left(\frac{1}{T_i}\sum_{t\in\mathcal{X}_i} s_{i,t,j}\right) &= \left(\prod_{t\in\mathcal{X}_i} \exp(s_{i,t,j})\right)^{1/T_i} \\ &= \left(\prod_{t\in\mathcal{X}_i} Z_{i,t}^{\text{vis}}\right)^{1/T_i} \left(\prod_{t\in\mathcal{X}_i} P_{i,t}^{\text{vis}}(j)\right)^{1/T_i},\end{aligned} \tag{24}$$

the visual partition functions cancel under normalization. Consequently,

$$\boxed{\bar{P}_i^{\text{vis}}(j) = \frac{\left(\prod_{t\in\mathcal{X}_i} P_{i,t}^{\text{vis}}(j)\right)^{1/T_i}}{\sum_{u\in\mathcal{V}} \left(\prod_{t\in\mathcal{X}_i} P_{i,t}^{\text{vis}}(u)\right)^{1/T_i}}.} \tag{25}$$

Therefore, computing visual-only softmax from the mean-pooled Query is mathematically equivalent to taking the normalized geometric mean of the token-level conditional visual attention distributions.

Furthermore, this distribution is the solution to the following KL-divergence minimization problem:

$$\bar{P}_i^{\text{vis}} = \arg\min_{p\in\Delta^{N_v-1}} \frac{1}{T_i}\sum_{t\in\mathcal{X}_i} D_{\text{KL}}\left(p \,\|\, P_{i,t}^{\text{vis}}\right), \tag{26}$$

where Δ^{N_v-1} denotes the probability simplex over the visual tokens.

This result shows that Query pooling yields a representative conditional visual distribution that geometrically aggregates the relative visual attention patterns across the selected query positions. The resulting distribution is subsequently averaged across the selected attention heads and intermediate Transformer layers, as defined in Eq. equation 1.

Implications for cross-agent visual coordination. In latent collaboration, agents operate under different reasoning contexts, and subsequent agents may attend to accumulated latent states in addition to the shared visual tokens. Consequently, their absolute visual attention masses can differ substantially, even when they exhibit similar relative attention patterns within the shared visual representation. Our conditional normalization intentionally separates these two quantities. Specifically, the conditional distribution characterizes the relative allocation of visual attention across spatial locations, while the absolute visual attention mass describes the proportion of full-sequence attention assigned to visual tokens. `V-Guide` focuses on the former because its objective is to identify spatial agreement among role-conditioned visual attention patterns. By comparing distributions normalized over the same shared visual-token set, our method captures which visual locations are consistently prioritized across agents without directly weighting their contributions by differences in overall visual attention allocation. This interpretation does not imply that the resulting distributions measure the absolute strength of visual attention or the reliability of individual visual tokens. Rather, they provide a common representation of relative visual attention patterns for subsequent spatial marginalization and cross-agent consensus.

E.2 Spatial Uncertainty Marginalization

We next derive the discrete probabilistic formulation of spatial uncertainty marginalization and establish its contraction properties.

Discrete Spatial Uncertainty Model. Let $r_j \in \mathbb{R}^2$ denote the spatial coordinate of visual token j on the visual token grid. We introduce two discrete random variables: $K \in \mathcal{V}$, representing the original attention location, and $J \in \mathcal{V}$, representing its uncertainty-perturbed location.

For agent A_i, the original attention location follows

$$\Pr(K = k) = p_i(k). \tag{27}$$

Conditioned on $K = k$, we model the perturbed location using a Gaussian transition kernel:

$$\begin{aligned} \Pr(J = j \mid K = k) &= G_\sigma(j \mid k) \\ &= \frac{\exp\left(-\frac{\|r_j - r_k\|_2^2}{2\sigma^2}\right)}{\sum_{m=1}^{N_v} \exp\left(-\frac{\|r_m - r_k\|_2^2}{2\sigma^2}\right)}. \end{aligned} \tag{28}$$

Here, $\sigma > 0$ controls the spatial extent of the perturbation. The normalization is performed over all valid destination tokens for each original location k, accounting for the finite visual token grid.

