Latent-Lens: Visual Perception in Small Language Models Through Latent Communication

Mo Shahdloo
Microsoft Applied Sciences Group (ASG)

Faegheh Sardari
Microsoft Applied Sciences Group (ASG)

Mathew Salvaris
Microsoft Applied Sciences Group (ASG)

Abstract

Bridging independent vision-language models (VLMs) and small language models (SLMs) in a modular multi-agent system typically forces them to exchange information as discrete text, which creates a lossy bottleneck for spatial and fine-grained details. This constraint is especially acute for SLMs, whose limited capacity makes them more reliant on the quality of the input signal. Recent latent-communication methods exchange states between language models. How a VLM can pass visual information to a separately pretrained text-only SLM though remains unexplored. We propose **Latent-Lens**, a lightweight codec that bridges the representation gap between a frozen VLM and a frozen SLM directly in latent space, enabling them to communicate purely through hidden states. The codec consists of a domain encoder that projects hidden states from the VLM into the SLM's embedding space, and an attention decoder formalised as a LoRA that teaches the SLM to attend to these visual prefixes, while training only $\sim$1.6M parameters (0.2% of the combined model size). On visual question answering (CLEVR, GQA), Latent-Lens achieves $81.0 \pm 7.9\%$ and $65.4 \pm 5.4\%$, compared with $58.3 \pm 5.7\%$ and $61.8 \pm 1.0\%$ for text communication with the same LoRA budget (+30.2–69.8pp relative to frozen text communication), while on image captioning (Flickr8k) it improves over the same baseline by +16.5 (+18.8 frozen). We also evaluate Latent-Lens with TinyLlama-1.1B and InternVL2-1B, which use different architectures, vocabularies, and hidden dimensions.

1 Introduction

Small language models (SLMs) are increasingly deployed on-device and at the edge, where compute and memory budgets can preclude accommodating a monolithic vision-language model (VLM), leaving these systems without visual perception. Tightly coupled VLMs such as LLaVA [Liu et al., 2023] and Qwen-VL [Bai et al., 2023] fuse visual and linguistic representations within a single end-to-end trained system. While effective, this tight coupling is inflexible. Upgrading the vision encoder or the language model requires retraining the entire system, and scaling such VLMs down to fit on-device budgets often degrades visual capabilities disproportionately. For example, TinyLLaVA [Zhou et al., 2024] reports substantial benchmark drops below 1B parameters. Multi-agent approaches offer modularity by pairing a small VLM with a SLM, but have so far relied on text-based communication [Du et al., 2024]. However, text constitutes a lossy channel. Spatial layouts and fine-grained visual attributes are difficult to verbalise faithfully. SLMs amplify this loss, since with little parametric capacity to fall back on they have no reliable way to recover details the text channel has already discarded. For instance, BeMyEyes [Huang et al., 2025] pairs a VLM with a large proprietary LLM through text conversation, but even with substantially more capable models, the generated descriptions inevitably discard fine-grained spatial and visual details.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Recent works on latent multi-agent communication aim to address this bottleneck [Zou et al., 2025, Zheng et al., 2025, Dery et al., 2026]. However, existing methods operate exclusively within the language modality, and their expressiveness is limited by reliance on representations from a single transformer layer pooling the latent information. Extending these latent-space protocols to cross-modal vision-language multi-agent systems is not yet studied in the literature.

We propose **Latent-Lens**, a lightweight, model-agnostic codec that bridges the representation gap between a frozen VLM and a frozen SLM, allowing them to communicate directly in latent space and thus bypass the lossy text bottleneck. In contrast to discrete text descriptions (as in BeMyEyes [Huang et al., 2025]), our method transmits rich, uncompressed visual representations directly. By extracting multiband latent prefixes, our approach enables cross-modal perception while preserving the depth-wise hierarchy of features encoded across the VLM's layers.

Latent-Lens consists of two functionally distinct components, jointly trained end-to-end. The first is a lightweight MLP domain encoder that extracts and projects hidden states from three semantic bands of the frozen VLM. The second is an attention decoder, instantiated as LoRA adapters on the frozen SLM's attention projections, that teaches the SLM to attend to these projected prefix tokens. We evaluate Latent-Lens with two SLMs, Qwen2.5-0.5B (896-dim [Alibaba Cloud, 2024]) and TinyLlama-1.1B (2048-dim [Zhang et al., 2024]), and two small VLMs, SmolVLM-256M (576-dim [HuggingFace, 2024]) and InternVL2-1B (896-dim [Chen et al., 2024]), to demonstrate generality across different families of models. Only 1.6-6.1M parameters are trained per pair, representing less than 1% of the combined frozen parameters in all cases.

We evaluate on free-form VQA tasks using CLEVR [Johnson et al., 2017] and GQA [Hudson and Manning, 2019], and an open-ended captioning task using Flickr8k [Hodosh et al., 2013], comparing against VLM-only, SLM-only, text-communication, and LLaMA-Adapter [Zhang et al., 2023] deeper-injection baselines. Latent-Lens reaches $81.0 \pm 7.9\%$ on CLEVR, $65.4 \pm 5.4\%$ on GQA, and 18.9 ± 1.0 BLEU-4 on Flickr8k with Qwen2.5-0.5B, and 78.2% / 66.3% VQA with TinyLlama-1.1B, with consistent margins over text communication and LLaMA-Adapter across both SLM families.

2 Related Work

Text-based inter-model communication. Multi-agent systems built from independently trained components most commonly exchange information through natural language. Du et al. [2024] show that multi-round text debate among LLM instances improves factuality and reasoning. Liang et al. [2024] extend this by encouraging divergent thinking across agents to enhance answer diversity, and ChatDev [Qian et al., 2024] applies text-based conversation to coordinate specialised models for software development. In the visual domain, BeMyEyes [Huang et al., 2025] pairs a small VLM with a large proprietary LLM, where the VLM describes the image in text and the LLM reasons over the description. While modular, all these approaches are limited by the expressiveness of the text channel, which discards spatial layout and fine-grained visual attributes.

