SEMANTIC ROLE PROJECTION FOR COMPOSITIONAL REASONING IN MULTIMODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) rely on learned projectors to translate visual patch features into the language model's token embedding space. We identify a fundamental limitation of standard projectors, which we call *semantic role flattening*: object identity, attribute, count, and spatial relation information are collapsed into an undifferentiated embedding region, even though the LLM's own text embeddings maintain clear role-specific subspace structure. We propose the **Semantic Role Projector (SRP)**, which replaces the standard MLP projector with four parallel typed projection heads supervised by role-alignment and inter-role orthogonality losses derived from the LLM's frozen text-embedding subspaces. A lightweight learned router produces soft role assignments per visual patch, and a mixing layer preserves drop-in compatibility with existing MLLM architectures. We prove that, under mild spectral assumptions, SRP increases the attention logit gap for role-selective heads whenever the alignment residual is small and head selectivity is high, providing a mechanistic explanation for its compositional advantage. Experiments on CLEVR-CoGenT, CLOSURE, GQA, TallyQA, and VQAv2 demonstrate that SRP achieves state-of-the-art compositional generalization among learned projector-based MLLMs, improving CoGenT Condition B accuracy by $+3.6$ pp over the best modular baseline (Honeybee) and $+5.3$ pp over a capacity-matched Mixture-of-Experts, while adding 3.4% wall-clock training overhead. Ablations confirm that each geometric loss component is essential, and cross-architecture evaluation on Qwen-VL-7B verifies that the gains transfer across LLM backbones.

1 INTRODUCTION

Multimodal large language models (MLLMs) have rapidly advanced visual question answering by coupling frozen vision encoders with large language models through learned visual projectors (Liu et al., 2024b; Bai et al., 2023; 2025). The projector, typically a two-layer MLP, maps visual patch embeddings into the LLM's token space, where they participate in the same self-attention computation as text tokens. While this simple interface works well for holistic scene understanding, it creates a bottleneck for *compositional* visual reasoning: questions such as "How many red cubes are to the left of the large blue sphere?" require the model to simultaneously resolve object identity, attributes, spatial relations, and numerosity from the projected visual tokens.

We hypothesize that standard projectors suffer from **semantic role flattening**: they map all visual information into a single, undifferentiated region of embedding space, collapsing the role-specific structure that the LLM's own text representations naturally maintain. This collapse is consequential because the LLM's attention heads cannot selectively retrieve, say, spatial relations without also activating entangled object and attribute features, a prerequisite for compositional multi-step reasoning. To test this hypothesis, we train linear probes to classify token embeddings into four semantic role types (object, attribute, count, and relation) and find that text probes achieve $>90\%$ four-way accuracy while visual probes plateau at $<60\%$ (Section 3.2). This diagnostic gap motivates our approach.

We propose the **Semantic Role Projector (SRP)** (Figure 1), which replaces the standard MLP with four parallel typed projection heads, each aligned to the corresponding role subspace of the LLM's text embeddings. A lightweight router produces soft per-patch role assignments, and a gated mixing

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.

layer preserves the standard token interface, so no LLM changes are needed. The subspaces are estimated once from the frozen LLM's contextual embeddings via PCA and remain fixed.

Our theory (Section 3.4) shows that role-structured projections enlarge the attention logit gap for role-selective heads whenever the alignment residual is small and head selectivity is high (Proposition 1).

Extensive experiments across synthetic (CLEVR-CoGenT, CLOSURE) and naturalistic (GQA, TallyQA, VQAv2) benchmarks demonstrate that SRP achieves substantial improvements on compositional reasoning while matching baselines on non-compositional tasks. We make the following contributions:

- **Diagnosis.** We identify *semantic role flattening* and quantify it with linear probes: text embeddings separate the four roles at $>90\%$ accuracy, while visual embeddings from standard MLP projectors plateau below 60% (Section 3.2).
- **Method.** We introduce SRP, a drop-in projector that maps each patch through four typed heads aligned to the LLM's frozen text-embedding subspaces, with a learned router and mixing layer (Section 3.3).
- **Theory.** We prove that role-selective heads obtain a larger attention logit gap on SRP tokens than on MLP tokens under explicit spectral and alignment conditions (Proposition 1).
- **Compositional generalization.** SRP attains the best non-oracle accuracy on CoGenT-B (78.4%, $+5.3$ pp over a capacity-matched MoE and $+3.6$ pp over Honeybee) and on the CLOSURE seven-split average ($+7.0$ pp over the MoE), and raises role probe accuracy from 58.1% to 82.4% (Tables 1 and 4).
- **Efficiency and transfer.** The gains cost 3.4% wall-clock training overhead, transfer to Qwen-VL-7B, and degrade gracefully under 20% bounding-box noise (Tables 5 and 4).