By marginalizing over the original attention location K, we obtain

$$\begin{aligned} \widetilde{p}_i(j) &= \Pr(J = j) \\ &= \sum_{k=1}^{N_v} \Pr(J = j \mid K = k) \Pr(K = k) \\ &= \sum_{k=1}^{N_v} G_\sigma(j \mid k) p_i(k). \end{aligned} \tag{29}$$

This establishes Eq. 2 directly as a discrete marginalization, without requiring a continuous attention density or interpolation between visual tokens.

In the limiting case $\sigma \to 0^+$, the kernel approaches the identity operator on a grid with distinct token coordinates, recovering the original attention distribution.

Probability Preservation. For any original location k, the Gaussian transition kernel satisfies

$$G_\sigma(j \mid k) \geq 0, \qquad \sum_{j=1}^{N_v} G_\sigma(j \mid k) = 1. \tag{30}$$

Consequently, for any $p_i \in \Delta$,

$$\begin{aligned} \sum_{j=1}^{N_v} \widetilde{p}_i(j) &= \sum_{j=1}^{N_v} \sum_{k=1}^{N_v} G_\sigma(j \mid k) p_i(k) \\ &= \sum_{k=1}^{N_v} p_i(k) \sum_{j=1}^{N_v} G_\sigma(j \mid k) \\ &= \sum_{k=1}^{N_v} p_i(k) = 1. \end{aligned} \tag{31}$$

Together with non-negativity, this shows that $\widetilde{p}_i \in \Delta$. Thus, G_σ is a column-stochastic Markov operator that preserves the total probability mass.

L_1 **Contraction.** For any two distributions $p, q \in \Delta$, we have

$$\begin{aligned}\|G_\sigma p - G_\sigma q\|_1 &= \sum_{j=1}^{N_v} \left| \sum_{k=1}^{N_v} G_\sigma(j \mid k)(p(k) - q(k)) \right| \\ &\le \sum_{j=1}^{N_v} \sum_{k=1}^{N_v} G_\sigma(j \mid k)|p(k) - q(k)| \\ &= \sum_{k=1}^{N_v} |p(k) - q(k)| \sum_{j=1}^{N_v} G_\sigma(j \mid k) \\ &= \|p - q\|_1.\end{aligned} \tag{32}$$

The inequality follows from the triangle inequality and the non-negativity of the transition probabilities. Hence, applying the same spatial marginalization operator to two attention distributions cannot increase their L_1 distance.

KL Contraction. We further establish that spatial marginalization cannot increase the KL divergence between two distributions.

Let

$$P_j = \sum_{k=1}^{N_v} G_\sigma(j \mid k)p(k), \qquad Q_j = \sum_{k=1}^{N_v} G_\sigma(j \mid k)q(k). \tag{33}$$

For each destination token j, the log-sum inequality gives

$$\begin{aligned}P_j \log \frac{P_j}{Q_j} &= \left(\sum_{k=1}^{N_v} G_\sigma(j \mid k)p(k) \right) \log \frac{\sum_{k=1}^{N_v} G_\sigma(j \mid k)p(k)}{\sum_{k=1}^{N_v} G_\sigma(j \mid k)q(k)} \\ &\le \sum_{k=1}^{N_v} G_\sigma(j \mid k)p(k) \log \frac{p(k)}{q(k)}.\end{aligned} \tag{34}$$

Summing over j and using Eq. 30, we obtain

$$\begin{aligned}D_{\mathrm{KL}}(G_\sigma p \| G_\sigma q) &= \sum_{j=1}^{N_v} P_j \log \frac{P_j}{Q_j} \\ &\le \sum_{j=1}^{N_v} \sum_{k=1}^{N_v} G_\sigma(j \mid k)p(k) \log \frac{p(k)}{q(k)} \\ &= \sum_{k=1}^{N_v} p(k) \log \frac{p(k)}{q(k)} \sum_{j=1}^{N_v} G_\sigma(j \mid k) \\ &= D_{\mathrm{KL}}(p \| q).\end{aligned} \tag{35}$$

The result follows from the data-processing inequality for a common Markov transition kernel.

Together, Eqs. 32 and 35 establish that spatial uncertainty marginalization is non-expansive under both L_1 distance and KL divergence.