Latent representation transfer between models. An alternative to text is to communicate directly through hidden states, motivated by the observation that natural language is structurally misaligned with LLMs' high-dimensional representations and causes irreversible information compression [Zhou et al., 2025]. In line with this, Yue et al. [2025] show that 28-73% of text-based inter-agent communication is redundant. The idea originated in multi-agent reinforcement learning. DIAL [Foerster et al., 2016] introduced differentiable inter-agent messaging, instantiated by CommNet [Sukhbaatar et al., 2016] as a mean-pooled hidden-state channel. Subsequent work introduced selectivity over the channel via attention-based targeting in TarMAC [Das et al., 2019] and learnable gating in IC3Net [Singh et al., 2019], and structured topologies via relational graphs [Jiang et al., 2019].

More recently, the principle has been applied to transfer representations between frozen language models. LatentMAS [Zou et al., 2025] shares last-layer hidden embeddings through a latent working memory between *identical* LLMs, achieving up to 14.6% accuracy gains with lower latency than text debate, and ThoughtComm [Zheng et al., 2025] provides identifiability guarantees for such latent thought exchange. A further line of work targets heterogeneous LLMs with different architectures and vocabularies: Interlat [Du et al., 2025] learns compression adapters across architectures for a substantial inference speedup, C2C [Zhang et al., 2025] fuses KV caches across Qwen, Llama,

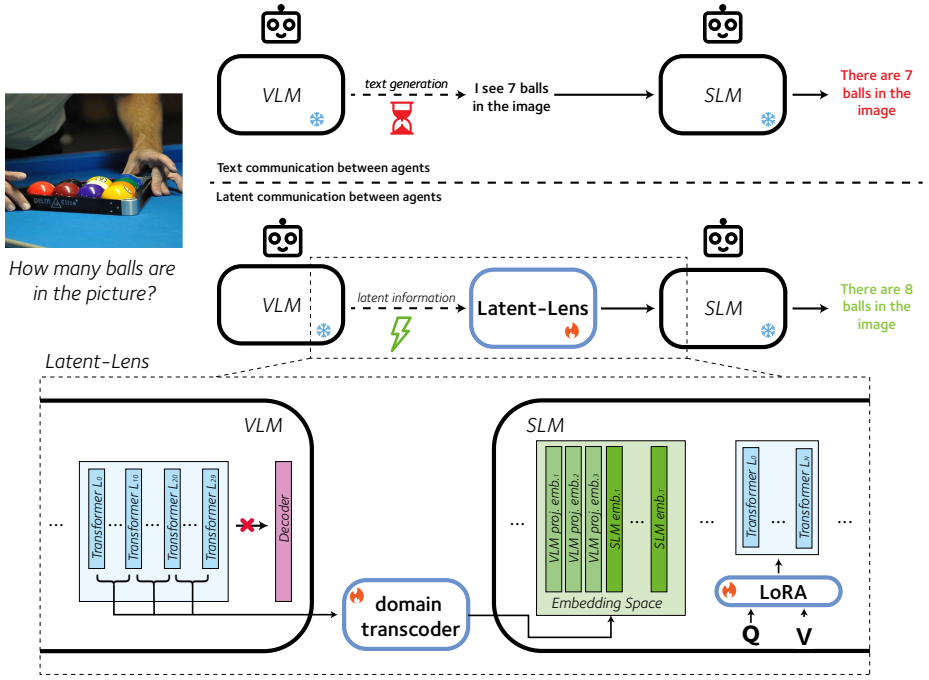


Figure 1: **Latent-Lens architecture.** Text communication passes a generated response from the VLM to the SLM. Latent-Lens passes projected VLM hidden states to the SLM. Black borders denote frozen components. The blue border denotes the trained codec.

and Gemma via learned gating, and Dery et al. [2026] learn KV-cache alignment adapters in the same spirit, SDE [Tang et al., 2025] takes a hybrid stance, augmenting text with hidden-state deltas, and Vision Wormhole [Liu et al., 2026] communicates reasoning traces between VLMs through the receiver's visual-token pathway. None of these methods address transfer from a VLM to a text-only SLM, which is the setting studied here. Moreover, these methods do not use the sender's depth hierarchy as a structured signal. Hidden-state methods exchange a single transformer layer, while KV-cache methods span all layers but pool or align them uniformly rather than separating depth-wise abstraction levels. Finally, although multi-depth structure has been explored within a *single* model - e.g., LLaMA-Adapter [Zhang et al., 2023] inserts gated visual residuals at several transformer depths of one VLM - multi-depth transfer across two separate models is unexplored.

Contributions.

- We introduce **Latent-Lens**, a latent-communication codec that passes visual information from a frozen VLM to a separately pretrained text-only SLM.
- Unlike prior inter-model latent methods that exchange a single transformer layer's representation, Latent-Lens extracts from multiple semantic depth bands of the sender VLM, preserving its perceptual hierarchy.
- Trained on only 1.6M parameters (0.2% of the frozen pair), Latent-Lens outperforms text-communication baselines on image-understanding tasks requiring fine-grained visual perception (VQA, captioning), and generalises to alternative SLM and VLM backbone combinations.

3 Methods

3.1 Problem Setup

Given an image and a question, we want a frozen SLM to generate a free-form answer that leverages visual information from a frozen VLM, without any text exchange between the two models (Figure 1). We denote the VLM by $\mathcal{V}$, with L_V transformer layers and hidden dimension d_V, and the SLM by $\mathcal{S}$, with L_S layers and hidden dimension d_S. The two have different tokenisers, vocabularies, and embedding dimensions ($d_V \neq d_S$ in general), so the latent channel must explicitly bridge their representation spaces. The specific backbones used in our experiments are described in Section 4.1.

3.2 Multiband Signal Decomposition

A single-layer representation collapses the VLM's hierarchical processing into one vector, discarding information that earlier or intermediate layers uniquely encode (see 4.4). To preserve this structure, we propose to decompose the VLM's hidden states across multiple depth bands, each spanning a contiguous group of transformer layers.

