CSM-VL: CONDITIONAL SUBSPACE MATRYOSHKA FOR VISION-LANGUAGE EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Matryoshka Representation Learning (MRL) enables a single encoder to support multiple dimensionality budgets through nested embedding prefixes. However, conventional MRL uses the same dimensional ordering for every input, limiting its ability to adapt to varying information needs in vision-language embeddings. We introduce **CSM-VL (Conditional Subspace Matryoshka for Vision-Language Embeddings)**, which replaces the fixed hierarchy with an input-dependent ordering of learned functional subspaces while preserving nested representations. **CSM-VL** partitions the embedding space into functional dimension groups and uses a conditional router to determine their order for each multimodal input. We further introduce **Utility-Guided Group Allocation**, which measures each group's contribution to the multimodal contrastive objective, and **Conditional Multimodal Interaction**, which encourages subspaces to capture informative cross-modal interactions. This allows different inputs to prioritize different subspaces under the same dimensional budget. Experiments across multimodal embeddings benchmarks and multiple vision-language backbones show that **CSM-VL** consistently improves the accuracy-dimensionality trade-off, particularly under constrained embedding budgets.

1 INTRODUCTION

Vision-language models (VLMs) have emerged as powerful models for understanding and reasoning over visual and linguistic information. Recent VLM2Vec-style approaches (Jiang et al., 2025) adapt VLMs into general-purpose multimodal embedding models for retrieval, classification, and visual question answering. However, their high-dimensional embeddings incur substantial memory, bandwidth, and similarity-computation costs, motivating compact representations that preserve useful information under reduced dimensionality. Matryoshka Representation Learning (MRL) (Kusupati et al., 2022) addresses this challenge by training a single representation whose prefixes form nested sub-embeddings, enabling flexible quality–efficiency trade-offs at inference time. ESE (Li et al., 2025) and MIPIC (Huy et al., 2026) further improve cross-dimensional information preservation, while M^3 (Cai et al., 2025) extends Matryoshka learning to multimodal settings. However, these methods retain a globally fixed dimensional hierarchy, assigning the same coordinates to each budget for every input. This is restrictive for multimodal embeddings, where different image–text pairs may emphasize visual content, linguistic semantics, or visual–linguistic interactions. Under a limited budget, a fixed prefix cannot adapt to these input-specific requirements, so dimensions useful for one input may be less informative for another. Thus, the challenge is not only to compress the embedding, but also to determine which functional dimensions should receive the limited budget for each multimodal input.

We propose **CSM-VL (Conditional Subspace Matryoshka for Vision-Language Embeddings)**, which replaces the fixed coordinate hierarchy with an input-dependent hierarchy over learned functional subspaces. **CSM-VL** groups dimensions according to their representation behavior and uses a conditional router to determine their ordering for each multimodal input. Larger representations retain previously selected groups, preserving the Matryoshka structure while allowing different inputs to prioritize different subspaces. We introduce two complementary objectives: **Utility-Guided Group Allocation**, which orders groups according to their contribution to the multimodal contrastive objective, and **Conditional Multimodal Interaction**, which encourages informative visual–linguistic interactions while reducing redundancy across groups. Thus, **CSM-VL** turns Matryoshka

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.

learning from fixed dimensional ordering into conditional subspace allocation while retaining nested representations.

Our contributions are as follows:

- We introduce **CSM-VL**, a conditional Matryoshka framework for VLM2Vec-style embeddings that learns functional groups and input-dependent group orderings while preserving nested representations.
- We propose **Utility-Guided Group Allocation**, linking group ordering to each group's contribution to the multimodal contrastive objective.
- We introduce **Conditional Multimodal Interaction**, promoting informative visual–linguistic interactions while reducing redundancy among candidate groups.
- We demonstrate improved accuracy–dimensionality trade-offs across multimodal benchmarks and multiple VLM backbones.

2 Related Work

Vision-language representation learning has evolved from contrastive models such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) to VLM-based multimodal embedding models such as VLM2Vec (Jiang et al., 2025). These approaches learn shared representation spaces across visual and linguistic inputs, enabling transfer across heterogeneous tasks. However, the resulting high-dimensional embeddings motivate methods that provide flexible representation budgets without training separate models. Matryoshka Representation Learning (MRL) (Kusupati et al., 2022) learns nested representation prefixes that support multiple dimensionality budgets with a single encoder. ESE (Li et al., 2025) extends this idea across dimensionality and model depth, while MIPIC (Huy et al., 2026) improves information preservation across nested dimensions through self-distilled intra-relational alignment and progressive information chaining. These methods primarily learn a globally shared dimensional hierarchy. Matryoshka learning has also been applied to multimodal models, such as Matryoshka Multimodal Models (M^3) (Cai et al., 2025), which vary the number of visual tokens rather than the dimensionality of the final multimodal embedding. Prior adaptive multimodal methods dynamically select visual tokens, adjust visual granularity, or allocate computation according to input characteristics. **CSM-VL** applies the same principle at the level of the final multimodal embedding. Rather than adapting encoder computation or pruning visual tokens, **CSM-VL** learns functional dimension groups and conditions their ordering on the input, allowing different samples to prioritize different subspaces under the same global dimension budget while preserving the nested structure of Matryoshka representations. This provides a complementary form of adaptivity that operates directly on the embedding dimensional hierarchy.

3 Methodology

3.1 Overview and Problem Formulation

As shown in Figure 1, we propose **CSM-VL**. Standard Matryoshka Representation Learning (MRL) applies a fixed coordinate ordering across samples, whereas **CSM-VL** conditionally orders learned functional groups based on the multimodal input.

Let $x = (V, L)$ be a multimodal input and $\mathbf{z} = f_\theta(x) \in \mathbb{R}^D$ the output of a vision-language encoder. For any dimension index set $\mathcal{I} \subseteq \{1, \dots, D\}$, let $\mathbf{z}_\mathcal{I}$ denote the subvector formed by the indexed coordinates. In standard MRL (Kusupati et al., 2022), a budget-d representation is the fixed prefix $\mathbf{z}^{(d)} = \mathbf{z}_{1:d}$. **CSM-VL** partitions the D dimensions into M disjoint functional groups $\mathcal{K} = \{G_1, \dots, G_M\}$ of equal size $|G_m| = r$, with $D = Mr$. For each input x, a lightweight router scores the groups and induces a permutation π_x. The selected groups at step k are

$$\mathcal{S}_k(x) = \{\pi_x(1), \dots, \pi_x(k)\}, \tag{1}$$

with corresponding dimension set

$$\mathcal{I}_k(x) = \bigcup_{m \in \mathcal{S}_k(x)} G_m, \qquad \mathbf{z}^{(k)}(x) = \mathbf{z}_{\mathcal{I}_k(x)}. \tag{2}$$

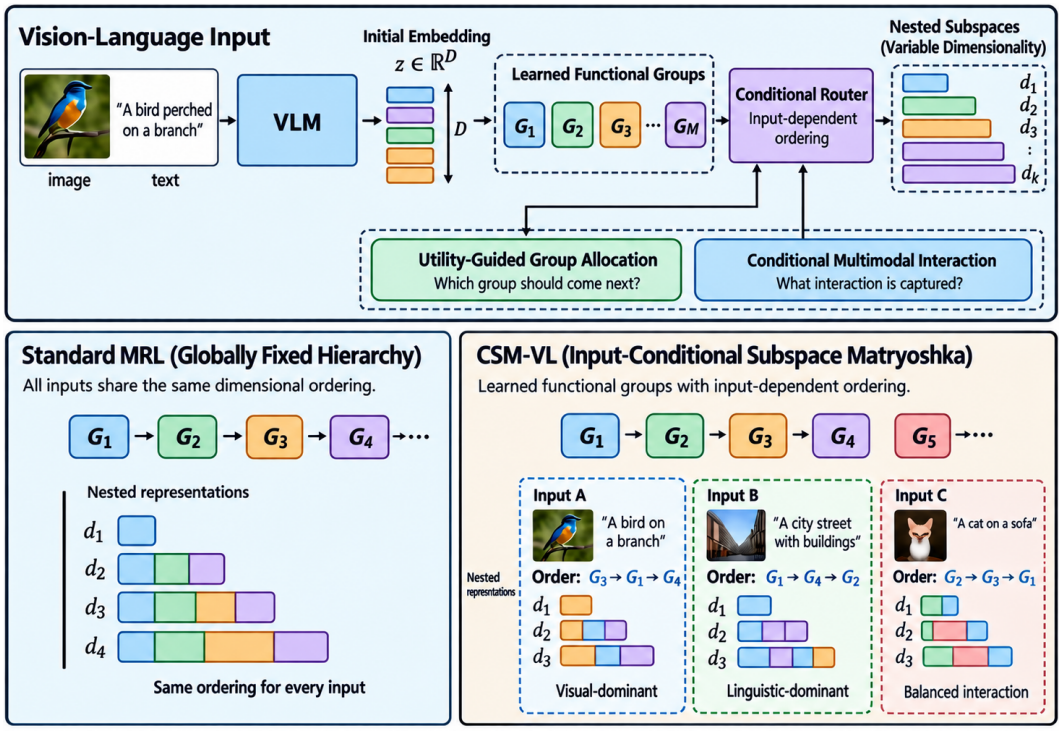

Figure 1: Overview of the **CSM-VL** framework. **Top:** a VLM extracts an embedding $\mathbf{z} \in \mathbb{R}^D$, partitioned into learned functional groups $\{G_1, \ldots, G_M\}$. A lightweight router predicts an input-dependent group ordering to form nested subspaces, optimized by *utility-guided group allocation* and *conditional multimodal interaction*. **Bottom-left:** standard MRL uses a fixed ordering across inputs. **Bottom-right: CSM-VL** adapts group ordering per sample while preserving nested representations.

We define $\mathcal{I}_0(x) = \varnothing$ and

$$\mathcal{R}_k(x) = \{1, \ldots, M\} \setminus \mathcal{S}_k(x). \tag{3}$$

The budget is $d_k = kr$, and $\mathcal{I}_k(x) \subseteq \mathcal{I}_{k+1}(x)$ by construction. For contrastive comparison at step k, the same query-selected set $\mathcal{I}_k(x_i)$ is applied to all target embeddings in the denominator, ensuring a common query-conditioned subspace.