2 Related Work

Multimodal large language models. The dominant MLLM paradigm connects a frozen vision encoder to an LLM through a learned projector. LLaVA (Liu et al., 2024b) uses a simple linear or two-layer MLP projection; subsequent work has explored more expressive interfaces including Q-Former's learned query tokens (Li et al., 2023), the two locality-enhanced abstractors introduced by Honeybee, namely the C-Abstractor and the D-Abstractor (Cha et al., 2024), and mixture-of-experts projectors (Shazeer et al., 2017; Lin et al., 2024); in our tables the *C-Abstractor* row denotes the convolutional variant and the *Honeybee* row its best-reported (D-Abstractor) configuration. These architectures increase projector capacity but treat visual information as an undifferentiated stream. SRP is complementary: it introduces *typed* projection heads that decompose visual information along semantic role axes, and can in principle be combined with any of these interfaces. The contrast with pure capacity is empirical: our Vanilla MoE control shares SRP's heads and router but receives no role losses and recovers only a small part of the gain (Table 4), so added capacity or expert routing alone cannot explain the improvement.

Compositional visual reasoning. Compositional generalization in vision-language models has been studied through synthetic benchmarks including CLEVR (Johnson et al., 2017), CLEVR-CoGenT (Johnson et al., 2017), and CLOSURE (Bahdanau et al., 2020), as well as naturalistic benchmarks such as GQA (Hudson & Manning, 2019). Neuro-symbolic approaches like NS-VQA (Yi et al., 2018) achieve near-perfect accuracy on synthetic tasks but require ground-truth programmatic supervision unavailable for natural images. Recent neural approaches explore modular networks (Andreas et al., 2016), program synthesis (Mao et al., 2019), and disentangled representations (Locatello et al., 2019), but none explicitly align the visual projector's output geometry with the LLM's text-embedding subspace structure. SRP bridges this gap by leveraging the LLM's own linguistic geometry as a structural prior for the projector.

Representation alignment and subspace methods. Cross-modal alignment has been explored through contrastive learning (Radford et al., 2021), optimal transport (Chen et al., 2020), and subspace-based methods. In NLP, subspace analysis has revealed role-specific structure in contextual embeddings (Tenney et al., 2019; Hewitt & Manning, 2019), and probing classifiers are widely used to assess representational geometry (Conneau et al., 2018). The mechanistic interpretability literature has documented syntactically specialized attention heads in transformers (Clark et al., 2019; Voita et al., 2019; Elhage et al., 2021). SRP uniquely combines these insights: it uses probing-derived

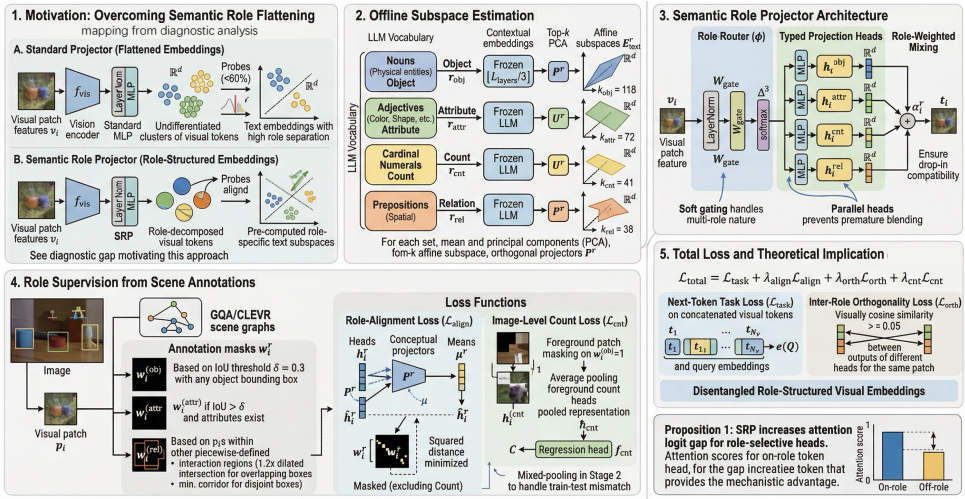

Figure 1: Overview of the Semantic Role Projector (SRP). (1) Motivation: SRP resolves semantic role flattening by maintaining distinct object, attribute, count, and spatial relation structures. (2) Estimation: fixed affine subspaces are derived from a frozen LLM's text embeddings. (3) Architecture: a learned router soft-gates visual patches into four parallel typed projection heads, and a mixing layer preserves drop-in compatibility. (4) Supervision: structured scene graphs provide supervision masks for role-alignment, count, and inter-role orthogonality losses. (5) Theory: combined with the task loss, this geometry widens the attention logit gap for role-selective heads under the stated spectral and alignment conditions.

subspace estimates to define explicit geometric training objectives for the visual projector, grounded in the attention head specialization literature.