To do this, the VLM processes the image and question, producing per-layer hidden states $\{h_\ell\}_{\ell=0}^{L_V-1}$ of the VLM's transformer stack, where $h_\ell \in \mathbb{R}^{T\times d_V}$ is the activation at layer ℓ and T is the input token sequence length. Then, we partition the L_V layers into $B=3$ bands of equal size and compute the band representation as the mean of last-token hidden states within each band:

$$\mathbf{v}_b = \frac{1}{|\mathcal{L}_b|}\sum_{\ell\in\mathcal{L}_b} h_\ell[:,-1,:] \in \mathbb{R}^{d_V}, \quad b \in \{1,2,3\} \tag{1}$$

where $\mathcal{L}_b$ is the set of layer indices in band b. Each band covers a contiguous third of the VLM's transformer stack: band 1 (early), band 2 (mid), and band 3 (late). As we show in Section 4.4, a simple classifier predicts the correct answer more accurately from later bands, while the early and late bands are less similar to each other than the middle and late bands.

3.3 Domain Transcoding

The three band representations are projected from the VLM's representation space into the SLM's embedding space using a shared 2-layer MLP that serves as a lightweight domain transcoder.

$$\tilde{\mathbf{v}}_b = W_2\,\mathrm{GELU}(W_1\mathbf{v}_b + b_1) + b_2, \quad W_1 \in \mathbb{R}^{d_S\times d_V},\ W_2 \in \mathbb{R}^{d_S\times d_S} \tag{2}$$

This yields three projected embeddings $\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, \tilde{\mathbf{v}}_3 \in \mathbb{R}^{d_S}$, one per band.

3.4 Prefix Delivery and SLM Adaptation

The transcoded band embeddings are prepended as soft prefix tokens to the SLM's input token sequence:

$$\mathbf{E}_{\text{input}} = [\tilde{\mathbf{v}}_1; \tilde{\mathbf{v}}_2; \tilde{\mathbf{v}}_3; \mathbf{e}_1, \ldots, \mathbf{e}_T] \in \mathbb{R}^{(3+T)\times d_S} \tag{3}$$

where $\mathbf{e}_1, \ldots, \mathbf{e}_T$ are the SLM's text token embeddings. The prefix tokens receive attention mask 1 so the SLM attends to them at every decoding step, but their positions are excluded from the SLM's next-token prediction loss (*i.e.*, cross-entropy loss, label = -100).

To adapt the SLM to attend to this novel visual signal, we add LoRA (rank 4, ablated in Section 4.6) to the query and value projections of every attention layer. The full trainable budget thus consists of the projection MLP plus the LoRA adapters. Per-pair parameter counts are reported in Section B of the Supplementary Materials.

4 Experiments

4.1 Setup

Models. The default set of models used here are SmolVLM-256M [HuggingFace, 2024] ($L_V = 30$, $d_V = 576$, 256M params) as the VLM agent and Qwen2.5-0.5B [Alibaba Cloud, 2024] ($L_S =$

24, $d_S = 896$, 500M params) as the SLM agent. For this pair, the multiband partition yields layer ranges L0-9 / L10-19 / L20-29 and the trainable budget is 1.6M parameters (1.3M projection MLP + 0.3M LoRA, 0.2% of the 752M combined). To test generality across heterogeneous backbones (Section 4.3), we additionally evaluate Latent-Lens with InternVL2-1B [Chen et al., 2024] ($L_V = 24$, $d_V = 896$, 900M params) as the VLM agent and TinyLlama-1.1B [Zhang et al., 2024] ($L_S = 22$, $d_S = 2048$, 1.1B params) as the SLM agent, spanning two model families (Qwen vs LLaMA), tokenisers, vocabularies, and hidden dimensions. Per-pair specifications and parameter counts are listed in Sections A and B.

Datasets. We use two free-form VQA benchmarks (*i.e.*, **CLEVR** [Johnson et al., 2017] and **GQA** [Hudson and Manning, 2019]) and one open-ended captioning benchmark (*i.e.*, **Flickr8k** [Hodosh et al., 2013]). **CLEVR**[1]: rendered 3D scenes with questions about counting, spatial relations, and attributes; 5,000 examples with a 4,000 / 1,000 train/test split. **GQA**[2]: real photographs with compositional questions requiring multi-step reasoning; 9,600 / 2,400 train/test split. **Flickr8k**[3]: open-ended image captioning; 4,000 / 1,000 train/test split.

Evaluation. For VQA we use free-form answer generation and compare predictions to ground truth with a normalized containment heuristic. We lowercase the prediction and reference, remove articles and punctuation, normalize whitespace, and convert number words to digits. We accept an exact match or a match in which every word in the shorter answer appears in the longer answer. Multiple-choice formatting with random distractors inflated SLM-only GQA accuracy in preliminary experiments, motivating the switch to free-form generation as a stricter test of visual understanding. Captioning on Flickr8k is evaluated with BLEU-4 against the reference captions.

Training. We use AdamW with learning rate 2×10^{-4}, cosine annealing, weight decay 10^{-4}, gradient clipping at 1.0. The effective batch size is 16 (4×4 gradient accumulation), and training runs for 5 epochs.

Baselines. (1) **VLM-only (frozen)**: SmolVLM-256M [HuggingFace, 2024] generates directly from the image and question. (2) **SLM-only (frozen)**: Qwen2.5-0.5B [Alibaba Cloud, 2024] with no trainable parameters, text input only. (3) **Text communication (frozen)**: the VLM produces a textual response to the question, which is prepended to the frozen SLM prompt. (4) **SLM-only + LoRA**: same as (2) with 0.3M LoRA [Hu et al., 2022] parameters (rank 4). (5) **Text communication + LoRA**: same as (3) with the same 0.3M rank-4 Q/V LoRA used by Latent-Lens. (6) **LLaMA-Adapter** [Zhang et al., 2023]: deeper-injection baseline that uses the same shared VLM→SLM projection as our method but additionally injects gated visual residuals at each of the top-12 SLM layers, with a total trainable budget matched to ours (1.6M). (7) **Latent-Lens**: our method (1.6M params: 1.3M projection + 0.3M rank-4 Q/V LoRA). The +LoRA variants isolate the effect of the communication channel from task adaptation by giving the SLM the same LoRA budget as in our method, while the LLaMA-Adapter baseline contrasts our input-prefix + LoRA design against layer-wise gated-residual injection at matched parameter budget.

4.2 Main Results

Table 1 reports the main comparison on CLEVR, GQA, and Flickr8k. Results show that Latent-Lens outperforms all baselines, with $81.0 \pm 7.9\%$ on CLEVR, $65.4 \pm 5.4\%$ on GQA, and 18.9 ± 1.0 BLEU-4 on Flickr8k.