3.2 Learning Functional Dimension Groups

We first learn a global partition of embedding dimensions into functional groups. For each dimension j, its functional signature is formed from activation statistics accumulated across training batches using an exponential moving average, avoiding dependence on individual mini-batch composition. Let ϕ_j denote the resulting stabilized signature.

An MLP h_ψ maps each signature to an embedding compared with learnable prototypes $\{c_m\}_{m=1}^M$:

$$A_{jm} = \frac{\exp\left(h_\psi(\phi_j)^\top c_m / \tau_g\right)}{\sum_{n=1}^M \exp\left(h_\psi(\phi_j)^\top c_n / \tau_g\right)}, \tag{4}$$

where τ_g is a temperature scalar. We convert these affinities into a balanced hard assignment $P \in \{0,1\}^{D \times M}$ satisfying

$$\sum_{m=1}^{M} P_{jm} = 1, \qquad \sum_{j=1}^{D} P_{jm} = r. \tag{5}$$

The functional groups are

$$G_m = \{j \in \{1, \ldots, D\} : P_{jm} = 1\}. \tag{6}$$

Thus, groups are disjoint and equal-sized, yielding predictable nested budgets. Group membership is fixed after the grouping stage; only their ordering is input-dependent during routing and inference.

3.3 Conditional Router

For each group G_m, the router uses activation statistics

$$\mathbf{s}_m(x) = \left[\mu_m(x), \sigma_m^2(x), \mathrm{RMS}_m(x)\right], \tag{7}$$

computed over its coordinates. A lightweight MLP h_ω produces

$$a_m(x) = h_\omega(\mathbf{s}_m(x)). \tag{8}$$

Sorting the scores induces π_x. At step k, the probability over remaining groups is

$$g_m^{(k)}(x) = \frac{\exp\left(a_m(x)/\tau_r\right)}{\sum_{n\in\mathcal{R}_k(x)} \exp\left(a_n(x)/\tau_r\right)}, \qquad m \in \mathcal{R}_k(x), \tag{9}$$

where τ_r is the routing temperature. This provides both an input-dependent ordering and the step-wise probabilities used by the routing objectives.

3.4 Objective Functions

After learning the functional groups, **CSM-VL** jointly optimizes the vision-language encoder and conditional router using nested contrastive learning, utility-guided routing, and conditional multi-modal interaction.

Contrastive Subspace Formulation. For a mini-batch $\mathcal{B}$ of size B, let $\mathbf{z}_i^q$ and $\mathbf{z}_i^{t^+}$ denote the full query and positive target embeddings. For $\mathcal{I} \subseteq \{1, \ldots, D\}$, the sample-level InfoNCE loss is

$$\ell_{\mathrm{InfoNCE}}(x_i; \mathcal{I}) = -\log \frac{\exp\left(\mathrm{sim}\left(\mathbf{z}_{i,\mathcal{I}}^q, \mathbf{z}_{i,\mathcal{I}}^{t^+}\right)/\tau\right)}{\sum_{j=1}^{B} \exp\left(\mathrm{sim}\left(\mathbf{z}_{i,\mathcal{I}}^q, \mathbf{z}_{j,\mathcal{I}}^{t}\right)/\tau\right)}, \tag{10}$$

where $\mathrm{sim}(\cdot,\cdot)$ is cosine similarity and τ is the temperature.

Base Multimodal Objective. The base loss optimizes all nested budgets:

$$\mathcal{L}_{\mathrm{base}} = \frac{1}{BM} \sum_{i=1}^{B} \sum_{k=1}^{M} \ell_{\mathrm{InfoNCE}}\left(x_i; \mathcal{I}_k(x_i)\right). \tag{11}$$

Utility-Guided Routing (UG). For an unselected group $m \in \mathcal{R}_k(x_i)$, define its marginal utility as

$$U_m^{(k)}(x_i) = \ell_{\mathrm{InfoNCE}}\left(x_i; \mathcal{I}_k(x_i)\right) - \ell_{\mathrm{InfoNCE}}\left(x_i; \mathcal{I}_k(x_i) \cup G_m\right). \tag{12}$$

Positive utility indicates a reduction in contrastive loss. We form the target distribution

$$q_m^{(k)}(x_i) = \frac{\exp\left(U_m^{(k)}(x_i)/\tau_u\right)}{\sum_{n\in\mathcal{R}_k(x_i)} \exp\left(U_n^{(k)}(x_i)/\tau_u\right)}, \tag{13}$$

where τ_u controls its sharpness. Using the router probabilities $g_m^{(k)}(x_i)$, the routing loss is

$$\mathcal{L}_{\mathrm{UG}} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{k=0}^{M-1} \sum_{m\in\mathcal{R}_k(x_i)} \mathrm{sg}\left(q_m^{(k)}(x_i)\right) \log g_m^{(k)}(x_i). \tag{14}$$

Conditional Multimodal Interaction (CMI). Utility alone may select groups with redundant cross-modal information. We therefore measure each group's interaction in a shared space. Let $\mathbf{v}_m(x_i)$, $\mathbf{l}_m(x_i)$, and $\mathbf{u}_m(x_i)$ denote its visual, linguistic, and joint modality summaries, obtained from the same multimodal forward pass. A shared projection W_C gives

$$\mathbf{c}_m(x_i) = W_C \left[\mathbf{u}_m(x_i) - \frac{1}{2} \left(\mathbf{v}_m(x_i) + \mathbf{l}_m(x_i) \right) \right]. \tag{15}$$

For selected groups, define the interaction subspace

$$C_k(x_i) = \operatorname{span} \left\{ \mathbf{c}_j(x_i) : j \in \mathcal{S}_k(x_i) \right\}. \tag{16}$$

For an unselected group $m \in \mathcal{R}_k(x_i)$, its residual interaction is

$$\mathbf{r}_m^{(k)}(x_i) = \mathbf{c}_m(x_i) - \operatorname{Proj}_{C_k(x_i)} \left(\mathbf{c}_m(x_i) \right), \tag{17}$$

with $C_0(x_i) = \{\mathbf{0}\}$. Pairwise residual redundancy is

$$\rho_{mn}^{(k)}(x_i) = \frac{\left\langle \mathbf{r}_m^{(k)}(x_i), \mathbf{r}_n^{(k)}(x_i) \right\rangle^2}{\left\| \mathbf{r}_m^{(k)}(x_i) \right\|^2 \left\| \mathbf{r}_n^{(k)}(x_i) \right\|^2 + \epsilon}, \tag{18}$$

where ϵ ensures numerical stability. The CMI loss penalizes high joint probability on redundant candidates:

$$\mathcal{L}_{\text{CMI}} = \frac{1}{B} \sum_{i=1}^{B} \sum_{k=0}^{M-1} \sum_{\substack{m,n \in \mathcal{R}_k(x_i) \\ m \neq n}} g_m^{(k)}(x_i) g_n^{(k)}(x_i) \rho_{mn}^{(k)}(x_i). \tag{19}$$

Overall Objective. We train **CSM-VL** with

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda_{\text{UG}} \mathcal{L}_{\text{UG}} + \lambda_{\text{CMI}} \mathcal{L}_{\text{CMI}}, \tag{20}$$

where the three terms respectively preserve multimodal quality, encourage utility-aware ordering, and reduce redundant cross-modal interactions. Here, $\lambda_{\text{UG}} = 0.45$ and $\lambda_{\text{CMI}} = 0.55$.

4 Theoretical Foundations

We provide theoretical foundations for three terms of **CSM-VL** that correspond directly to its training objectives. Complete proofs are provided in Appendix A.

4.1 Subspace Contrastive Learning and Mutual Information

Let X^q and X^t denote query and target inputs with joint distribution $p(x^q, x^t)$. For $\mathcal{I} \subseteq \{1, \ldots, D\}$, define $\mathbf{Z}_{\mathcal{I}}^q = f_\theta(X^q)_{\mathcal{I}}$ and $\mathbf{Z}_{\mathcal{I}}^t = f_\theta(X^t)_{\mathcal{I}}$.

Theorem 1 (Subspace Information Bound)**.** *Under the standard InfoNCE sampling setup,*

$$I(\mathbf{Z}_{\mathcal{I}}^q; \mathbf{Z}_{\mathcal{I}}^t) \geq \log B - \mathcal{L}_{\text{NCE}}(\mathcal{I}), \tag{21}$$

where

$$\mathcal{L}_{\text{NCE}}(\mathcal{I}) = \mathbb{E}_{(X^q, X^t)} \left[\ell_{\text{InfoNCE}}(X; \mathcal{I}) \right]. \tag{22}$$

Thus, minimizing the subspace InfoNCE term in **CSM-VL** maximizes a lower bound on the mutual information preserved within each selected subspace. See proofs in Appendix A.1.

4.2 Utility-Guided Routing and Greedy Allocation

Define the population information function

$$F(\mathcal{S}) = I \left(\mathbf{Z}_{\mathcal{I}(\mathcal{S})}^q; \mathbf{Z}_{\mathcal{I}(\mathcal{S})}^t \right), \qquad \mathcal{I}(\mathcal{S}) = \bigcup_{m \in \mathcal{S}} G_m, \tag{23}$$

with marginal gain

$$\Delta_m(\mathcal{S}) = F(\mathcal{S} \cup \{m\}) - F(\mathcal{S}). \tag{24}$$

The utility used by the **CSM-VL** router is

$$U_m^{(k)}(x_i) = \ell_{\text{InfoNCE}}\left(x_i; \mathcal{I}_k(x_i)\right) - \ell_{\text{InfoNCE}}\left(x_i; \mathcal{I}_k(x_i) \cup G_m\right), \tag{25}$$

which provides a sample-level proxy for the population marginal gain used to learn the input-dependent group ordering.

[Diminishing Information Returns] A$\subseteq \mathcal{B},\ m \notin \mathcal{B} \implies \Delta_m(\mathcal{A}) \geq \Delta_m(\mathcal{B})$.

Under this assumption, F is normalized, monotone, and submodular.