These properties hold for any common probability-preserving Markov kernel and are not specific to the Gaussian form. The Gaussian kernel additionally provides a spatially structured transition rule, assigning progressively lower probability to more distant visual tokens.

E.3 Derivation of Geometric Consensus

Given the spatially marginalized attention distributions $\{\widetilde{p}_i\}_{i=1}^M$, we seek a consensus distribution $c \in \Delta$ that minimizes its average reverse KL divergence to the agent-conditioned distributions:

$$c = \arg\min_{q \in \Delta} \frac{1}{M} \sum_{i=1}^{M} D_{\mathrm{KL}}(q \| \widetilde{p}_i). \tag{36}$$

Since the visual-only softmax and the Gaussian transition kernel assign strictly positive probability to every visual token, we have $\widetilde{p}_i(j) > 0$ for all agents i and visual tokens j.

Closed-Form Solution. Expanding the KL divergence, we obtain

$$\begin{aligned}\mathcal{J}(q) &= \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{N_v} q(j)\log\frac{q(j)}{\widetilde{p}_i(j)} \\ &= \sum_{j=1}^{N_v} q(j)\log q(j) - \sum_{j=1}^{N_v} q(j)\left(\frac{1}{M}\sum_{i=1}^{M}\log\widetilde{p}_i(j)\right).\end{aligned} \tag{37}$$

Define the unnormalized geometric mean

$$g_j = \exp\left(\frac{1}{M}\sum_{i=1}^{M}\log\widetilde{p}_i(j)\right) = \prod_{i=1}^{M}\widetilde{p}_i(j)^{1/M}. \tag{38}$$

Substituting this definition into Eq. 37 gives

$$\mathcal{J}(q) = \sum_{j=1}^{N_v} q(j)\log\frac{q(j)}{g_j}. \tag{39}$$

Let

$$Z = \sum_{j=1}^{N_v} g_j, \qquad \bar{g}_j = \frac{g_j}{Z}. \tag{40}$$

Since $g_j > 0$, we have $Z > 0$ and $\bar{g} \in \Delta$. Using the normalization constraint $\sum_j q(j) = 1$, we can rewrite the objective as

$$\begin{aligned}\mathcal{J}(q) &= \sum_{j=1}^{N_v} q(j)\log\frac{q(j)}{Z\bar{g}_j} \\ &= D_{\mathrm{KL}}(q\|\bar{g}) - \log Z.\end{aligned} \tag{41}$$

The term $-\log Z$ is independent of q. By the non-negativity of KL divergence,

$$D_{\mathrm{KL}}(q\|\bar{g}) \geq 0, \tag{42}$$

with equality if and only if $q = \bar{g}$.

Therefore, the unique minimizer of Eq. 36 is

$$\boxed{c_j = \frac{\prod_{i=1}^{M}\widetilde{p}_i(j)^{1/M}}{\sum_{k=1}^{N_v}\prod_{i=1}^{M}\widetilde{p}_i(k)^{1/M}}}. \tag{43}$$

This establishes the normalized geometric mean as the closed-form solution to the reverse-KL consensus objective.

Interpretation of Cross-Agent Agreement. The geometric consensus assigns each visual token an unnormalized weight

$$g_j = \prod_{i=1}^{M}\widetilde{p}_i(j)^{1/M}. \tag{44}$$

Its logarithm is the average log-probability assigned to that token across agents:

$$\log g_j = \frac{1}{M}\sum_{i=1}^{M}\log\widetilde{p}_i(j). \tag{45}$$

Consequently, a low probability assigned by an agent reduces the geometric weight of the corresponding visual token, even when other agents assign it relatively high attention.

In contrast, arithmetic averaging is the solution to the opposite KL orientation:

$$\arg\min_{q\in\Delta} \frac{1}{M}\sum_{i=1}^{M} D_{\mathrm{KL}}(\widetilde{p}_i\|q) = \frac{1}{M}\sum_{i=1}^{M}\widetilde{p}_i. \tag{46}$$

The two objectives therefore induce different aggregation rules. Arithmetic averaging combines attention mass from individual agents, whereas the reverse-KL formulation favors visual tokens receiving consistent support across agents.

Our formulation uses the latter to construct a shared visual attention distribution for subsequent visual Value calibration.