Frozen baselines. We first establish a floor by evaluating each component in isolation. The frozen SLM reaches 1.5%, 18.0%, and 11.8 BLEU-4 on CLEVR, GQA, and Flickr8k, respectively. The VLM in isolation reaches 58.6%, 41.8%, and 13.9 BLEU-4 on CLEVR, GQA, and Flickr8k, respectively. Prepending VLM-generated text responses to the frozen SLM (text communication, frozen) yields 11.2%, 35.2%, and 0.1 BLEU-4 on CLEVR, GQA, and Flickr8k, respectively. Latent-Lens reaches $81.0 \pm 7.9\%$, $65.4 \pm 5.4\%$, and 18.9 ± 1.0 BLEU-4 on CLEVR, GQA, and Flickr8k, respectively, exceeding frozen text communication by +69.8pp, +30.2pp, and +18.8 BLEU-4. The two

[1] `https://huggingface.co/datasets/MMInstruction/Clevr_CoGenT_ValA`

[2] `https://huggingface.co/datasets/lmms-lab/GQA`

[3] `https://huggingface.co/datasets/jxie/flickr8k`

Table 1: Main results on CLEVR, GQA (accuracy) and Flickr8k (BLEU-4). Source column lists the paper introducing each baseline. Trained conditions report mean ± sample standard deviation over three runs; frozen conditions are deterministic point estimates.

Condition	**Source**	**Params**	**CLEVR** (%)	**GQA** (%)	**Flickr8k** (BLEU-4)
VLM only (frozen)	SmolVLM-256M [HuggingFace, 2024]	0	58.6	41.8	13.9
SLM only (frozen)	Qwen2.5-0.5B [Alibaba Cloud, 2024]	0	1.5	18.0	11.8
Text communication (frozen)	this work	0	11.2	35.2	0.1
SLM only + LoRA	Hu et al. [2022]	0.3M	42.4 ± 0.2	46.2 ± 1.3	12.7 ± 0.9
Text communication + LoRA	this work	0.3M	58.3 ± 5.7	61.8 ± 1.0	2.4 ± 2.0
LLaMA-Adapter	Zhang et al. [2023]	1.6M	73.6 ± 3.5	61.9 ± 6.1	18.0 ± 0.7
Latent-Lens (ours, $B=3$)	this work	1.6M	$\mathbf{81.0 \pm 7.9}$	$\mathbf{65.4 \pm 5.4}$	$\mathbf{18.9 \pm 1.0}$

VQA scores under text communication remain well below VLM-only direct generation, indicating that the text channel does not preserve the visual information available to the VLM internally.

Controlling for adaptation. To disentangle the contribution of the communication channel from that of LoRA task adaptation, we add a 0.3M-parameter LoRA adapter (rank 4, matched to the Latent-Lens budget) on top of both the SLM-only and text-communication baselines. SLM+LoRA reaches $42.4 \pm 0.2\%$ / $46.2 \pm 1.3\%$ / 12.7 ± 0.9 BLEU-4, text-communication+LoRA reaches $58.3 \pm 5.7\%$ / $61.8 \pm 1.0\%$ / 2.4 ± 2.0 BLEU-4, and Latent-Lens reaches $81.0 \pm 7.9\%$ / $65.4 \pm 5.4\%$ / 18.9 ± 1.0 BLEU-4. The residual margin is therefore not attributable to the LoRA capacity but to the content of the channel. Changing only the VLM sender prompt to request one or two concise sentences of question-relevant visual evidence gives 69.5% on CLEVR and 58.2% on GQA with text communication + LoRA, still far below Latent-Lens.

Alternative VLM-SLM couplings. We compare Latent-Lens against LLaMA-Adapter [Zhang et al., 2023], an alternative method for conditioning a frozen language model on projected visual features through gated prompts attended at the upper transformer layers. In our adaptation we retain the shared VLM→SLM projection and replace only the input-prefix + LoRA Q/V adaptation with the LLaMA-Adapter mechanism at the top-12 SLM layers, under a matched 1.6M trainable parameter budget. The baseline reaches $73.6 \pm 3.5\%$, $61.9 \pm 6.1\%$, and 18.0 ± 0.7 BLEU-4 on CLEVR, GQA, and Flickr8k, respectively, against $81.0 \pm 7.9\%$, $65.4 \pm 5.4\%$, and 18.9 ± 1.0 for Latent-Lens. We attribute the gap to where the visual signal enters the SLM. LLaMA-Adapter introduces visual context only at the upper transformer layers, so the lower-layer text representations are computed without it. Latent-Lens injects the visual prefix at the input embedding, and all subsequent layers compute over visually-conditioned text representations through the SLM's pre-trained self-attention.

Per-benchmark analysis. **CLEVR** isolates spatial relations, counting, and attribute recognition over synthetic 3D scenes, where the latent channel was expected to separate most clearly from text since fine-grained structured visual content is hardest to verbalise. Latent-Lens achieves $81.0 \pm 7.9\%$, compared with $58.3 \pm 5.7\%$ for text communication + LoRA. **GQA** contains compositional questions over real photographs, on which language priors and short descriptions are expected to carry more signal than on CLEVR. Latent-Lens achieves $65.4 \pm 5.4\%$, compared with $61.8 \pm 1.0\%$ for text communication + LoRA, with SLM+LoRA ($46.2 \pm 1.3\%$) only modestly above VLM-only (41.8%), consistent with the partial sufficiency of text-based reasoning at this benchmark. **Flickr8k** is an open-ended captioning task that allows the VLM to emit substantially richer text than the terse VQA answers, which we hypothesised would narrow the gap between Latent-Lens and VLM-only generation. The VLM produces 40.6-word descriptions on average against 1.0 and 1.1 words on CLEVR and GQA, and the +5.0 BLEU-4 margin of Latent-Lens over VLM-only is the smallest in our experiments, consistent with this hypothesis. To examine whether the observed captioning improvement extends beyond BLEU-4, we rescore the saved captions from Latent-Lens and text communication + LoRA on a diagnostic 200-example Flickr8k subset using CIDEr [Vedantam et al., 2015], obtaining 0.0587 and 0.0215, respectively.