Theorem 2 (Greedy Approximation)**.** *Let $\mathcal{S}_k^*$ maximize F over subsets of k groups, and let $\mathcal{S}_k^{\text{greedy}}$ select the largest population marginal gain at each step. Then*

$$F\left(\mathcal{S}_k^{\text{greedy}}\right) \geq \left(1 - \left(1 - \frac{1}{k}\right)^k\right) F\left(\mathcal{S}_k^*\right) \geq \left(1 - \frac{1}{e}\right) F\left(\mathcal{S}_k^*\right). \tag{26}$$

This guarantee applies to ideal population-greedy allocation, while the learned **CSM-VL** router approximates this behavior using sample-level utility targets. See Appendix A.2.

4.3 Geometric Redundancy Reduction

The CMI objective in **CSM-VL** reduces redundancy between candidate groups in a shared interaction space. For selected groups $\mathcal{S}_k(x_i)$, define

$$\mathcal{C}_k(x_i) = \text{span}\left\{\mathbf{c}_j(x_i) : j \in \mathcal{S}_k(x_i)\right\}, \tag{27}$$

where $\mathbf{c}_m(x_i)$ is the projected interaction vector for group G_m. For an unselected group m,

$$\mathbf{r}_m^{(k)}(x_i) = \mathbf{c}_m(x_i) - \text{Proj}_{\mathcal{C}_k(x_i)}\left(\mathbf{c}_m(x_i)\right). \tag{28}$$

The residual similarity is

$$\rho_{mn}^{(k)}(x_i) = \frac{\left\langle \mathbf{r}_m^{(k)}(x_i), \mathbf{r}_n^{(k)}(x_i) \right\rangle^2}{\|\mathbf{r}_m^{(k)}(x_i)\|^2 \|\mathbf{r}_n^{(k)}(x_i)\|^2 + \epsilon}. \tag{29}$$

Proposition 3 (Residual Orthogonality)**.** *If $\left\langle \mathbf{r}_m^{(k)}(x_i), \mathbf{r}_n^{(k)}(x_i) \right\rangle = 0$, then $\rho_{mn}^{(k)}(x_i) = 0$.*

This is a geometric diversity result and does not imply statistical independence or zero conditional mutual information. Thus, the CMI term acts as a geometric regularizer that discourages highly aligned residual interactions among candidate groups.

Overall Interpretation. The three components of **CSM-VL** are complementary: subspace InfoNCE preserves multimodal information, utility-guided routing learns input-dependent group ordering from marginal gains, and CMI reduces geometric redundancy among candidate groups.

5 Experiments

5.1 Experimental Setup

Benchmark and Tasks. We evaluate **CSM-VL** on the Multimodal Embedding Benchmark (MMEB) (Jiang et al., 2025), covering Classification (CLS) and Visual Question Answering (VQA). Following the standard protocol, CLS includes ImageNet-1K, VOC2007, N24News, SUN397, and Hateful-Memes, while VQA includes OK-VQA, A-OKVQA, DocVQA, InfographicsVQA, ChartQA, and Visual7W. We report macro-averaged $P@1$ across datasets at each dimension, with mean $\pm$ standard deviation across independent runs. More experimental details are provided in Appendix B.

Backbones and baselines. We evaluate four backbones: FastVLM-0.5B (Vasu et al., 2025), B3-Qwen2-2B (Thirukovalluru et al., 2025), Qwen3-VL-2B (Bai et al., 2025), and Qwen3-VL-8B (Bai et al., 2025). The evaluated dimensions are $\{64, 128, 256, 512, 768, 896\}$ for FastVLM-0.5B (Vasu et al., 2025), $\{64, 128, 256, 512, 768, 896, 1536\}$ for B3-Qwen2-2B (Thirukovalluru et al., 2025), and $\{64, 128, 256, 512, 768, 1024, 2048\}$ for both Qwen3-VL-2B (Bai et al., 2025) and Qwen3-VL-8B (Bai et al., 2025). We compare **CSM-VL** on four backbones against MRL (Kusupati et al., 2022), ESE (Li et al., 2025), and MIPIC (Huy et al., 2026). For MIPIC (Huy et al., 2026), we adapt its intra-relational and progressive information-chaining objectives to the multimodal setting. All methods use the same training and evaluation protocol.

5.2 Main Results

Tables 1–4 show that **CSM-VL** generally improves over MRL (Kusupati et al., 2022) and MIPIC (Huy et al., 2026), particularly at constrained dimensions. On FastVLM-0.5B (Vasu et al., 2025), **CSM-VL** improves 64-dimensional Classification from 60.42% to 66.08% and VQA from 39.45% to 41.11% over MRL, reaching 66.82% and 51.90% at 256 dimensions. On B3-Qwen2-2B (Thirukovalluru et al., 2025), it reaches 77.40%/47.29% at 64 dimensions versus 72.92%/43.10% for MRL. On Qwen3-VL-2B (Bai et al., 2025), the corresponding results are 76.12%/55.39% versus 73.02%/53.37%, while on Qwen3-VL-8B (Bai et al., 2025), they are 77.11%/56.79% versus 74.62%/59.34%. **CSM-VL** also improves full-dimensional Classification and VQA performance. Since all groups are included at the full budget, these gains cannot be attributed to conditional ordering; instead, they reflect the effect of the proposed training objectives on the learned representation. Thus, the objectives improve the embedding overall, while conditional routing provides additional gains at compact budgets. Detailed results and computation overhead are reported in Appendices D and C.

Table 1: **Average accuracy of FastVLM-0.5B on embedding dimensions.** Results are reported for Classification and Visual Question Answering. Best performance is **bolded**.

Method	Embedding Dimension					
	64	128	256	512	768	896
Classification						
MRL	60.42±0.63	63.46±0.59	64.38±0.54	65.28±0.51	65.36±0.48	65.54±0.46
ESE	59.02±0.63	60.90±0.59	63.24±0.55	63.74±0.53	64.10±0.51	64.74±0.49
MIPIC	64.02±0.51	62.62±0.47	64.58±0.47	65.86±0.44	66.22±0.44	66.38±0.42
CSM-VL	**66.08±0.44**	**65.30±0.38**	**66.82±0.39**	**67.80±0.38**	**67.76±0.34**	**67.87±0.33**
Visual Question Answering						
MRL	39.45±0.78	46.27±0.73	50.12±0.68	50.88±0.63	51.63±0.60	52.03±0.57
ESE	38.45±0.83	42.77±0.78	44.68±0.73	46.17±0.68	46.93±0.65	47.05±0.61
MIPIC	39.94±0.73	46.54±0.67	50.27±0.61	51.17±0.57	51.64±0.54	51.91±0.51
CSM-VL	**41.11±0.62**	**48.62±0.56**	**51.90±0.56**	**52.61±0.53**	**53.05±0.51**	**53.30±0.50**

5.3 Ablation Study

Table 5 compares a base model without the auxiliary objectives, models using only $\mathcal{L}_{\text{UG}}$ or $\mathcal{L}_{\text{CMI}}$, and the full **CSM-VL** objective. Both objectives improve over the base model across dimensions, with $\mathcal{L}_{\text{UG}}$ providing the larger standalone gain. Their combination performs best at every budget, reaching 66.08% at 64 dimensions and 67.87% at full dimension, supporting their complementary roles. Further ablation results are provided in Appendix E.

5.4 Loss-Weight Sensitivity

Table 6 shows smooth performance variation across the nine weight configurations. The setting $(\lambda_{\text{UG}}, \lambda_{\text{CMI}}) = (0.45, 0.55)$ performs best across dimensions, while differences shrink at larger budgets. Additional results are provided in Appendix F.

Table 2: **Average accuracy of B3-Qwen2-2B on embedding dimensions.** Results are reported for Classification and Visual Question Answering. Best performance is **bolded**.

Method	Embedding Dimension						
	64	128	256	512	768	896	1536
Classification							
MRL	72.92±0.25	74.56±0.20	75.78±0.17	76.60±0.16	76.72±0.21	76.88±0.27	77.32±0.24
ESE	75.32±0.25	75.90±0.22	76.86±0.23	77.88±0.22	77.62±0.24	77.68±0.24	77.70±0.25
MIPIC	76.14±0.25	76.70±0.22	77.62±0.20	78.34±0.16	78.40±0.16	78.48±0.16	78.62±0.13
CSM-VL	**77.40±0.18**	**77.84±0.14**	**78.62±0.16**	**79.66±0.16**	**79.98±0.16**	**80.06±0.17**	**80.22±0.17**
Visual Question Answering							
MRL	43.10±0.33	48.95±0.36	51.58±0.33	53.40±0.28	54.18±0.24	54.20±0.28	54.78±0.27
ESE	45.38±0.39	49.55±0.37	51.97±0.35	53.05±0.31	53.72±0.27	53.55±0.27	53.80±0.27
MIPIC	45.86±0.37	49.85±0.30	52.49±0.29	54.30±0.26	54.69±0.26	54.83±0.25	55.25±0.24
CSM-VL	**47.29±0.26**	**51.71±0.26**	**53.62±0.27**	**55.40±0.25**	**55.52±0.23**	**55.68±0.23**	**56.04±0.23**

Table 3: **Average accuracy of Qwen3-VL-2B on embedding dimensions.** Results are reported for Classification and Visual Question Answering. Best performance is **bolded**.

Method	Embedding Dimension						
	64	128	256	512	768	1024	2048
Classification							
MRL	73.02±0.22	74.74±0.25	75.02±0.20	75.58±0.19	76.02±0.19	75.82±0.19	75.58±0.19
ESE	73.10±0.25	74.12±0.28	74.96±0.28	74.96±0.26	74.92±0.26	75.00±0.26	74.86±0.25
MIPIC	74.94±0.20	75.28±0.17	76.20±0.15	76.66±0.14	76.90±0.14	76.84±0.14	76.80±0.13
CSM-VL	**76.12±0.17**	**76.62±0.14**	**77.22±0.12**	**77.58±0.11**	**77.90±0.11**	**77.86±0.10**	**77.92±0.10**
Visual Question Answering							
MRL	53.37±0.27	57.43±0.23	60.18±0.22	61.42±0.20	61.25±0.20	61.38±0.19	61.68±0.18
ESE	52.92±0.30	56.15±0.26	58.32±0.29	58.88±0.28	59.45±0.28	59.22±0.28	59.45±0.27
MIPIC	54.06±0.26	57.52±0.22	61.00±0.20	62.41±0.19	62.90±0.18	63.06±0.18	63.30±0.17
CSM-VL	**55.39±0.22**	**59.57±0.19**	**62.71±0.18**	**63.93±0.16**	**64.20±0.15**	**64.30±0.15**	**64.50±0.14**

5.5 Conditional Allocation Analysis

CSM-VL conditionally orders functional groups according to the input. For partial budgets, we measure the Jaccard overlap (Jaccard, 1901) between selected group sets as $J(S_i, S_j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}$, and define diversity as $1 - J(S_i, S_j)$. Lower overlap indicates greater input dependence. Table 7 shows decreasing overlap and increasing diversity for FastVLM-0.5B (Vasu et al., 2025) and Qwen3-VL-2B (Bai et al., 2025). We restrict this metric to partial budgets because the full budget selects all groups, yielding Jaccard overlap exactly 1. We provide extra experiments with other backbones in Appendix G.