3 Methodology

3.1 Notation and Problem Formulation

Notation. An MLLM consists of a frozen vision encoder $f_{\text{vis}} : \mathcal{V} \to \mathbb{R}^{N_v \times d_v}$, a **projector** $g : \mathbb{R}^{N_v \times d_v} \to \mathbb{R}^{N_v \times d}$, and an LLM $\mathcal{M}$ with L_{layers} transformer layers (Vaswani et al., 2017) over d-dimensional embeddings. We define four **semantic role types** $\mathcal{R} = \{r_{\text{obj}}, r_{\text{attr}}, r_{\text{cnt}}, r_{\text{rel}}\}$ (object, attribute, count, relation), with soft assignments in the 3-simplex Δ^3. Let $\mathbf{E}^{(r)}_{\text{text}} \subseteq \mathbb{R}^d$ denote the empirical subspace of role r, estimated from middle-layer text embeddings (Section 3.3.1), and $\mathcal{S}_j$ the spatial relations of scene-graph object j. Given image I and query Q, the MLLM produces $A = \mathcal{M}([g(f_{\text{vis}}(I)); \mathbf{e}(Q)])$.

Objective. We seek a projector g^* producing **role-structured visual embeddings** that decompose each patch's information along semantic role axes into the corresponding subspaces $\mathbf{E}^{(r)}_{\text{text}}$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{orth}}\mathcal{L}_{\text{orth}} + \lambda_{\text{cnt}}\mathcal{L}_{\text{cnt}}, \tag{1}$$

where $\mathcal{L}_{\text{align}}$ encourages subspace alignment, $\mathcal{L}_{\text{orth}}$ enforces inter-role separation, and $\mathcal{L}_{\text{cnt}}$ supervises counting (Section 3.3.3).

3.2 Diagnostic Analysis: Semantic Role Flattening

Before presenting our method, we formalize the diagnostic that motivates it. We measure the **semantic role separability** of embeddings in $\mathbb{R}^d$ by training linear probes (logistic regression classifiers) to predict the role $r \in \mathcal{R}$ from individual token embeddings.

Text embeddings. We sample visually relevant vocabulary (concrete nouns, visual adjectives, cardinal numerals, and spatial prepositions; Section 3.3.1) and extract **contextual embeddings** by passing each token through the LLM in a minimal template ("This is a `[TOKEN]`") and recording

activations at layer $\ell^* = \lfloor L_{\text{layers}}/3 \rfloor$, which best balances lexical identity with learned compositional structure (Tenney et al., 2019). We train a probe $h_{\text{text}} : \mathbb{R}^d \to \mathcal{R}$ on these contextual embeddings.

Visual embeddings. Using scene graphs from GQA and CLEVR, we assign role labels to visual patches as described in Section 3.3.4 (excluding count, which is image-level). We pass images through f_{vis} followed by the baseline MLP projector g_{MLP} and train an analogous probe $h_{\text{vis}} : \mathbb{R}^d \to \mathcal{R}$.

Across LLaVA-1.5-7B and Qwen-VL-7B, text probes achieve $>90\%$ four-way classification accuracy while visual probes plateau at $<60\%$, confirming **semantic role flattening**: the MLP projector collapses role-distinguishing structure. An ablation comparing static embeddings (from the token embedding matrix $\mathbf{W}_E$) versus contextual embeddings shows that contextual embeddings yield more separable clusters (probe accuracy: 93.2% contextual vs. 88.7% static; see Appendix E), supporting our use of contextual embeddings for subspace estimation.

3.3 Proposed Method: Semantic Role Projector (SRP)

3.3.1 Text-Embedding Subspace Estimation

For each role $r \in \mathcal{R}$, we collect a reference set $\mathcal{S}^{(r)}$ of **contextual** text token embeddings. Tokens are selected from the LLM's vocabulary and filtered to retain only **visually-relevant** entries:

- $r_{\textbf{obj}}$ **(object):** Concrete nouns denoting physical objects, selected by intersecting the vocabulary with WordNet (Miller, 1995) synsets under *physical_entity.n.01*. Resulting set: $\sim$6,800 tokens.
- $r_{\textbf{attr}}$ **(attribute):** Visual adjectives restricted to curated semantic categories: color (148 terms), shape (43), size (26), material/texture (51). Total: $\sim$270 unique tokens.
- $r_{\textbf{cnt}}$ **(count):** Cardinal numerals and number words: digits "0"–"99" plus spelled-out forms and quantifiers ("few," "many," "several," "none"). Total: $\sim$260 tokens.
- $r_{\textbf{rel}}$ **(relation):** A closed set of 23 spatial prepositions and adverbs: {*above, below, left, right, behind, front, near, far, inside, outside, between, beside, over, under, across, along, around, through, toward, away, adjacent, opposite, atop*}.