F Multi-Agent Role Prompts

We provide the role-specific prompt templates used for both multi-agent architectures.

In the following templates, `[Question]` denotes the input question, and the input image is supplied as part of each agent's multimodal context.

For intermediate agents, the instructions specify their intended reasoning roles. The agents perform latent reasoning without explicitly generating the requested textual plans or analyses. Only the final agent produces a textual response.

F.1 Sequential MAS Prompts

Visual Planner.

> You are a Planner Agent.
> Question: `[Question]`
> Determine how to solve the current question from the image. Identify the relevant entities, decisive visual cues, required comparisons, and reasoning steps.
> Form a question-specific visual instruction and an initial hypothesis. Ground the hypothesis in the image rather than language priors.

Visual Critic.

> You are a Critic Agent.
> Question: `[Question]`
> The visual instruction and an initial hypothesis is provided in latent KV representation format.
> Independently inspect the image and challenge the current hypothesis.
> Look for overlooked details, ambiguous references, conflicting visual evidence, and competing answers.
> Determine which observation would most strongly confirm or reject the hypothesis.

Visual Refiner.

> You are a Refiner Agent.
> Question: `[Question]`
> Integrate the plan, hypothesis, critique, and original image.
> Resolve the strongest uncertainty using visible evidence.
> Compare the leading candidate answers, refine the reasoning, and form the best-supported answer hypothesis.

Visual Verifier.

You are a helpful assistant. You are provided with latent information for reference.

Question: `[Question]`

The latent information might contain irrelevant contents. Ignore it if it is not helpful for solving the target question.

Verify the refined hypothesis against the original image and question.

Check the decisive visual evidence and the strongest competing answer. Correct any unsupported conclusion.

You must reason step-by-step without outputting other irrelevant information and then output the final answer inside `\boxed{YOUR_FINAL_ANSWER}`.

F.2 HIERARCHICAL MAS PROMPTS

Visual Perception Agent.

You are a Perception Agent.

Question: `[Question]`

Focus on VISUAL DETAILS in the image. Carefully identify and describe:

- Key objects, their attributes (color, size, shape, text), and spatial relationships.
- Fine-grained visual details that are relevant to answering the question.
- Any subtle or easily-missed visual cues.

Provide your visual analysis concisely. Do not give the final answer.

Visual Logic Agent.

You are a Logic Agent.

Question: `[Question]`

Focus on LOGICAL REASONING. Analyze:

- The logical structure of the question and what it requires.
- Cause-and-effect relationships visible in the image.
- Compare and contrast relevant options or possibilities.
- Identify potential reasoning traps or ambiguities.

Provide your logical analysis concisely. Do not give the final answer.

Contextual Knowledge Agent.

You are a Knowledge Agent.

Question: `[Question]`

Focus on applying DOMAIN KNOWLEDGE and COMMONSENSE. Consider:

- Relevant world knowledge, facts, or conventions that help interpret the image.
- Common sense reasoning about the scene or situation depicted.
- Cultural, scientific, or practical knowledge relevant to the question.

Provide your knowledge-based analysis concisely. Do not give the final answer.

Visual Synthesizer.

You are a Synthesizer.

Question: `[Question]`

You are provided with latent representations from three specialized expert agents (Perception, Logic, and Knowledge) who each analyzed the question from different perspectives.

The latent information might contain irrelevant or conflicting contents.

Weigh the evidence from all perspectives and resolve any conflicts.

You must reason step-by-step based on the image and synthesized information, then output the final answer inside `\boxed{YOUR_FINAL_ANSWER}`.

G LIMITATIONS

We currently instantiate all agents from the same MLLM backbone. This homogeneous setting allows layer-wise KV states to be reused directly across agents without additional projection or alignment, and is well suited to the training-free latent communication paradigm considered in this work. Our focus is therefore on how agents coordinate visual information within a shared representation space, rather than on cross-model representation alignment.

An interesting extension is to support heterogeneous agents based on different MLLM backbones. Because their internal states may differ in dimensionality, layer organization, attention structure, and representation space, direct KV-cache reuse is generally not available in this setting. Future work could explore lightweight alignment modules or a shared latent interface that enables heterogeneous models to communicate through compatible internal representations while retaining the efficiency advantages of latent collaboration.