6 Conclusion

We presented **CSM-VL**, a conditional subspace Matryoshka framework for compact vision-language embeddings. **CSM-VL** replaces the fixed dimension ordering of conventional MRL with input-dependent functional group allocation, guided by group utility and conditional multimodal interactions. Across four VLM backbones and both Classification and VQA, **CSM-VL** consistently improves performance at compact embedding budgets, with the largest gains at lower dimensions. The method introduces additional training-time computation for routing and auxiliary objectives, while its gains become smaller as more dimensions are retained. These trade-offs highlight the value of adaptive subspace allocation while also motivating future work on more efficient routing and deployment strategies for large-scale multimodal retrieval.

Table 4: **Average accuracy of Qwen3-VL-8B on embedding dimensions.** Results are reported for Classification and Visual Question Answering. Best performance is **bolded**.

Method	Embedding Dimension						
	64	128	256	512	768	1024	2048
Classification							
MRL	74.62±0.50	75.94±0.46	76.80±0.42	77.38±0.38	77.70±0.36	77.88±0.34	78.00±0.33
ESE	74.67±0.44	75.59±0.44	76.51±0.40	77.11±0.36	77.44±0.34	77.64±0.32	77.78±0.31
MIPIC	76.65±0.47	76.50±0.39	77.28±0.33	77.92±0.30	78.22±0.28	78.38±0.26	78.98±0.25
CSM-VL	**77.11±0.37**	**77.00±0.34**	**78.02±0.30**	**78.64±0.27**	**79.66±0.26**	**79.54±0.25**	**80.28±0.24**
Visual Question Answering							
MRL	59.34±0.69	63.02±0.65	65.65±0.59	67.29±0.55	68.03±0.52	68.35±0.49	68.59±0.48
ESE	59.34±0.69	63.02±0.65	65.65±0.59	67.29±0.55	68.03±0.52	68.35±0.49	68.59±0.48
MIPIC	55.65±0.65	59.12±0.60	61.39±0.57	63.65±0.52	64.39±0.48	64.75±0.46	65.01±0.45
CSM-VL	**56.79±0.58**	**60.34±0.53**	**62.65±0.48**	**64.81±0.45**	**65.65±0.42**	**66.01±0.40**	**66.29±0.38**

Table 5: **Ablation of CSM-VL objectives for FastVLM-0.5B on Classification.** Results are reported across embedding dimensions. Best performance at each dimension is **bolded**.

Method	Embedding Dimension					
	64	128	256	512	768	896
Base	60.84±0.54	63.58±0.50	65.02±0.47	65.58±0.45	65.66±0.43	65.78±0.42
+$\mathcal{L}_{\mathrm{UG}}$	62.18±0.51	64.82±0.47	66.28±0.44	66.92±0.41	67.08±0.39	67.20±0.38
+$\mathcal{L}_{\mathrm{CMI}}$	61.62±0.52	64.34±0.48	65.82±0.45	66.42±0.42	66.56±0.40	66.68±0.39
CSM-VL	**66.08±0.44**	**65.30±0.38**	**66.82±0.39**	**67.80±0.38**	**67.76±0.34**	**67.87±0.33**

Table 6: **Loss-weight sensitivity on Visual Question Answering with Qwen3-VL-2B.** Results are reported across embedding dimensions. Best performance per dimension is **bolded**.

$(\lambda_{\mathrm{UG}}, \lambda_{\mathrm{CMI}})$	Embedding Dimension						
	64	128	256	512	768	1024	2048
(0.30, 0.70)	55.15±0.27	59.06±0.25	61.62±0.23	62.72±0.22	63.06±0.21	63.26±0.21	63.46±0.20
(0.35, 0.65)	55.30±0.26	59.26±0.24	61.82±0.22	62.92±0.21	63.26±0.20	63.46±0.20	63.66±0.19
(0.40, 0.60)	55.18±0.25	59.46±0.23	62.02±0.21	63.12±0.20	63.46±0.19	63.66±0.19	63.86±0.18
(0.45, 0.55)	**55.39±0.22**	**59.57±0.19**	**62.71±0.18**	**63.93±0.16**	**64.20±0.15**	**64.30±0.15**	**64.50±0.14**
(0.50, 0.50)	55.22±0.25	59.50±0.23	62.06±0.21	63.16±0.20	63.50±0.19	63.70±0.19	63.90±0.18
(0.55, 0.45)	55.16±0.26	59.34±0.24	61.90±0.22	63.00±0.21	63.34±0.20	63.54±0.20	63.74±0.19
(0.60, 0.40)	55.26±0.27	59.14±0.25	61.70±0.23	62.80±0.22	63.14±0.21	63.34±0.21	63.54±0.20
(0.65, 0.35)	55.02±0.28	58.90±0.26	61.46±0.24	62.56±0.23	62.90±0.22	63.10±0.22	63.30±0.21
(0.70, 0.30)	54.74±0.29	58.62±0.27	61.18±0.25	62.28±0.24	62.62±0.23	62.82±0.23	63.02±0.22

Table 7: **Conditional allocation diversity at partial embedding dimensions.** Group overlap measures allocation similarity across inputs, while allocation diversity is its complement. Lower overlap and higher diversity indicate more input-dependent allocation.

	FastVLM-0.5B		**Qwen3-VL-2B**	
Dim.	Overlap ↓	Diversity ↑	Overlap ↓	Diversity ↑
64	0.71	0.29	0.68	0.32
128	0.64	0.36	0.61	0.39
256	0.57	0.43	0.54	0.46
512	0.51	0.49	0.48	0.52
768	0.48	0.52	0.45	0.55
1024	–	–	0.42	0.58

7 AI USE STATEMENT

In this work, we used generative AI tools for proofreading, grammar correction, and improving the clarity and readability of the manuscript text. We have not used generative AI tools for research ideation, problem formulation, methodological design, data collection, data analysis, experimental design, implementation, or generation of experimental results, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for no other tasks requiring disclosure. We have reviewed all AI-assisted text and verified that the final manuscript accurately reflects the authors' work and conclusions. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

8 REPRODUCIBILITY STATEMENT

We have made our best effort to ensure that the experiments are reproducible. We provide the implementation of **CSM-VL**, training and evaluation scripts, configuration files, and instructions for reproducing the reported experiments in the supplementary material. The accompanying source code will be included as an anonymous supplementary ZIP file for review and made publicly available after publication. The main paper and appendix describe the model formulation, training objectives, experimental settings, evaluation protocol, and additional results, while the supplementary materials provide the details needed to reproduce the experiments and compare against the evaluated baselines.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Li, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Phung Gia Huy, Hai An Vu, Minh-Phuc Truong, Thang Duc Tran, Linh Ngo Van, Thanh Hong Nguyen, and Trung Le. Mipic: Matryoshka representation learning via self-distilled intra-relational and progressive information chaining. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13824–13838, 2026.

Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 4904–4916, 2021.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. In *The Thirteenth International Conference on Learning Representations*, 2025.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30233–30249, 2022.

Xianming Li, Zongxi Li, Jing Li, Haoran Xie, and Qing Li. Ese: Espresso sentence embeddings. In *The Thirteenth International Conference on Learning Representations*, 2025.

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 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Raghuveer Thirukovalluru, Rui Meng, Ye Liu, K. Karthikeyan, Mingyi Su, Ping Nie, Semih Yavuz, Yingbo Zhou, Wenhu Chen, and Bhuwan Dhingra. Breaking the batch barrier (b3) of contrastive learning via smart batch mining. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://openreview.net/forum?id=M8zmlixh9y.

Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokula Santhanam, James Gabriel, Peter Grasch, Oncel Tuzel, and Hadi Pouransari. Fastvlm: Efficient vision encoding for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19769–19780, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Vasu_FastVLM_Efficient_Vision_Encoding_for_Vision_Language_Models_CVPR_2025_paper.pdf.

A Theoretical Proofs and Derivations

A.1 Proof of Theorem 1 (Subspace Information Bound)

Proof. Let $\mathcal{I} \subseteq \{1, \dots, D\}$ and define the restricted random variables $\mathbf{Z}^q_{\mathcal{I}}$ and $\mathbf{Z}^t_{\mathcal{I}}$ as in Theorem 1. Under the standard InfoNCE sampling scheme, positive pairs are drawn from the joint distribution $p(\mathbf{z}^q_{\mathcal{I}}, \mathbf{z}^t_{\mathcal{I}})$, while negative targets are sampled independently from the marginal $p(\mathbf{z}^t_{\mathcal{I}})$. Using the density-ratio critic

$$h^*(\mathbf{z}^q_{\mathcal{I}}, \mathbf{z}^t_{\mathcal{I}}) = \log \frac{p(\mathbf{z}^q_{\mathcal{I}}, \mathbf{z}^t_{\mathcal{I}})}{p(\mathbf{z}^q_{\mathcal{I}})p(\mathbf{z}^t_{\mathcal{I}})},$$

the standard InfoNCE mutual-information bound gives

$$I(\mathbf{Z}^q_{\mathcal{I}}; \mathbf{Z}^t_{\mathcal{I}}) \geq \log B - \mathcal{L}_{\mathrm{NCE}}(\mathcal{I}), \tag{30}$$

where

$$\mathcal{L}_{\mathrm{NCE}}(\mathcal{I}) = \mathbb{E}_{(X^q, X^t)} \left[\ell_{\mathrm{InfoNCE}}(X; \mathcal{I}) \right]. \tag{31}$$

For completeness, the bound follows by taking the expectation over the negative samples and applying Jensen's inequality to the logarithm of the negative-sample denominator, followed by the density-ratio expression above. The resulting expected contrastive objective is therefore an upper bound on $\log B - I(\mathbf{Z}^q_{\mathcal{I}}; \mathbf{Z}^t_{\mathcal{I}})$. Rearranging yields the claimed inequality. The mini-batch average used in training is an empirical estimator of $\mathcal{L}_{\mathrm{NCE}}(\mathcal{I})$. □

A.2 Proof of Theorem 2 (Greedy Approximation)

Proof. Let $F(\mathcal{S})$ be the population information function defined in Theorem 2. Under Assumption 4.2, F is normalized, monotone, and submodular.