For each token in $\mathcal{S}^{(r)}$, we extract its contextual embedding at layer $\ell^* = \lfloor L_{\text{layers}}/3 \rfloor$ (layer 10 of 32 for both 7B models). We compute the empirical mean $\boldsymbol{\mu}^{(r)}$ and perform PCA on the centered embeddings to obtain the top-k principal components $\mathbf{U}^{(r)} \in \mathbb{R}^{d \times k}$, where $\mathbf{U}^{(r)\top}\mathbf{U}^{(r)} = \mathbf{I}_k$. The **role subspace** is the affine subspace $\mathbf{E}^{(r)}_{\text{text}} = \{\boldsymbol{\mu}^{(r)} + \mathbf{U}^{(r)}\mathbf{z} : \mathbf{z} \in \mathbb{R}^k\}$, and the orthogonal projector onto the linear part is $\mathbf{P}^{(r)} = \mathbf{U}^{(r)}\mathbf{U}^{(r)\top} \in \mathbb{R}^{d \times d}$. These parameters $\{(\boldsymbol{\mu}^{(r)}, \mathbf{U}^{(r)})\}_{r \in \mathcal{R}}$ are computed once and frozen.

Choice of k. We set k such that $\geq 90\%$ of each role's variance is explained. For LLaVA-1.5-7B ($d = 4096$), this yields $k_{\text{obj}} = 118$, $k_{\text{attr}} = 72$, $k_{\text{cnt}} = 41$, $k_{\text{rel}} = 38$. Sensitivity analysis (Appendix E) shows downstream accuracy varies by <1.8 percentage points across $k \in \{32, 64, 128, 256\}$, while unfiltered POS-tagged sets yield $\sim$3 pp lower downstream accuracy, confirming that domain-appropriate filtering is necessary.

3.3.2 Architecture

The **Semantic Role Projector** replaces the standard two-layer MLP projector with four parallel **typed projection heads** and a learned **role router**, followed by a mixing layer. Given a visual patch feature $\mathbf{v}_i \in \mathbb{R}^{d_v}$:

1. **Role Router.** A lightweight gating network $\phi : \mathbb{R}^{d_v} \to \Delta^{|\mathcal{R}|-1}$ produces soft role assignments:
$$\boldsymbol{\alpha}_i = \text{softmax}(\mathbf{W}_{\text{gate}}\ \text{LayerNorm}(\mathbf{v}_i) + \mathbf{b}_{\text{gate}}) \in \mathbb{R}^{|\mathcal{R}|}, \tag{2}$$
where $\mathbf{W}_{\text{gate}} \in \mathbb{R}^{|\mathcal{R}| \times d_v}$, $\mathbf{b}_{\text{gate}} \in \mathbb{R}^{|\mathcal{R}|}$.
2. **Typed Projection Heads.** For each role $r \in \mathcal{R}$, a two-layer MLP with GELU activation:
$$\mathbf{h}_i^{(r)} = \mathbf{W}_2^{(r)}\ \text{GELU}\Big(\mathbf{W}_1^{(r)}\,\mathbf{v}_i + \mathbf{b}_1^{(r)}\Big) + \mathbf{b}_2^{(r)}, \quad \mathbf{h}_i^{(r)} \in \mathbb{R}^d, \tag{3}$$

with parameters $\theta^{(r)} = \{\mathbf{W}_1^{(r)} \in \mathbb{R}^{d_m \times d_v}, \mathbf{b}_1^{(r)}, \mathbf{W}_2^{(r)} \in \mathbb{R}^{d \times d_m}, \mathbf{b}_2^{(r)}\}$ and intermediate dimension d_m.

3. **Role-Weighted Mixing.** The final visual token embedding is:

$$\mathbf{t}_i = \sum_{r \in \mathcal{R}} \alpha_i^{(r)} \, \mathbf{h}_i^{(r)} \in \mathbb{R}^d. \tag{4}$$

The output $(\mathbf{t}_1, \ldots, \mathbf{t}_{N_v})$ has exactly the same shape as a standard MLP projector's output, ensuring *drop-in compatibility*. Parallel heads ensure role-specific information is never blended prematurely; soft gating handles the inherent multi-role nature of visual patches; and the mixing step preserves the standard MLLM token interface, requiring no LLM architectural changes.

3.3.3 Loss Functions

Task loss. Standard autoregressive cross-entropy $\mathcal{L}_{\text{task}} = -\sum_j \log p_{\mathcal{M}}(a_j \mid [\mathbf{t}_{1:N_v}; \mathbf{e}(Q); a_{<j}])$.

Role-alignment loss. For each patch i and role r, let $\hat{\mathbf{h}}_i^{(r)} = \boldsymbol{\mu}^{(r)} + \mathbf{P}^{(r)}(\mathbf{h}_i^{(r)} - \boldsymbol{\mu}^{(r)})$ be the subspace projection. The alignment loss penalizes the residual:

$$\mathcal{L}_{\text{align}} = \frac{1}{\sum_{i,r} w_i^{(r)}} \sum_{i=1}^{N_v} \sum_{r \in \{r_{\text{obj}}, r_{\text{attr}}, r_{\text{rel}}\}} w_i^{(r)} \left\| \mathbf{h}_i^{(r)} - \hat{\mathbf{h}}_i^{(r)} \right\|_2^2, \tag{5}$$

where $w_i^{(r)} \in \{0, 1\}$ is a supervision mask from scene annotations (Section 3.3.4). The count role is excluded and handled separately below.