Table 2: Generality across all four VLM $\times$ SLM pairings. Score is the mean of CLEVR (%), GQA (%), and Flickr8k (BLEU-4). VLM-only and SLM+LoRA columns are matched-VLM and matched-SLM baselines averaged the same way. These are single-run compatibility results for each pairing. Table 1 reports three-run means for the default SmolVLM/Qwen pair.

VLM	SLM	VLM-only	SLM+LoRA	Latent-Lens
SmolVLM-256M	Qwen2.5-0.5B	38.1	33.2	60.1
SmolVLM-256M	TinyLlama-1.1B	38.1	38.2	54.0
InternVL2-1B	Qwen2.5-0.5B	51.5	33.2	57.0
InternVL2-1B	TinyLlama-1.1B	51.5	38.2	63.4

Table 3: Linear probe accuracy per band on CLEVR and GQA.

Representation	CLEVR (%)	GQA (%)
Band 1 (early, L0-9)	59.6	43.1
Band 2 (mid, L10-19)	67.9	53.2
Band 3 (late, L20-29)	70.1	57.8
Last layer only	68.6	56.5

4.3 Generality across model backbones

To test generality across heterogeneous backbones, we evaluate Latent-Lens on all four pairings of two VLMs (SmolVLM-256M, InternVL2-1B) and two SLMs (Qwen2.5-0.5B, TinyLlama-1.1B). The two VLMs differ in scale and feature dimensionality, while the two SLMs differ in family (Qwen vs LLaMA), tokeniser, vocabulary, and hidden dimension (896 vs 2048). Each pairing only retrains the projection MLP for the relevant VLM-SLM dimensions and the SLM LoRA adapters, leaving both backbones frozen.

Across all four pairings, Latent-Lens exceeds both the VLM-only and SLM+LoRA baselines on the average score (Table 2). The gain over SLM+LoRA stays in the +15.8 to +26.9 range across all four pairings, indicating that the latent channel adds visual content the text-only SLM cannot recover from its language prior. The smallest gain over VLM-only (+5.5, InternVL2-1B + Qwen2.5-0.5B) occurs when the VLM is already strong in isolation. The Flickr8k scores additionally reflect the SLM's text-only fit to the caption distribution, since TinyLlama-1.1B reaches 27.1 BLEU-4 with no visual input against 11.7 for Qwen2.5-0.5B, which is why the TinyLlama pairings reach higher Flickr8k scores. The InternVL2-1B/TinyLlama-1.1B pairing achieves 87.4% on CLEVR and 72.4% on GQA.

4.4 Representation Analysis

We measure two properties of the multiband latent channel, the task-relevant information contained in each individual band and the overlap between bands.

Per-band information content. To estimate how much task-relevant information each depth band carries, we fit a logistic regression probe on each band's frozen representation, predicting the correct VQA answer from the set observed in the training split. Linear-probe accuracy measures the decodability of task-relevant information from each band.

Higher bands carry more task-relevant information (Table 3), consistent with the depth-abstraction hierarchy [Tenney et al., 2019]. Prior inter-model latent methods extract from a single transformer layer (typically the last), corresponding to the last-layer probe at 68.6 / 56.5. The Band 3 probe (mean of L20-29) reaches 70.1 / 57.8, above the last layer alone, while the full Latent-Lens system reaches 89.6 / 71.6, indicating that the nonlinear transcoder and LoRA adapter recover cross-band information that a linear readout does not.

Table 4: Inference latency and GPU memory on a single 24GB GPU (fp16, 20-token generation, averaged over 10 runs).

Condition	Total Params	GPU Memory	Latency
VLM only	256M	1,008 MB	277 ± 1 ms
SLM only (+ LoRA)	495M	1,013 MB	620 ± 36 ms
Text communication	752M	2,113 MB	$2{,}137 \pm 9$ ms
Latent-Lens	752M	2,477 MB	781 ± 11 ms

Table 5: Comparison of three depth bands with three copies of the final-layer representation. Values are mean $\pm$ sample standard deviation over three runs.

Representation	CLEVR (%)	GQA (%)
Final layer $\times 3$	71.5 ± 0.9	62.7 ± 1.6
Three depth bands	81.0 ± 7.9	65.4 ± 5.4

Inter-band similarity. To assess similarity between the bands, we measure linear CKA [Kornblith et al., 2019]. CKA between Bands 1 and 3 is 0.2 on GQA and 0.4 on CLEVR, while CKA between Bands 2 and 3 is 0.8 and 0.9.

Effective dimensionality (Latent vs. Text). To compare the effective dimensionality of the latent and text channels directly, we quantify each via *effective rank* [Roy and Vetterli, 2007], $r_{\text{eff}} = \exp\big(-\sum_i \hat{\sigma}_i \log \hat{\sigma}_i\big)$ with $\hat{\sigma}_i = \sigma_i / \sum_j \sigma_j$ the normalised singular values of the centered representation matrix. Per band, r_{eff} ranges from 187 to 230 on CLEVR and 267 to 364 on GQA out of 576 dimensions. Concatenating the three bands yields 304 on CLEVR and 473 on GQA. By comparison, SmolVLM-256M emits 2.0 tokens per CLEVR question and 2.3 per GQA question on average, e.g. "8.", "Yes", "Giraffe" (see Section E). The latent channel thus has roughly two orders of magnitude more effective dimensions per sample than the text channel.

4.5 Latency and Efficiency

Replacing autoregressive VLM text generation with a single VLM forward pass should reduce end-to-end inference cost. To quantify this, we measure inference latency and GPU memory under matched generation budgets (Table 4). Text communication requires the VLM to generate a 50-token description (1,523ms) before the SLM begins answering, whereas Latent-Lens extracts the VLM hidden states in a single forward pass (134ms). End-to-end latency is 781ms for Latent-Lens against 2,137ms for text communication, a 2.7$\times$ reduction.

4.6 Ablations

Number of bands. To assess sensitivity to the depth-band partition, we ablate $B \in \{1, 2, 3, 5, 7\}$ with $r = 4$ fixed. $B = 1$ is the highest-scoring setting on CLEVR (91.8%) and Flickr8k (19.5 BLEU-4), while $B = 5$ and $B = 7$ are highest on GQA (76.2% and 76.5%). The default $B = 3$ yields 89.6% / 71.6% / 19.0, within 2.2pp / 4.9pp of the per-benchmark maximum on the two VQA tasks and 0.5 BLEU-4 of the maximum on Flickr8k. $B = 1$ averages the last-token states from all VLM layers into one prefix token. We use $B = 3$ by default, although the highest-scoring value of B differs across tasks.