Let $\mathcal{S}^*_k$ be an optimal subset of k groups and let $\mathcal{S}_j$ denote the subset produced after j greedy selections, with $\mathcal{S}_0 = \varnothing$. By submodularity and monotonicity,

$$F(\mathcal{S}^*_k) \leq F(\mathcal{S}_j) + \sum_{m \in \mathcal{S}^*_k \setminus \mathcal{S}_j} \left[F(\mathcal{S}_j \cup \{m\}) - F(\mathcal{S}_j) \right]. \tag{32}$$

There are at most k terms in the sum. Since the greedy rule chooses the remaining group with the largest marginal gain,

$$F(\mathcal{S}^*_k) \leq F(\mathcal{S}_j) + k \left[F(\mathcal{S}_{j+1}) - F(\mathcal{S}_j) \right]. \tag{33}$$

Rearranging gives

$$F(\mathcal{S}^*_k) - F(\mathcal{S}_{j+1}) \leq \left(1 - \frac{1}{k}\right) \left[F(\mathcal{S}^*_k) - F(\mathcal{S}_j) \right]. \tag{34}$$

Applying this recurrence for $j = 0, \dots, k-1$ and using $F(\mathcal{S}_0) = F(\varnothing) = 0$ yields

$$F(\mathcal{S}^*_k) - F(\mathcal{S}_k) \leq \left(1 - \frac{1}{k}\right)^k F(\mathcal{S}^*_k). \tag{35}$$

Therefore,

$$F(\mathcal{S}_k) \geq \left[1 - \left(1 - \frac{1}{k}\right)^k\right] F(\mathcal{S}^*_k) \geq \left(1 - \frac{1}{e}\right) F(\mathcal{S}^*_k). \tag{36}$$

Thus, the ideal population-greedy allocation admits the standard $1 - 1/e$ approximation guarantee. This guarantee applies to the ideal population marginal gains, while **CSM-VL** approximates them using sample-level utility targets. □

A.3 Proof of Proposition 3 (Residual Orthogonality)

Proof. For fixed input x_i and routing step k, let $\mathcal{C}_k(x_i)$ denote the shared interaction subspace spanned by the selected interaction vectors. By definition,

$$\mathbf{r}^{(k)}_m(x_i) = \mathbf{c}_m(x_i) - \mathrm{Proj}_{\mathcal{C}_k(x_i)} \left(\mathbf{c}_m(x_i) \right). \tag{37}$$

Thus, each residual is the component of the candidate interaction vector orthogonal to the current selected interaction subspace.

For two candidate groups m and n, the redundancy penalty is

$$\rho_{mn}^{(k)}(x_i) = \frac{\left\langle \mathbf{r}_m^{(k)}(x_i), \mathbf{r}_n^{(k)}(x_i) \right\rangle^2}{\left\| \mathbf{r}_m^{(k)}(x_i) \right\|^2 \left\| \mathbf{r}_n^{(k)}(x_i) \right\|^2 + \epsilon}. \tag{38}$$

Since $\epsilon > 0$, the denominator is strictly positive. Hence

$$\rho_{mn}^{(k)}(x_i) = 0 \iff \left\langle \mathbf{r}_m^{(k)}(x_i), \mathbf{r}_n^{(k)}(x_i) \right\rangle = 0, \tag{39}$$

which is precisely orthogonality of the two residual interaction vectors in the shared interaction space.

Therefore, minimizing $\rho_{mn}^{(k)}$ suppresses geometric alignment between candidate residual interactions. This result is a deterministic geometric statement and does not imply statistical independence or zero conditional mutual information. □

B Experimental Details

B.1 Implementation Details

Models. We evaluate our approach on four vision-language model backbones spanning different model scales and architectures: **FastVLM-0.5B** (Vasu et al., 2025), **B3-Qwen2-2B** (Thirukovalluru et al., 2025), **Qwen3-VL-2B**, and **Qwen3-VL-8B** (Bai et al., 2025). The Qwen3-VL family provides dense models at multiple parameter scales, allowing us to evaluate whether the proposed conditional subspace allocation generalizes across substantially different model capacities (Bai et al., 2025).

For all models, we obtain a fixed-dimensional representation by applying `eos` pooling to the final hidden states. The resulting embeddings are subsequently used for contrastive training and evaluation across different dimensional budgets. Detailed training configurations are reported in Table 8.

Training and Evaluation. We train and evaluate the models on two meta-task categories from the MMEB benchmark (Jiang et al., 2025): **Classification** and **Visual Question Answering**. Classification tasks evaluate whether the learned embeddings preserve discriminative information for selecting the correct class label, while VQA tasks evaluate whether the representation captures the semantic correspondence between an image-question query and its textual answer. We run training and evaluation on two H100 GPUs.

Unless otherwise specified, the same training datasets and evaluation protocol are used across model architectures. This enables a controlled comparison of representation quality across model scales and embedding dimensions.

Functional Groups and Conditional Router. For each backbone, we partition the full embedding dimension D into $M = 8$ disjoint groups of uniform size $r = D/8$. Thus, $D = Mr$, and retaining k groups corresponds to a budget of $d_k = kr$ dimensions. A lightweight MLP with hidden dimension 256 assigns a scalar score to each group. The scores are normalized with a softmax over the available groups to determine the input-dependent group ordering. No additional backbone or teacher model is introduced.

Table 8: **Training configurations for all evaluated VLM backbones.** The Qwen3-VL-8B configuration follows the same optimization settings as Qwen3-VL-2B unless otherwise specified.

Setting	**FastVLM-0.5B**	**B3-Qwen2-2B**	**Qwen3-VL-2B**	**Qwen3-VL-8B**
Epoch	1	1	1	1
Learning Rate	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}
Projector LR	5×10^{-4}	5×10^{-4}	5×10^{-4}	5×10^{-4}
Batch Size	32	16	16	16
LR Scheduler	Cosine	Cosine	Cosine	Cosine
Warmup Ratio	0.03	0.03	0.03	0.03
Weight Decay	0.01	0.01	0.01	0.01
LoRA Rank	64	64	64	64
LoRA Alpha	64	64	64	64
Image Resolution	448	336	336	336

Table 9: **Statistics of the training and evaluation datasets.** Query and target denote the input and retrieval target modalities, respectively.

Meta-Task	Dataset	Query	Target	#Training	#Eval	#Candidates
Classification	ImageNet-1K	I	T	100K	1000	1000
	N24News	I + T	I	49K	1000	24
	HatefulMemes	I	T	8K	1000	2
	VOC2007	I	T	8K	1000	20
	SUN397	I	T	20K	1000	397
VQA	OK-VQA	I + T	T	9K	1000	1000
	A-OKVQA	I + T	T	17K	1000	1000
	DocVQA	I + T	T	40K	1000	1000
	InfographicVQA	I + T	T	24K	1000	1000
	ChartQA	I + T	T	28K	1000	1000
	Visual7W	I + T	T	70K	1000	1000

B.2 Detailed Dataset Statistics

We follow the MMEB training and evaluation protocol (Jiang et al., 2025). The datasets used in our experiments are divided into two meta-task categories: Classification and Visual Question Answering. Their statistics are summarized in Table 9.

For Classification, each query consists primarily of an image together with an instruction and, for datasets containing multimodal inputs, associated textual information. The target is a textual class label. Evaluation is performed by ranking the candidate class labels according to their embedding similarity, with the number of candidates determined by the number of classes in each dataset.

For VQA, each query contains an image, a natural-language question, and the corresponding task instruction. The target is a textual answer. Following the MMEB protocol (Jiang et al., 2025), evaluation is formulated as retrieval over 1,000 answer candidates, containing one positive answer and 999 distractors. This formulation evaluates the quality of the learned multimodal representation independently of free-form generation.

Embedding Extraction. For each input, the VLM produces a sequence of hidden states, from which we use the final `EOS` representation as the full-dimensional embedding. **CSM-VL** subsequently partitions the embedding dimensions into learned groups and predicts an input-dependent ordering over these groups. For a given budget, the representation is obtained by retaining the corresponding prefix of the ordered groups, yielding nested subspaces with different dimensional budgets. Importantly, the grouping and ordering are determined during training and do not require an additional encoder or teacher model at inference time.

C Computational Overhead

CSM-VL introduces a conditional allocation module and two auxiliary objectives, resulting in additional computation during training. We therefore compare its computational cost with MRL (Kusupati et al., 2022), ESE (Li et al., 2025), and MIPIC (Huy et al., 2026) across four VLM backbones, including **FastVLM-0.5B** (Vasu et al., 2025), **B3-Qwen2-2B** (Thirukovalluru et al., 2025), **Qwen3-VL-2B** (Bai et al., 2025), and **Qwen3-VL-8B** (Bai et al., 2025). All methods are evaluated using the same training configuration and hardware for each backbone. As shown in Table 10, MRL provides the reference training cost, while ESE incurs substantially higher cost due to its additional depth-wise elasticity. MIPIC (Huy et al., 2026) and **CSM-VL** introduce additional operations for learning structured representations, but these operations are performed around the extracted VLM embeddings rather than requiring an additional backbone model. The computational overhead of **CSM-VL** is primarily incurred during training. At inference time, **CSM-VL** does not require an

Table 10: **Computational overhead across VLM backbones.** Training time is measured under the same experimental configuration.