Inter-role orthogonality loss. $\mathcal{L}_{\text{orth}} = \frac{1}{N_v} \sum_i \sum_{r \neq r'} \max(0, \cos(\mathbf{h}_i^{(r)}, \mathbf{h}_i^{(r')}) - \tau)$ with margin $\tau = 0.05$ (below the observed inter-subspace cosine of $\sim$0.08).

Image-level count loss. We average-pool the count head over foreground patches $\bar{\mathbf{h}}_{\text{cnt}} = \text{mean}_{i: w_i^{(r_{\text{obj}})} = 1} \mathbf{h}_i^{(r_{\text{cnt}})}$ and regress via $f_{\text{cnt}} : \mathbb{R}^d \to \mathbb{R}$: $\mathcal{L}_{\text{cnt}} = (f_{\text{cnt}}(\bar{\mathbf{h}}_{\text{cnt}}) - C)^2$, where C is the scene-graph object count. At inference, the count head pools over **all** N_v patches; we bridge the train–test mismatch via a **mixed-pooling schedule** that randomly substitutes all-patch pooling with probability 0.5 during Stage 2 training (Appendix G).

3.3.4 Role Supervision from Scene Annotations

We derive supervision masks $w_i^{(r)}$ from structured scene annotations in CLEVR (programmatic scene graphs) and GQA (human-annotated scene graphs). Let $\{(o_j, b_j, \mathcal{A}_j, \mathcal{S}_j)\}_{j=1}^M$ be the scene graph for an image. For each visual patch i with spatial extent p_i:

- $w_i^{(r_{\text{obj}})} = 1$ if $\text{IoU}(p_i, b_j) > \delta$ for any object j;
- $w_i^{(r_{\text{attr}})} = 1$ if $\text{IoU}(p_i, b_j) > \delta$ and $|\mathcal{A}_j| > 0$ for some j;
- $w_i^{(r_{\text{rel}})} = 1$ if patch p_i lies within the **interaction region** of two related objects (j, k) linked by a spatial relation $s \in \mathcal{S}_j$.

Note that a single patch may receive positive masks for multiple roles simultaneously (e.g., a patch overlapping an attributed object near a relation corridor), which is handled correctly by the per-role summation in $\mathcal{L}_{\text{align}}$. The count role r_{cnt} has no patch-level mask; it is supervised exclusively at the image level via $\mathcal{L}_{\text{cnt}}$.

The **interaction region** is defined piecewise: for overlapping boxes ($\text{IoU}(b_j, b_k) > 0$) it is the $1.2\times$-dilated intersection $\text{dilate}(b_j \cap b_k, 1.2)$, and for disjoint boxes it is $\text{dilate}(\text{bbox}(\text{convexhull}(b_j \cup b_k)), 1.2)$, forming a minimal corridor between them. We set $\delta = 0.3$; sensitivity analysis (Appendix E) shows stable performance for $\delta \in [0.2, 0.4]$, and on average 68% of patches receive at least one positive role mask per CLEVR image and 54% per GQA image, with unsupervised patches trained only via $\mathcal{L}_{\text{task}}$.

3.4 Theoretical Grounding

We formalize the advantage of role-structured projections. We state three assumptions (formal versions in Appendix C): *(i) Subspace separation:* the role subspaces $\{\mathbf{P}^{(r)}\}$ are pairwise separated in Frobenius norm. *(ii) Role-selective attention head:* there exists an attention head whose key projection $\mathbf{W}_K$ concentrates energy on a target role subspace r^* while attenuating others by a factor $\kappa > 1$, and whose query-key bilinear form is positive definite on the r^* subspace (Clark et al., 2019; Voita et al., 2019; Elhage et al., 2021). *(iii) Alignment convergence:* after training, each SRP head's output lies within ϵ of its target subspace, while the MLP projector retains only a ρ-fraction ($\rho \approx 0.4$) of the in-subspace energy.

Proposition 1 (Attention Logit Gap)**.** *Under Assumptions (i)–(iii), for a role-r^*-selective attention head and a co-directional text query $\mathbf{e}_q \in \mathbf{E}_{\text{text}}^{(r^*)}$ with $\|\mathbf{z}_q\|_2 = \sigma_q > 0$:*

$$\frac{\mathbf{q}^\top \mathbf{W}_K^\top \mathbf{t}_i^{\text{SRP}}}{\sqrt{d_h}} - \frac{\mathbf{q}^\top \mathbf{W}_K^\top \mathbf{t}_i^{\text{MLP}}}{\sqrt{d_h}} \geq \frac{1}{\sqrt{d_h}} \left[\alpha_{\min} \left(\beta_{qk}^2 \sigma_q s_i (1-\rho) - \beta_q \beta_k \epsilon\right) - B_{\text{cross}} - B_\perp\right], \quad (6)$$

where $B_{\text{cross}} = \mathcal{O}(\beta_q \beta_k / \kappa)$ bounds cross-role interference and $B_\perp$ bounds the out-of-subspace MLP contribution. When ϵ is sufficiently small and κ sufficiently large, the SRP logit strictly exceeds the MLP logit.