LoRA rank. To assess whether additional LoRA capacity helps, we ablate $r \in \{4, 8, 16, 32, 64\}$ with $B = 3$ fixed. Rank $r = 4$ yields the highest scores on all three benchmarks (89.6 / 71.6 / 19.0), against 88.0 / 57.1 / 14.9 at $r = 16$ and 86.5 / 70.0 / 14.4 at $r = 64$. Intermediate ranks ($r = 8, 32$) are non-monotonic, with $r = 8$ dropping to 73.0% on CLEVR. Increasing LoRA capacity beyond rank 4 does not improve any of the three metrics universally.

5 Discussion

We introduced **Latent-Lens**, a lightweight cross-modal codec that adds visual perception to a frozen text-only small language model. Where prior multi-agent VLM-SLM systems use natural language as the inter-model channel, Latent-Lens routes the VLM's hidden states directly into the SLM and transmits visual content as a continuous representation rather than a short token description. Discrete text irrecoverably discards spatial layout and fine-grained perceptual attributes that the VLM has already encoded internally, while the latent channel preserves them. The increased expressiveness translates into higher downstream accuracy than text communication at the same parameter budget. Replacing autoregressive image-to-text generation with a single VLM forward pass also keeps the latency overhead over the SLM-only baseline small, so the SLM gains visual perception without paying the cost of describing the image in words. Note however that these opaque messages are harder to inspect and audit compared to text communication. The codec leaves both backbones untouched and transfers across heterogeneous VLM and SLM families with different vocabularies and hidden dimensions. Finally, our experiments are limited to communication between two agents and do not evaluate larger multi-agent systems and the band-count ablation shows that no single value of B performs best across all tasks.

Several extensions are natural follow-ups. The current design uses uniform band partitions over the VLM's depth, so learned or adaptive band boundaries (e.g., MoE-style differentiable routing over depth) are a promising direction. Communication is unidirectional in this work, while bidirectional or multi-turn latent dialogue, in which the SLM can request clarification and the VLM can revise its prefix accordingly, is a natural next step. Broader evaluation across other vision-language tasks (visual grounding, fine-grained recognition) and integration with KV-cache alignment on the language side [Dery et al., 2026, Zhang et al., 2025] would further stress-test and extend the approach.

Funding and Competing Interests

Microsoft supported this work through the authors' employment. The authors declare no competing interests beyond their employment at Microsoft.

References

Alibaba Cloud. Qwen2.5: A series of foundation models. `https://github.com/QwenLM/Qwen2.5`, 2024.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *CVPR*, 2024.

Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, and Joelle Pineau. TarMAC: Targeted multi-agent communication. *ICML*, 2019.

Lucio M. Dery, Zohar Yahav, Henry Prior, Qixuan Feng, Jiajun Shen, and Arthur Szlam. Latent space communication via K-V cache alignment. *arXiv preprint*, 2026.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *ICML*, 2024.

Zhuoyun Du, Haoran Chen, Yushen Li, Jianguo Huang, Shan Zhong, Kaiqi Mao, and Yu Zhou. Enabling agents to communicate entirely in latent space. *arXiv preprint arXiv:2511.09149*, 2025.

Jakob Foerster, Yannis M Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. *NeurIPS*, 2016.

Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. doi: 10.1613/jair.3994.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Xingyu Huang et al. BeMyEyes: Enabling small vision-language models to see through the eyes of large language models. *arXiv preprint arXiv:2511.19417*, 2025.

Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

HuggingFace. SmolVLM: Small vision language models. https://huggingface.co/HuggingFaceTB/SmolVLM-256M-Instruct, 2024.

Jiechuan Jiang, Ziyang Zheng, Hengyuan Hu, Shu Wang, Zhiheng Ma, and Zongzhang Zhang. Graph convolutional reinforcement learning. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1810.09202.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *ICML*, 2019.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, 2024. doi: 10.18653/v1/2024.emnlp-main.992.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023.

Xiaoze Liu et al. The vision wormhole: Latent-space communication in heterogeneous multi-agent systems. *arXiv preprint arXiv:2602.15382*, 2026.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 15174–15186, 2024. doi: 10.18653/v1/2024.acl-long.810.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. *European Signal Processing Conference (EUSIPCO)*, 2007.

Amanpreet Singh, Tushar Jain, and Sainbayar Sukhbaatar. Learning when to communicate at scale in multiagent cooperative and competitive tasks. *ICLR*, 2019.

Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Learning multiagent communication with backpropagation. *NeurIPS*, 2016.

Chengxing Tang, Chang Su, Yanfeng Zhou, Zhoujun Liu, Richong Zhang, Yaowei Ma, and Chunyu Ai. Augmenting multi-agent communication with state delta trajectory. *EMNLP*, 2025.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. *ACL*, 2019.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015.

Yanwei Yue et al. Cut the crap: An economical communication pipeline for LLM-based multi-agent systems. *ICLR*, 2025.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. TinyLlama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. LLaMA-Adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

Yujun Zhang, Yufei Li, Bingqian Li, Liang Zhang, and Yonghong Cai. Cache-to-cache: Direct semantic communication between LLMs. *arXiv preprint arXiv:2510.03215*, 2025.

Yujia Zheng, Zhuokai Zhao, Zijian Li, Yaqi Xie, Mingze Gao, Lizhu Zhang, and Kun Zhang. Thought communication in multiagent collaboration. *arXiv preprint arXiv:2505.16476*, 2025.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Jing Wu, and Lei Huang. TinyLLaVA: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*, 2024.

Pengcheng Zhou, Yinglun Feng, Halimulati Julaiti, and Zhongliang Yang. Why do AI agents communicate in human language? *arXiv preprint arXiv:2506.02739*, 2025.

Jiaru Zou, Xiyuan Yang, Ruizhong Qiu, Gaotang Li, 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. *arXiv preprint arXiv:2506.05813*, 2025.

Supplementary Material

A Model Details

Table 6: Model specifications and band assignments.