Model	Method	Training Time (h)	Relative Cost
FastVLM-0.5B	MRL	4.1	$\times 1.00$
	ESE	7.4	$\times 1.80$
	MIPIC	4.5	$\times 1.10$
	CSM-VL	4.3	$\times 1.05$
B3-Qwen2-2B	MRL	7.8	$\times 1.00$
	ESE	13.4	$\times 1.72$
	MIPIC	9.1	$\times 1.17$
	CSM-VL	8.9	$\times 1.14$
Qwen3-VL-2B	MRL	8.5	$\times 1.00$
	ESE	14.9	$\times 1.75$
	MIPIC	9.9	$\times 1.16$
	CSM-VL	9.4	$\times 1.11$
Qwen3-VL-8B	MRL	16.8	$\times 1.00$
	ESE	29.2	$\times 1.74$
	MIPIC	19.5	$\times 1.16$
	CSM-VL	19.1	$\times 1.14$

additional VLM forward pass for each representation dimension. Once the VLM embedding and conditional allocation are computed, the nested representations can be constructed from the learned groups. Thus, the method separates the expensive VLM forward computation from the lightweight dimension-selection operation, which is particularly important when multiple representation budgets must be served from the same embedding.

D Detailed Main Results

The per-dataset results in this section provide a detailed view of CSM-VL across datasets, embedding dimensions, and VLM backbones. Across both Classification and Visual Question Answering, CSM-VL consistently improves over MIPIC across the evaluated dimensions, complementing the aggregate main results.

E More Ablation Experiments

We further evaluate the contribution of the two training objectives on the larger Qwen3-VL-8B (Bai et al., 2025) backbone using the Visual Question Answering task. Table 18 compares a base model without either auxiliary objective, models using only $\mathcal{L}_{\mathrm{UG}}$ or $\mathcal{L}_{\mathrm{CMI}}$, and the full **CSM-VL** objective across the complete nested representation hierarchy. Both individual objectives improve over the base model, while their combination consistently achieves the best performance across all embedding dimensions, supporting their complementary roles.

F More Loss-Weight Sensitivity experiments

We further study the sensitivity of **CSM-VL** to the relative weights of the two training objectives for B3-Qwen2-2B (Thirukovalluru et al., 2025) on Classification. We evaluate nine configurations of $(\lambda_{\mathrm{UG}}, \lambda_{\mathrm{CMI}})$ whose weights sum to one, covering values from 0.30 to 0.70. Table 19 shows that performance varies smoothly across the tested configurations, with no sharp degradation toward either objective. The setting $(0.45, 0.55)$ achieves the best performance across all embedding dimensions, suggesting that a slight emphasis on $\mathcal{L}_{\mathrm{CMI}}$ is beneficial for this backbone. The gains over nearby configurations are relatively small and become less pronounced at larger dimensions, indicating that the precise loss weighting is less critical as representation capacity increases.

Table 11: **Per-dataset classification accuracy of FastVLM-0.5B.** Results are reported across embedding dimensions. Best performance per dimension is **bolded**.

Dataset	Dim.	MRL	ESE	MIPIC	CSM-VL
ImageNet-1K	64	41.00 ± 0.55	41.60 ± 0.60	59.76 ± 0.49	**61.90 ± 0.41**
	128	48.70 ± 0.51	47.70 ± 0.55	61.82 ± 0.45	**64.52 ± 0.37**
	256	50.50 ± 0.46	49.70 ± 0.51	63.66 ± 0.42	**65.82 ± 0.34**
	512	52.40 ± 0.43	52.60 ± 0.47	64.50 ± 0.39	**66.36 ± 0.31**
	768	54.60 ± 0.40	53.80 ± 0.45	64.82 ± 0.37	**66.52 ± 0.29**
	896	55.20 ± 0.38	54.80 ± 0.42	64.98 ± 0.35	**66.64 ± 0.28**
VOC2007	64	77.80 ± 0.60	75.30 ± 0.65	61.82 ± 0.46	**64.40 ± 0.41**
	128	80.00 ± 0.56	75.60 ± 0.60	63.42 ± 0.42	**66.02 ± 0.34**
	256	77.80 ± 0.51	77.80 ± 0.56	64.06 ± 0.39	**66.62 ± 0.31**
	512	80.80 ± 0.48	76.80 ± 0.52	66.30 ± 0.36	**68.16 ± 0.28**
	768	77.50 ± 0.45	76.20 ± 0.50	66.42 ± 0.34	**68.52 ± 0.25**
	896	79.00 ± 0.43	77.50 ± 0.47	66.58 ± 0.32	**68.64 ± 0.24**
N24News	64	68.40 ± 0.64	67.30 ± 0.60	65.06 ± 0.51	**67.70 ± 0.43**
	128	68.20 ± 0.60	67.70 ± 0.57	62.62 ± 0.50	**65.32 ± 0.39**
	256	69.00 ± 0.55	68.60 ± 0.53	65.06 ± 0.51	**67.22 ± 0.43**
	512	69.40 ± 0.52	68.20 ± 0.57	65.90 ± 0.49	**68.16 ± 0.43**
	768	69.40 ± 0.49	68.70 ± 0.56	66.22 ± 0.49	**67.92 ± 0.39**
	896	69.00 ± 0.46	68.50 ± 0.57	66.38 ± 0.47	**68.02 ± 0.38**
SUN397	64	54.10 ± 0.67	53.20 ± 0.64	66.18 ± 0.53	**67.78 ± 0.45**
	128	61.00 ± 0.63	57.10 ± 0.60	62.62 ± 0.50	**65.32 ± 0.39**
	256	63.90 ± 0.58	62.70 ± 0.57	65.06 ± 0.51	**67.22 ± 0.43**
	512	64.40 ± 0.55	62.70 ± 0.52	66.30 ± 0.49	**68.16 ± 0.43**
	768	66.30 ± 0.52	63.00 ± 0.50	66.82 ± 0.49	**67.92 ± 0.39**
	896	66.70 ± 0.49	64.60 ± 0.50	66.98 ± 0.47	**68.02 ± 0.38**
Hateful-Memes	64	60.80 ± 0.71	57.70 ± 0.65	67.26 ± 0.55	**68.60 ± 0.49**
	128	59.40 ± 0.67	56.40 ± 0.62	62.62 ± 0.50	**65.32 ± 0.39**
	256	60.70 ± 0.62	57.40 ± 0.60	65.06 ± 0.51	**67.22 ± 0.43**
	512	59.40 ± 0.59	58.40 ± 0.55	66.30 ± 0.49	**68.16 ± 0.43**
	768	59.00 ± 0.56	58.80 ± 0.52	66.82 ± 0.49	**67.92 ± 0.39**
	896	57.80 ± 0.53	58.30 ± 0.50	66.98 ± 0.47	**68.02 ± 0.38**

Table 12: **Per-dataset visual question answering accuracy of FastVLM-0.5B.** Results are reported across embedding dimensions. Best performance per dimension is **bolded**.

Dataset	Dim.	MRL	ESE	MIPIC	CSM-VL
OK-VQA	64	37.50 ± 0.70	35.60 ± 0.76	38.62 ± 0.63	$\mathbf{39.86 \pm 0.53}$
	128	45.30 ± 0.65	39.30 ± 0.70	45.34 ± 0.59	$\mathbf{46.92 \pm 0.49}$
	256	49.80 ± 0.60	43.00 ± 0.67	49.24 ± 0.54	$\mathbf{50.76 \pm 0.50}$
	512	51.30 ± 0.55	44.80 ± 0.62	50.14 ± 0.50	$\mathbf{51.58 \pm 0.47}$
	768	52.00 ± 0.52	45.40 ± 0.59	50.72 ± 0.47	$\mathbf{52.06 \pm 0.45}$
	896	51.80 ± 0.49	45.50 ± 0.54	50.98 ± 0.46	$\mathbf{52.32 \pm 0.44}$
InfographicsVQA	64	24.80 ± 0.74	20.60 ± 0.80	38.62 ± 0.67	$\mathbf{39.86 \pm 0.57}$
	128	29.00 ± 0.69	25.20 ± 0.74	46.04 ± 0.63	$\mathbf{48.42 \pm 0.53}$
	256	34.60 ± 0.64	27.50 ± 0.71	50.24 ± 0.61	$\mathbf{52.38 \pm 0.57}$
	512	34.30 ± 0.59	27.90 ± 0.67	51.14 ± 0.57	$\mathbf{52.58 \pm 0.54}$
	768	34.70 ± 0.56	28.20 ± 0.64	51.02 ± 0.54	$\mathbf{52.86 \pm 0.52}$
	896	35.40 ± 0.53	28.00 ± 0.61	51.38 ± 0.52	$\mathbf{53.08 \pm 0.51}$
A-OKVQA	64	37.10 ± 0.77	36.20 ± 0.83	39.92 ± 0.69	$\mathbf{40.76 \pm 0.61}$
	128	44.10 ± 0.72	40.50 ± 0.77	47.24 ± 0.63	$\mathbf{48.22 \pm 0.53}$
	256	47.50 ± 0.67	41.50 ± 0.71	50.54 ± 0.61	$\mathbf{52.06 \pm 0.57}$
	512	49.10 ± 0.62	43.20 ± 0.66	51.44 ± 0.57	$\mathbf{52.88 \pm 0.54}$
	768	50.20 ± 0.59	45.10 ± 0.63	52.02 ± 0.54	$\mathbf{53.34 \pm 0.52}$
	896	50.70 ± 0.56	44.50 ± 0.60	52.28 ± 0.52	$\mathbf{53.60 \pm 0.51}$
ChartQA	64	40.40 ± 0.80	41.60 ± 0.82	40.62 ± 0.76	$\mathbf{41.86 \pm 0.64}$
	128	45.70 ± 0.75	43.80 ± 0.78	47.34 ± 0.70	$\mathbf{49.12 \pm 0.56}$
	256	48.50 ± 0.70	44.90 ± 0.72	50.54 ± 0.65	$\mathbf{52.06 \pm 0.57}$
	512	49.60 ± 0.65	46.20 ± 0.67	51.44 ± 0.60	$\mathbf{52.88 \pm 0.54}$
	768	49.30 ± 0.62	47.10 ± 0.64	52.02 ± 0.56	$\mathbf{53.34 \pm 0.52}$
	896	49.60 ± 0.59	46.90 ± 0.61	52.28 ± 0.52	$\mathbf{53.60 \pm 0.51}$
DocVQA	64	57.40 ± 0.82	55.30 ± 0.84	40.92 ± 0.79	$\mathbf{42.16 \pm 0.69}$
	128	68.10 ± 0.77	62.30 ± 0.80	46.64 ± 0.71	$\mathbf{49.52 \pm 0.61}$
	256	72.70 ± 0.72	65.90 ± 0.74	50.54 ± 0.65	$\mathbf{52.06 \pm 0.57}$
	512	73.60 ± 0.67	68.90 ± 0.69	51.44 ± 0.60	$\mathbf{52.88 \pm 0.54}$
	768	74.40 ± 0.64	69.00 ± 0.66	52.02 ± 0.56	$\mathbf{53.34 \pm 0.52}$
	896	74.90 ± 0.61	70.30 ± 0.63	52.28 ± 0.52	$\mathbf{53.60 \pm 0.51}$
Visual7W	64	39.50 ± 0.85	41.40 ± 0.91	40.92 ± 0.82	$\mathbf{42.16 \pm 0.69}$
	128	45.40 ± 0.80	45.50 ± 0.86	46.64 ± 0.74	$\mathbf{49.52 \pm 0.61}$
	256	47.60 ± 0.75	45.30 ± 0.81	50.54 ± 0.61	$\mathbf{52.06 \pm 0.57}$
	512	47.40 ± 0.70	46.00 ± 0.76	51.44 ± 0.57	$\mathbf{52.88 \pm 0.54}$
	768	49.20 ± 0.67	46.80 ± 0.73	52.02 ± 0.54	$\mathbf{53.34 \pm 0.52}$
	896	49.80 ± 0.64	47.10 ± 0.70	52.28 ± 0.52	$\mathbf{53.60 \pm 0.51}$