For compositional queries requiring multi-role attention across successive layers, we expect the per-head logit gap to compound along the reasoning chain; we treat this as a motivating observation rather than a formal guarantee. The full proof, formal assumption statements, and multi-head extension are provided in Appendix C.

4 Experiments

We evaluate SRP on synthetic benchmarks (CLEVR-CoGenT, CLOSURE) and naturalistic benchmarks (GQA, GQA-OOD, TallyQA, VQAv2) against an extensive baseline suite spanning foundational controls (Linear, MLP-2L, LoRA-only), capacity-matched controls (MLP-4×, Vanilla MoE at 37.8M parameters), and modular interfaces (C-Abstractor, Q-Former, Honeybee). All results report exact-match accuracy. Each training configuration is repeated over five random seeds and we report the mean, while probing results average over five outer cross-validation folds (Appendix D). Complete dataset descriptions, baseline specifications, evaluation metrics, and training hyperparameters are provided in Appendix D.

4.1 Main Results on Compositional Generalization

Table 1 presents the primary compositional generalization results. **(i) Capacity alone is insufficient.** The capacity-matched MLP-4× and Vanilla MoE gain little over MLP-2L despite a 3× parameter increase, so unconditionally scaling projector capacity has diminishing returns for compositional reasoning. **(ii) SRP achieves substantial gains.** On the critical CoGenT-B out-of-distribution split, SRP surpasses the best modular interface (Honeybee) by +3.6 pp and the capacity-matched MoE by +5.3 pp, showing that aligning the visual latent space to textual subspace geometry helps deeply nested multi-entity reasoning. **(iii) Cross-architecture transfer.** The same recipe on Qwen-VL-7B (Bai et al., 2023) gains +4.2 pp on CoGenT-B over Qwen-MoE, so the alignment is architecture-agnostic. **(iv) GQA-OOD stress test.** SRP widens its lead on rare relational compositions (+2.8 pp over Honeybee on GQA-OOD).

Ablation analysis (Table 2) isolates each component. Removing $\mathcal{L}_{\text{align}}$ causes the largest single-factor drop (−3.6 pp on CoGenT-B), so semantic subspace alignment is the dominant driver; removing $\mathcal{L}_{\text{orth}}$ costs −1.6 pp, showing that inter-role separation limits representational interference; and collapsing to a shared head eliminates the gains entirely, proving that distinct parameter pathways are obligatory. Static subspaces cost −2.8 pp versus contextual subspaces, validating contextual estimation (Section 3.3.1), and unfiltered vocabulary costs −3.0 pp, consistent with the filtering analysis.

Table 1: Main compositional generalization results. **Bold** marks the best result per column within each backbone group; the oracle row is excluded. CL-avg: CLOSURE seven-split average. Params: projector + LoRA total.

Method	Params	CoGenT-A	CoGenT-B	CL-avg	GQA	GQA-comp	GQA-OOD
Foundational controls							
Linear	4.2M	69.1	65.3	58.4	55.2	46.1	40.2
MLP-2L	12.6M	82.5	72.1	65.2	61.3	52.4	46.5
LoRA-only	4.2M	75.3	69.8	62.1	58.7	49.8	43.1
Capacity-matched controls							
MLP-4×	37.8M	83.1	72.5	65.8	61.7	52.9	47.1
Vanilla MoE	37.8M	83.6	73.1	66.4	62.1	53.5	47.8
Modular interfaces							
C-Abstractor	64.2M	81.4	74.3	68.1	62.8	54.1	48.4
Q-Former	112.2M	84.1	73.9	67.5	62.5	53.8	48.1
Honeybee	∼50M	84.5	74.8	68.7	63.2	54.5	48.7
NS-VQA (Oracle)	N/A	99.8	99.0	–	–	–	–
Ours (LLaVA-1.5-7B)							
SRP (Ours)	**37.8M**	**87.8**	**78.4**	**73.4**	**64.8**	**56.1**	**51.5**
Cross-architecture (Qwen-VL-7B)							
Qwen-MLP-2L	12.6M	83.4	74.5	67.8	62.4	53.6	47.4
Qwen-MoE	37.8M	84.1	75.6	68.5	63.4	54.1	48.5
Qwen-SRP	**37.8M**	**88.6**	**79.8**	**74.5**	**65.7**	**57.2**	**52.8**

Table 2: Ablation study. All variants use 37.8M parameters (LLaVA-1.5-7B backbone).