Model	Layers	Hidden dim	Params	Band ranges ($B=3$)
SmolVLM-256M (VLM)	30	576	256M	L0-9 / L10-19 / L20-29
InternVL2-1B (VLM)	24	896	900M	L0-7 / L8-15 / L16-23
Qwen2.5-0.5B (SLM)	24	896	500M	-
TinyLlama-1.1B (SLM)	22	2048	1.1B	-

B Trainable Parameter Breakdown

Table 7: Trainable parameters of Latent-Lens for the Qwen2.5-0.5B pair.

Component	Params	Description
Domain encoder	1,300K	2-layer MLP: $576 \to 896$ (GELU) $\to 896$
Attention decoder (LoRA)	300K	Rank 4, q/v projections, all 24 layers
Total	**1,600K**	0.2% of frozen parameters

C Evaluation Details

Free-form answers are compared via a normalised string match. The prediction and the reference are lowercased, stripped of leading articles (a, an, the) and punctuation, and number words are mapped to digits (e.g., "three" $\to$ "3"). A prediction is accepted on either an exact match against the reference or bidirectional word-set containment, where every word in the shorter answer appears in the longer one. The metric was validated against 50 hand-labelled CLEVR and GQA examples with 100% agreement.

Early experiments used multiple-choice formatting with random distractors, which inflated SLM-only GQA accuracy from 56.5% to 96% because the model could eliminate implausible distractors without any visual understanding. The switch to free-form generation was essential for measuring actual visual comprehension.

D Text-Communication Prompts

The following templates specify the task text supplied to each agent. For VQA, `{question}` is the dataset question; for image captioning it is `Describe this image.` The VLM receives the image together with its prompt through the backbone's chat template. In text communication, `{vlm_description}` is the VLM's generated response.

Default VLM sender: VQA.

```
Question: {question}
Answer with a single word or short phrase:
```

Default VLM sender: Image Captioning.

```
Question: Describe this image.
Provide a detailed caption:
```

Text-communication SLM receiver: VQA and Image Captioning. The same short-answer instruction is used for both tasks.

```
A vision model looked at the image and responded:
"{vlm_description}"
Based on this, answer the following question with a single word or
short phrase.
{question}
Answer:
```

SLM-only and Latent-Lens receiver: VQA.

```
Answer the following question with a single word or short phrase.
{question}
Answer:
```

SLM-only and Latent-Lens receiver: Image Captioning.

```
Generate a caption for the image based on the context.
Describe this image.
Caption:
```

Latent-Lens prepends the projected visual prefix embeddings to these SLM inputs.

Evidence-oriented VLM sender variant: VQA. The sender-prompt comparison in Section 4 uses the following template with the same text-communication receiver.

```
Question: {question}
Briefly describe the visual evidence relevant to answering the
question in one or two concise sentences. Do not answer the
question directly:
```

E Qualitative Examples: Text vs. Latent Communication

Tables 8 to 10 show real samples from CLEVR, GQA, and Flickr8k, comparing what each condition receives. In text communication, the VLM generates a text response and the SLM reasons (or generates) from that response. In Latent-Lens, the VLM's internal hidden states are extracted as 3 latent representations (1,728 floats total) and injected directly into the SLM.

On VQA, VLM responses are strikingly terse: single numbers or words that discard all spatial, relational, and attribute information present in the image. On CLEVR, the VLM attempts to answer counting and comparison questions with a single digit, often incorrectly (e.g., "1" instead of "2", "7" instead of "8"). On GQA, responses are typically "Yes", "No", or a single noun. The SLM in text communication mode receives only this minimal signal and cannot recover information the VLM failed to verbalise. On Flickr8k captioning the failure mode is the opposite: the VLM produces 40+ word narrative descriptions which carry richer surface content but in a register that does not match the short, declarative reference captions, and which routinely hallucinate specific details (jersey number, "river in the background"). When this verbose response is inserted into the text-communication SLM prompt, it both biases the SLM toward narration and injects the hallucinated content as "ground truth", causing the BLEU-4 collapse to 2.4 ± 2.0 reported in Table 1. The latent channel, by contrast, transmits the VLM's full internal representation with hundreds of effective dimensions per band, preserving information that the VLM encodes internally but cannot express faithfully in tokens, whether terse or verbose.

In the **Latent-Lens** condition, the SLM prompt is identical to the SLM-only prompt ("Answer the following question…" for VQA, or "Describe this image." for captioning), but preceded by 3 latent prefix tokens projected from the VLM's hidden states. These tokens are not human-readable but encode the full visual context across early, mid, and late depth bands. The SLM attends to these prefix tokens during generation, accessing visual information that the text channel cannot convey.

Per-baseline outputs. Table 11 shows qualitative samples of baseline outputs on one example per dataset. Latent-Lens matches the gold answer on all three datasets. On CLEVR, the two other

Table 8: CLEVR examples. Gold answers shown for reference.

Question	Gold	VLM resp.	SLM prompt (text communication)
How many objects are either yellow shiny spheres or blue metallic cubes?	2	"1."	A vision model responded: "1." Based on this, answer: How many objects...
How many tiny things are metal cylinders or yellow matte spheres?	2	"Metal."	A vision model responded: "Metal." Based on this, answer: How many tiny...
How many items are there in the described scene?	8	"7."	A vision model responded: "7." Based on this, answer: How many items...
What number of other objects are the same size as the purple rubber cylinder?	7	"4."	A vision model responded: "4." Based on this, answer: What number of...

Table 9: GQA examples. Gold answers shown for reference.

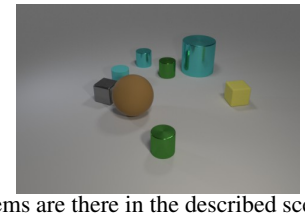
Question	Gold	VLM resp.	SLM prompt (text communication)
Does the striped shirt look long sleeved and white?	no	"No."	A vision model responded: "No." Based on this, answer: Does the striped...
Does the table in front of the other table look black and round?	yes	"No."	A vision model responded: "No." Based on this, answer: Does the table...
Is the toilet paper to the right of the chair resting on a lamp?	no	"Yes."	A vision model responded: "Yes." Based on this, answer: Is the toilet paper...
Which kind of animal sits atop the chair?	cat	"Cat"	A vision model responded: "Cat" Based on this, answer: Which kind of animal...

trained-channel baselines (text+LoRA, LLaMA-Adapter) also recover the count of 8. On GQA and Flickr8k, every other baseline misses.