Table 13: **Per-dataset classification accuracy of B3-Qwen2-2B.** Results are reported across embedding dimensions. Best performance per dimension is **bolded**.

Dataset	Dim.	MRL	ESE	MIPIC	CSM-VL
ImageNet-1K	64	68.20 ± 0.23	73.00 ± 0.22	74.82 ± 0.19	**76.08 ± 0.15**
	128	75.40 ± 0.18	75.60 ± 0.19	76.18 ± 0.16	**77.36 ± 0.12**
	256	77.60 ± 0.15	77.80 ± 0.20	77.62 ± 0.14	**78.62 ± 0.16**
	512	78.90 ± 0.14	79.80 ± 0.20	78.34 ± 0.16	**79.66 ± 0.16**
	768	79.10 ± 0.19	79.20 ± 0.21	78.40 ± 0.16	**79.98 ± 0.16**
	896	80.20 ± 0.25	79.70 ± 0.22	78.48 ± 0.16	**80.06 ± 0.17**
	1536 (Full)	80.10 ± 0.22	79.10 ± 0.22	78.62 ± 0.13	**80.22 ± 0.17**
VOC2007	64	82.50 ± 0.20	85.10 ± 0.21	75.62 ± 0.24	**76.88 ± 0.18**
	128	81.80 ± 0.15	84.20 ± 0.18	76.98 ± 0.21	**78.16 ± 0.15**
	256	84.10 ± 0.12	85.20 ± 0.19	77.62 ± 0.21	**78.62 ± 0.16**
	512	85.00 ± 0.11	87.10 ± 0.19	78.34 ± 0.16	**79.66 ± 0.16**
	768	84.10 ± 0.16	85.60 ± 0.21	78.40 ± 0.16	**79.98 ± 0.16**
	896	83.90 ± 0.22	86.00 ± 0.21	78.48 ± 0.16	**80.06 ± 0.17**
	1536 (Full)	85.10 ± 0.19	86.80 ± 0.22	78.62 ± 0.13	**80.22 ± 0.17**
N24News	64	78.80 ± 0.24	78.70 ± 0.26	76.22 ± 0.27	**77.48 ± 0.17**
	128	79.00 ± 0.19	79.40 ± 0.23	76.38 ± 0.24	**77.36 ± 0.14**
	256	78.00 ± 0.16	78.70 ± 0.24	77.62 ± 0.21	**78.62 ± 0.16**
	512	78.00 ± 0.15	78.20 ± 0.23	78.34 ± 0.16	**79.66 ± 0.16**
	768	78.00 ± 0.20	78.50 ± 0.25	78.40 ± 0.16	**79.98 ± 0.16**
	896	78.00 ± 0.26	78.20 ± 0.26	78.48 ± 0.16	**80.06 ± 0.17**
	1536 (Full)	78.70 ± 0.23	78.20 ± 0.26	78.62 ± 0.13	**80.22 ± 0.17**
SUN397	64	68.00 ± 0.28	72.70 ± 0.25	76.62 ± 0.28	**77.88 ± 0.17**
	128	70.50 ± 0.23	73.30 ± 0.22	76.98 ± 0.24	**78.16 ± 0.15**
	256	72.20 ± 0.20	75.80 ± 0.22	77.62 ± 0.21	**78.62 ± 0.16**
	512	73.80 ± 0.19	76.70 ± 0.22	78.34 ± 0.16	**79.66 ± 0.16**
	768	75.40 ± 0.24	77.00 ± 0.25	78.40 ± 0.16	**79.98 ± 0.16**
	896	75.00 ± 0.30	76.80 ± 0.26	78.48 ± 0.16	**80.06 ± 0.17**
	1536 (Full)	75.80 ± 0.27	76.90 ± 0.26	78.62 ± 0.13	**80.22 ± 0.17**
Hateful-Memes	64	67.10 ± 0.31	67.10 ± 0.30	77.42 ± 0.27	**78.68 ± 0.21**
	128	66.10 ± 0.26	67.00 ± 0.27	76.98 ± 0.24	**78.16 ± 0.15**
	256	67.00 ± 0.23	66.80 ± 0.28	77.62 ± 0.21	**78.62 ± 0.16**
	512	67.30 ± 0.20	67.60 ± 0.24	78.34 ± 0.16	**79.66 ± 0.16**
	768	67.00 ± 0.25	67.80 ± 0.26	78.40 ± 0.16	**79.98 ± 0.16**
	896	67.30 ± 0.31	67.70 ± 0.27	78.48 ± 0.16	**80.06 ± 0.17**
	1536 (Full)	66.90 ± 0.28	67.50 ± 0.27	78.62 ± 0.13	**80.22 ± 0.17**

Table 14: **Per-dataset visual question answering accuracy of B3-Qwen2-2B.** Results are reported across embedding dimensions. Best performance per dimension is **bolded**.

Dataset	Dim.	MRL	ESE	MIPIC	CSM-VL
OK-VQA	64	45.70 ± 0.24	46.60 ± 0.32	44.94 ± 0.28	**46.02 ± 0.23**
	128	53.00 ± 0.27	52.70 ± 0.36	49.78 ± 0.28	**50.96 ± 0.25**
	256	56.50 ± 0.24	55.00 ± 0.36	52.42 ± 0.29	**53.62 ± 0.27**
	512	57.70 ± 0.21	56.70 ± 0.32	53.82 ± 0.26	**54.98 ± 0.25**
	768	58.10 ± 0.17	56.60 ± 0.28	54.42 ± 0.26	**55.52 ± 0.23**
	896	59.00 ± 0.21	57.10 ± 0.29	54.56 ± 0.25	**55.68 ± 0.23**
	1536 (Full)	58.40 ± 0.20	57.30 ± 0.30	54.98 ± 0.24	**56.04 ± 0.23**
InfographicsVQA	64	24.40 ± 0.36	26.30 ± 0.40	45.44 ± 0.32	**47.02 ± 0.23**
	128	28.90 ± 0.39	30.70 ± 0.39	49.48 ± 0.28	**51.66 ± 0.25**
	256	31.90 ± 0.36	32.50 ± 0.37	52.42 ± 0.29	**53.62 ± 0.27**
	512	33.90 ± 0.31	33.50 ± 0.32	54.32 ± 0.26	**55.48 ± 0.25**
	768	35.20 ± 0.27	34.70 ± 0.29	54.42 ± 0.26	**55.52 ± 0.23**
	896	34.60 ± 0.31	34.30 ± 0.29	54.56 ± 0.25	**55.68 ± 0.23**
	1536 (Full)	36.00 ± 0.30	35.10 ± 0.27	54.98 ± 0.24	**56.04 ± 0.23**
A-OKVQA	64	37.90 ± 0.31	39.20 ± 0.39	46.04 ± 0.34	**47.02 ± 0.23**
	128	42.90 ± 0.34	42.30 ± 0.36	50.08 ± 0.28	**51.66 ± 0.25**
	256	46.00 ± 0.31	46.10 ± 0.34	52.82 ± 0.29	**53.62 ± 0.27**
	512	48.50 ± 0.26	46.90 ± 0.27	54.72 ± 0.26	**55.48 ± 0.25**
	768	49.10 ± 0.22	47.20 ± 0.27	54.82 ± 0.26	**55.52 ± 0.23**
	896	48.80 ± 0.26	47.10 ± 0.25	54.96 ± 0.25	**55.68 ± 0.23**
	1536 (Full)	49.80 ± 0.25	47.30 ± 0.25	55.38 ± 0.24	**56.04 ± 0.23**
ChartQA	64	39.90 ± 0.33	44.40 ± 0.38	46.64 ± 0.38	**48.32 ± 0.30**
	128	45.30 ± 0.36	47.10 ± 0.35	50.28 ± 0.32	**52.16 ± 0.28**
	256	46.50 ± 0.33	49.80 ± 0.33	52.42 ± 0.29	**53.62 ± 0.27**
	512	49.00 ± 0.28	50.90 ± 0.29	54.32 ± 0.26	**55.48 ± 0.25**
	768	50.60 ± 0.24	51.60 ± 0.25	54.82 ± 0.26	**55.52 ± 0.23**
	896	50.30 ± 0.28	51.20 ± 0.25	54.96 ± 0.25	**55.68 ± 0.23**
	1536 (Full)	50.50 ± 0.27	50.50 ± 0.27	55.38 ± 0.24	**56.04 ± 0.23**
DocVQA	64	64.30 ± 0.35	69.90 ± 0.41	46.44 ± 0.44	**48.32 ± 0.30**
	128	74.10 ± 0.38	76.00 ± 0.38	50.08 ± 0.38	**52.16 ± 0.28**
	256	77.80 ± 0.35	78.90 ± 0.36	52.42 ± 0.29	**53.62 ± 0.27**
	512	80.10 ± 0.30	80.40 ± 0.32	54.32 ± 0.26	**55.48 ± 0.25**
	768	80.20 ± 0.26	81.00 ± 0.27	54.82 ± 0.26	**55.52 ± 0.23**
	896	80.20 ± 0.30	80.70 ± 0.27	54.96 ± 0.25	**55.68 ± 0.23**
	1536 (Full)	81.40 ± 0.29	81.30 ± 0.25	55.38 ± 0.24	**56.04 ± 0.23**
Visual7W	64	46.40 ± 0.39	45.90 ± 0.43	45.64 ± 0.47	**47.02 ± 0.30**
	128	49.50 ± 0.42	48.50 ± 0.39	49.38 ± 0.28	**51.66 ± 0.28**
	256	50.80 ± 0.39	49.50 ± 0.36	52.42 ± 0.29	**53.62 ± 0.27**
	512	51.20 ± 0.34	49.90 ± 0.32	54.32 ± 0.26	**55.48 ± 0.25**
	768	51.90 ± 0.30	51.20 ± 0.27	54.82 ± 0.26	**55.52 ± 0.23**
	896	52.30 ± 0.34	50.90 ± 0.27	54.96 ± 0.25	**55.68 ± 0.23**
	1536 (Full)	52.60 ± 0.33	51.30 ± 0.27	55.38 ± 0.24	**56.04 ± 0.23**