Variant	CoGenT-A	CoGenT-B	CL-avg	GQA	GQA-comp	GQA-OOD
SRP (Full)	**87.8**	**78.4**	**73.4**	**64.8**	**56.1**	**51.5**
w/o $\mathcal{L}_{\text{align}}$	85.4	74.8	68.6	62.5	53.6	48.2
w/o $\mathcal{L}_{\text{orth}}$	86.5	76.8	71.3	63.8	54.8	49.6
w/o $\mathcal{L}_{\text{cnt}}$	86.8	77.1	71.6	64.2	55.4	50.1
Static subspaces	86.2	75.6	69.8	63.4	54.2	49.2
Hard routing	86.6	76.5	70.5	63.9	54.7	49.8
Shared head	84.8	74.2	68.2	62.1	53.8	47.5
w/o mixed pooling	87.1	77.4	72.1	64.5	55.6	50.4
Unfiltered vocab	85.8	75.4	69.3	62.8	54.4	48.8

4.2 Fine-Grained Question-Type and Counting Analysis

Table 3 disaggregates performance by question type. Gains are *asymmetrically concentrated* on multi-role queries: SRP's advantage over MoE is $+5.6$ pp on CLEVR-Compare and $+5.4$ pp on CLEVR-Spatial, and reaches $+6.3$ pp on TallyQA-Complex, which requires resolving spatial coordinates and integer comparisons simultaneously. Notably, CLEVR-Spatial improves by $+5.4$ pp over MoE, consistent with the relation head receiving the strongest alignment signal from the dedicated spatial-preposition subspace. This pattern corroborates the attention logit gap analysis (Proposition 1): the per-head improvement compounds precisely on queries requiring multi-role composition. Gains on both synthetic (CLEVR) and naturalistic (TallyQA) counting indicate that the count head captures cardinality structure rather than dataset-specific shortcuts.

Removing $\mathcal{L}_{\text{cnt}}$ triggers a -3.7 pp drop on CLEVR-Count, confirming that holistic contrastive alignment alone fails to isolate cardinality. Reverting to foreground-only pooling (w/o mixed pooling) degrades TallyQA-Complex by -3.9 pp, as the train–test distribution mismatch resurfaces. The two ablations address complementary failure modes, with each degradation isolated to the counting metrics it is designed to protect. Exposing the count head to background patches during training via mixed pooling successfully suppresses this mismatch (Section 3.3.3).

Table 3: Fine-grained question-type breakdown and counting analysis.

Method	CLEVR Count	CLEVR Compare	CLEVR Spatial	TallyQA Simple	TallyQA Complex
MLP-2L	66.8	71.2	69.8	71.2	54.3
Vanilla MoE	67.5	72.0	71.4	72.5	55.8
SRP (Ours)	**72.1**	**77.6**	**76.8**	**75.4**	**62.1**
w/o $\mathcal{L}_{\text{cnt}}$	68.4	75.1	76.2	73.1	56.5
w/o mixed pooling	70.1	75.2	76.4	74.5	58.2

Table 4: Diagnostic probing, non-compositional control (VQAv2), and robustness evaluations. "GQA 20% Noise" reports accuracy under 20% bounding-box corruption (cf. clean GQA in Table 1).

Method	Probe Acc (%)	VQAv2 Acc (%)	GQA 20% Noise (%)	Zero-Shot GQA (%)
MLP-2L	58.1	81.2	55.8	48.5
LoRA-only	55.4	79.8	54.1	45.2
Vanilla MoE	65.2	81.0	58.2	50.4
SRP (task loss only)	66.8	81.3	59.1	51.1
SRP (Full)	**82.4**	80.9	**63.5**	**54.1**

4.3 Diagnostic Probing, Control, and Robustness

Table 4 validates three critical properties. **(i) Latent reorganization.** Linear probe accuracy jumps from 58.1% (MLP-2L) to 82.4% (SRP Full), confirming that the geometric losses enforce strict role-subspace boundaries in the embedding space. The Vanilla MoE achieves an intermediate 65.2%, showing that multi-head capacity alone provides partial but insufficient separation: the $+17.2$ pp gap between MoE and SRP Full is attributable entirely to the alignment and orthogonality losses. The SRP (task loss only) variant, trained without auxiliary losses, achieves only 66.8%, demonstrating that the multi-head architecture alone is insufficient; the geometric losses $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{orth}}$ are essential for disentanglement. We additionally observe that SRP Full's router entropy stabilizes at $H \approx 0.82$ nats (vs. 1.39 nats for uniform), indicating learned specialization without collapse to hard routing.

(ii) Non-compositional safety. SRP matches MLP-2L on VQAv2 (80.9% vs. 81.2%), so role-structured projections do not degrade holistic scene understanding in general-purpose pipelines.

(iii) Robustness. Under 20% bounding box noise corruption, SRP degrades by only -1.3 pp (64.8% $\rightarrow$ 63.5% on GQA), while MLP-2L drops -5.5 pp (61.3% $\rightarrow$ 55.8%). This $\sim 4\times$ robustness advantage arises because the soft router dynamically redistributes role weights when patch-level annotations are noisy, effectively averaging out localization errors across the gating distribution (full noise curve in Appendix H). Zero-shot transfer to unseen GQA distributions yields $+5.6$ pp over MLP-2L, confirming domain-agnostic generalization of the learned role structure.