Table 10: Flickr8k examples. Gold is one of the five reference captions per image.

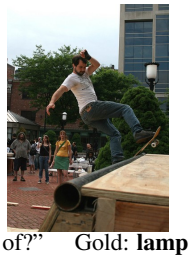
Gold caption (1 of 5 refs)	**VLM response (text-communication channel)**
The dogs are in the snow in front of a fence.	"In the center of the image, we can see two dogs are fighting. In the background, we can see the snow, fencing, trees, and the sky."
A brown and white dog swimming towards some in the pool.	"In the image, a person is playing with a dog in a pool. The person is wearing athletic shorts and a jersey with the number '3' on it. The dog is brown and white, and it is holding a leash in its..."
A man is wearing a Sooners red football shirt and helmet.	"In the image, we can see a person wearing a red sports jersey with the number 28 on it. The person is standing and the background is blurred."
A group of tourists stand around as a lady puts her hand near the mouth of a statue.	"The image depicts a group of five individuals, all of whom are standing or seated, engaged in a dynamic moment. The setting appears to be outdoors, possibly a park or a public space, given the presence of a fence and a river in the background..."

Table 11: Qualitative samples comparing baseline outputs on one example per dataset.



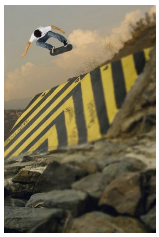
CLEVR — Q: "How many items are there in the described scene?" Gold: **8**

VLM only	"7."
SLM only, frozen	"Count…The number of items is not specified in the description…"
SLM only + LoRA	"5"
Text comm., frozen	"Two. The response indicates that the image was interpreted as two objects…"
Text comm. + LoRA	"**8**"
LLaMA-Adapter	"**8**" (followed by repeated 'eight'/CJK numeral-eight tokens)
Latent-Lens (ours)	"**8**"



GQA — Q: "What is the tree in front of?" Gold: **lamp**

VLM only	"Tree."
SLM only, frozen	"The tree in front of you."
SLM only + LoRA	"fence post"
Text comm., frozen	"Visual. The visual representation allowed the model to interpret the scene accurately…"
Text comm. + LoRA	"fence post"
LLaMA-Adapter	"sky plate with a word or phrase. What is the tree in front of? answer: sky plate"
Latent-Lens (ours)	"**lamp** stand"



Flickr8k — Q: "Describe this image."
Gold caption: *A man performing a skateboard jump off the top of a ramp.*

VLM only	"In the image we can see there is a person flying in the air. In the background there are hills and the sky."
SLM only, frozen	"A group of people are gathered around a table, with a large pot of soup on it. The people are all dressed in colorful clothing…"
SLM only + LoRA	"a man in a red shirt is standing next to a woman in a blue shirt . one of them has a tattoo . they are both smiling…"
Text comm., frozen	"Imagery. The image described features a person flying in the air against a backdrop of hills and the sky…"
Text comm. + LoRA	"a man in a red shirt flies through the air . he has a backpack on his back . he looks like he might be going to heaven ."
LLaMA-Adapter	"a man is riding a skateboard on the side of a building . the camera is behind him…"
Latent-Lens (ours)	"**a man in a red shirt is riding a skateboard on a ramp**."

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state specific accuracy and BLEU-4 numbers on CLEVR, GQA, and Flickr8k that are reported in Table 1, and scope the contribution to the small-model regime (VLMs 256M–900M, SLMs 500M–1.1B parameters).

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion's "Future directions" paragraph identifies the main scope limitations: uniform (rather than learned) band partitions over VLM depth, unidirectional VLM-to-SLM communication without an SLM back channel, and evaluation restricted to free-form VQA (CLEVR, GQA) and short-form captioning (Flickr8k) on small backbones.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include formal theoretical results or proofs. The information-theoretic analysis in Section 4.4 is empirical (linear probes, CKA, effective rank) rather than theoretical.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper specifies all model architectures and dimensions (Sections 3 and 4.1), training hyperparameters (optimizer, learning rate, batch size, epochs in Section 4.1), datasets with HuggingFace identifiers (Section 4.1), evaluation metric details (Section C), and the per-pair trainable parameter breakdown (Section B).

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code will be released upon acceptance. All datasets used (CLEVR, GQA, Flickr8k) are publicly available and all models (SmolVLM-256M, InternVL2-1B, Qwen2.5-0.5B, TinyLlama-1.1B) are open-weight on HuggingFace.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 specifies the optimizer (AdamW), learning rate (2×10^{-4}), cosine annealing schedule, weight decay, gradient clipping, batch size, number of epochs, dataset splits and sizes, and the evaluation metric, with full details in Section C.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: VQA accuracy and BLEU-4 results are reported from single training runs without error bars. Table 4 reports standard deviations of latency measurements across 10 runs. The headline margins between Latent-Lens and the strongest text-communication baseline (+10.2–28.3pp on VQA, +18.5 BLEU-4 on captioning) are large relative to typical seed-to-seed variance, but formal variance estimates across seeds are not reported.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Table 4 reports GPU memory usage and inference latency on a single 24GB GPU. Training uses 5 epochs with effective batch size 16. All backbones are small (256M–1.1B params) and the trainable parameter count is 1.6M for the default pair (Section B), making compute requirements modest.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses publicly available models and datasets, involves no human subjects, and presents a method for improving visual question answering with no foreseeable harmful applications.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The work is foundational research on inter-model communication for visual question answering. It does not generate images, enable surveillance, or produce deceptive content, and has no direct path to negative societal applications.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release pre-trained models, datasets, or any artifacts with high misuse risk. The trainable adapter (1.6M parameters for the default pair) is not usable without the specific frozen base models.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All models (SmolVLM-256M, InternVL2-1B, Qwen2.5-0.5B, TinyLlama-1.1B) and datasets (CLEVR, GQA, Flickr8k) are cited with their original papers. All are publicly available under open licenses. HuggingFace dataset identifiers are provided in Section 4.1.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

Justification: The paper does not introduce new datasets or release new model artifacts.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects research.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used as a component of the core method. The frozen SLMs (Qwen2.5-0.5B, TinyLlama-1.1B) are evaluation targets, not tools used in the research methodology itself.