4.4 Efficiency, Scalability, and Convergence

Table 5 demonstrates that SRP's structured decomposition imposes minimal computational overhead. SRP adds only 3.4% wall-clock training overhead and 4.0% inference throughput reduction relative to MLP-2L, while achieving substantially higher accuracy. Peak GPU memory remains within 1.0 GB of the capacity-matched Vanilla MoE, confirming negligible memory impact from the auxiliary loss computation. In contrast, Q-Former incurs 37% higher training time and 24% lower inference throughput with lower compositional accuracy. The key efficiency insight is that SRP shifts the computational burden to *pre-computed* textual subspace extraction ($\sim$45 min, amortized once), converting an expensive per-forward-pass cost into a one-time initialization.

The auxiliary losses converge monotonically, and hyperparameters tuned on CLEVR transfer to GQA without recalibration, indicating that the alignment geometry reflects linguistic structure rather than dataset-specific artifacts.

Table 5: Computational efficiency comparison. Total Params: projector + LoRA. Train: seconds per iteration. Infer: images per second. Peak Mem: GPU memory in GB.

Method	Total Params	Train (s/iter)	Infer (img/s)	Peak Mem (GB)
MLP-2L	12.6M	1.18	15.1	38.4
Vanilla MoE	37.8M	1.20	14.8	39.1
C-Abstractor	64.2M	1.45	12.8	41.2
Q-Former	112.2M	1.62	11.4	45.2
SRP (Ours)	**37.8M**	**1.22**	**14.5**	**39.4**

5 Conclusion

We identified semantic role flattening, the collapse of role-specific structure during visual-to-language projection, as a bottleneck for compositional reasoning in multimodal LLMs: an LLM's own text embeddings separate object, attribute, count, and relation structure, while standard MLP projectors discard it. The **Semantic Role Projector** restores that structure by decomposing visual projection into four typed heads aligned to the LLM's frozen text-embedding subspaces. We proved that this geometry widens the attention logit gap for role-selective heads under explicit spectral and alignment conditions, and we showed empirically that SRP improves compositional generalization on synthetic and naturalistic benchmarks and transfers to a second backbone, at 3.4% wall-clock training overhead and without degrading holistic performance.

Limitations

SRP needs structured scene annotations during training to build the role supervision masks; on domains without them it reduces to a gated multi-head projector, and we claim no geometric benefit there. Its gains are also concentrated on compositional queries: on holistic benchmarks such as VQAv2 it matches rather than exceeds the MLP baseline (Table 4). The interaction-region heuristic for the relation head loses precision in dense scenes, where gains diminish beyond twelve annotated objects (Appendix B.4), and the role set is fixed to four roles, so finer-grained or language-specific inventories remain untested. Finally, we evaluate only 7B backbones and English benchmarks, and the oracle neuro-symbolic baseline remains far above every learned projector, indicating that role-structured projection narrows but does not close the compositional gap.

Reproducibility statement

We provide a complete specification of our results. The derivation of the role subspaces, the supervised losses, and the training algorithm are given in Section 3 and Algorithm 1; formal assumptions and the full proof of Proposition 1 appear in Appendix C. Dataset splits, baseline specifications, evaluation metrics, optimization hyperparameters, random seeds, and software versions are documented in Appendix D, and the analyses validating the subspace and supervision choices are in Appendix E. All experiments use publicly available datasets and pretrained backbones.

References

Manoj Acharya, Kushal Kafle, and Christopher Kanan. TallyQA: Answering complex counting questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8076–8084, 2019.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39–48, 2016.

Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron Courville. CLOSURE: Assessing systematic generalization of CLEVR models. *arXiv preprint arXiv:1912.05783*, 2020.

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.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Shuai Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal LLM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision*, pp. 104–120. Springer, 2020.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? An analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP*, pp. 276–286. Association for Computational Linguistics, 2019.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2126–2136. Association for Computational Linguistics, 2018.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4129–4138. Association for Computational Linguistics, 2019.

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.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning*, pp. 19730–19742. PMLR, 2023.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Jinfa Zhang, Munan Ning, and Yuan Li. MoE-LLaVA: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2024b.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the International Conference on Machine Learning*, pp. 4114–4124. PMLR, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *International Conference on Learning Representations*, 2019.

George A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11): 39–41, 1995.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. In *International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, 2020.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Mazyarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601. Association for Computational Linguistics, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808. Association for Computational Linguistics, 2019.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. Neural-symbolic VQA: Disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

A Algorithm

Inference. At test time, scene-graph annotations are *not required*. The router and heads operate unconditionally; all auxiliary losses are not computed. The count head pools over all N_v patches. SRP adds no annotation burden at deployment.

Evaluation protocol. We evaluate on CLEVR-CoGenT (Condition A $\rightarrow$ B transfer), CLOSURE (seven systematic generalization splits), and GQA compositional splits. All experiments report exact-match accuracy.