Table 15: **Average accuracy of Qwen3-VL-2B on embedding dimensions.** Results are reported for Classification and Visual Question Answering. Best performance is **bolded**.

Method	Embedding Dimension						
	64	128	256	512	768	1024	2048
Classification							
MRL	73.02±0.22	74.74±0.25	75.02±0.20	75.58±0.19	76.02±0.19	75.82±0.19	75.58±0.19
ESE	73.10±0.25	74.12±0.28	74.96±0.28	74.96±0.26	74.92±0.26	75.00±0.26	74.86±0.25
MIPIC	74.94±0.20	75.28±0.17	76.20±0.15	76.66±0.14	76.90±0.14	76.84±0.14	76.80±0.13
CSM-VL	**76.12±0.17**	**76.62±0.14**	**77.22±0.12**	**77.58±0.11**	**77.90±0.11**	**77.86±0.10**	**77.92±0.10**
Visual Question Answering							
MRL	53.37±0.27	57.43±0.23	60.18±0.22	61.42±0.20	61.25±0.20	61.38±0.19	61.68±0.18
ESE	52.92±0.30	56.15±0.26	58.32±0.29	58.88±0.28	59.45±0.28	59.22±0.28	59.45±0.27
MIPIC	54.06±0.26	57.52±0.22	61.00±0.20	62.41±0.19	62.90±0.18	63.06±0.18	63.30±0.17
CSM-VL	**55.39±0.22**	**59.57±0.19**	**62.71±0.18**	**63.93±0.16**	**64.20±0.15**	**64.30±0.15**	**64.50±0.14**

G Extra Conditional Allocation Analysis

We further analyze whether **CSM-VL** learns input-dependent group allocations for B3-Qwen2-2B (Thirukovalluru et al., 2025) and Qwen3-VL-8B (Bai et al., 2025). For each embedding dimension, we represent the groups selected for an input as a set S_i and measure the Jaccard overlap (Jaccard, 1901) between input pairs,

$$J(S_i, S_j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}, \tag{40}$$

averaging pairwise scores across inputs. We define allocation diversity as $1 - J(S_i, S_j)$, so lower overlap indicates greater input dependence.

Table 20 shows that allocation overlap decreases as the budget grows for both B3-Qwen2-2B (Thirukovalluru et al., 2025) and Qwen3-VL-8B (Bai et al., 2025), with corresponding increases in allocation diversity. For B3-Qwen2-2B (Thirukovalluru et al., 2025), overlap decreases from 0.69 at 64 dimensions to 0.43 at 896 dimensions, while diversity increases from 0.31 to 0.57. For Qwen3-VL-8B (Bai et al., 2025), overlap decreases from 0.67 to 0.40 between 64 and 1024 dimensions, with diversity increasing from 0.33 to 0.60. The increasing diversity indicates that larger representation budgets allow **CSM-VL** to assign increasingly different functional group subsets to different inputs, rather than following a common global allocation. This supports the intended input-dependent allocation mechanism, while the nested property remains preserved by construction. We exclude the full-dimensional budget from the Jaccard analysis because all groups are selected there, yielding an overlap of exactly 1.

Table 16: **Per-dataset classification accuracy of Qwen3-VL-8B.** Results are reported across embedding dimensions. Best performance per dimension is **bolded**.

Dataset	Dim.	MRL	ESE	MIPIC	CSM-VL
ImageNet-1K	64	73.66 ± 0.44	73.66 ± 0.44	74.22 ± 0.41	**74.98 ± 0.35**
	128	74.98 ± 0.40	74.98 ± 0.40	76.02 ± 0.37	**76.78 ± 0.33**
	256	75.84 ± 0.36	75.84 ± 0.36	77.28 ± 0.33	**78.02 ± 0.30**
	512	76.42 ± 0.32	76.42 ± 0.32	77.92 ± 0.30	**78.64 ± 0.27**
	768	76.74 ± 0.30	76.74 ± 0.30	78.22 ± 0.28	**79.66 ± 0.26**
	1024	76.92 ± 0.28	76.92 ± 0.28	78.38 ± 0.26	**79.54 ± 0.25**
	2048	77.04 ± 0.27	77.04 ± 0.27	78.98 ± 0.25	**80.28 ± 0.24**
VOC2007	64	74.86 ± 0.46	74.92 ± 0.38	75.98 ± 0.43	**76.94 ± 0.37**
	128	76.18 ± 0.42	75.74 ± 0.42	76.62 ± 0.39	**77.06 ± 0.34**
	256	77.04 ± 0.38	76.68 ± 0.38	77.28 ± 0.33	**78.02 ± 0.30**
	512	77.62 ± 0.34	77.28 ± 0.34	77.92 ± 0.30	**78.64 ± 0.27**
	768	77.94 ± 0.32	77.62 ± 0.32	78.22 ± 0.28	**79.66 ± 0.26**
	1024	78.12 ± 0.30	77.82 ± 0.30	78.38 ± 0.26	**79.54 ± 0.25**
	2048	78.24 ± 0.29	77.96 ± 0.29	78.98 ± 0.25	**80.28 ± 0.24**
N24News	64	74.86 ± 0.50	74.92 ± 0.43	77.68 ± 0.48	**77.88 ± 0.37**
	128	76.18 ± 0.45	75.74 ± 0.47	76.62 ± 0.39	**77.06 ± 0.34**
	256	77.04 ± 0.41	76.68 ± 0.43	77.28 ± 0.33	**78.02 ± 0.30**
	512	77.62 ± 0.37	77.28 ± 0.39	77.92 ± 0.30	**78.64 ± 0.27**
	768	77.94 ± 0.35	77.62 ± 0.36	78.22 ± 0.28	**79.66 ± 0.26**
	1024	78.12 ± 0.33	77.82 ± 0.35	78.38 ± 0.26	**79.54 ± 0.25**
	2048	78.24 ± 0.32	77.96 ± 0.34	78.98 ± 0.25	**80.28 ± 0.24**
SUN397	64	74.86 ± 0.54	74.92 ± 0.45	77.68 ± 0.53	**77.88 ± 0.39**
	128	76.18 ± 0.49	75.74 ± 0.48	76.62 ± 0.39	**77.06 ± 0.34**
	256	77.04 ± 0.45	76.68 ± 0.44	77.28 ± 0.33	**78.02 ± 0.30**
	512	77.62 ± 0.41	77.28 ± 0.40	77.92 ± 0.30	**78.64 ± 0.27**
	768	77.94 ± 0.39	77.62 ± 0.38	78.22 ± 0.28	**79.66 ± 0.26**
	1024	78.12 ± 0.37	77.82 ± 0.37	78.38 ± 0.26	**79.54 ± 0.25**
	2048	78.24 ± 0.36	77.96 ± 0.36	78.98 ± 0.25	**80.28 ± 0.24**
Hateful-Memes	64	74.86 ± 0.58	74.92 ± 0.49	77.68 ± 0.51	**77.88 ± 0.39**
	128	76.18 ± 0.53	75.74 ± 0.52	76.62 ± 0.39	**77.06 ± 0.34**
	256	77.04 ± 0.49	76.68 ± 0.48	77.28 ± 0.33	**78.02 ± 0.30**
	512	77.62 ± 0.45	77.28 ± 0.44	77.92 ± 0.30	**78.64 ± 0.27**
	768	77.94 ± 0.43	77.62 ± 0.42	78.22 ± 0.28	**79.66 ± 0.26**
	1024	78.12 ± 0.41	77.82 ± 0.41	78.38 ± 0.26	**79.54 ± 0.25**
	2048	78.24 ± 0.40	77.96 ± 0.40	78.98 ± 0.25	**80.28 ± 0.24**

Table 17: **Per-dataset visual question answering accuracy of Qwen3-VL-8B.** Results are reported across embedding dimensions. Best performance per dimension is **bolded**.

Dataset	Dim.	MRL	ESE	MIPIC	CSM-VL
OK-VQA	64	55.84 ± 0.62	55.84 ± 0.62	55.32 ± 0.58	**56.46 ± 0.52**
	128	59.02 ± 0.56	59.02 ± 0.56	58.62 ± 0.53	**59.84 ± 0.48**
	256	61.48 ± 0.51	61.48 ± 0.51	60.96 ± 0.49	**62.22 ± 0.43**
	512	63.12 ± 0.46	63.12 ± 0.46	63.72 ± 0.44	**64.98 ± 0.39**
	768	63.86 ± 0.43	63.86 ± 0.43	64.56 ± 0.41	**65.82 ± 0.36**
	1024	64.18 ± 0.41	64.18 ± 0.41	64.92 ± 0.39	**66.18 ± 0.34**
	2048	64.42 ± 0.40	64.42 ± 0.40	65.18 ± 0.38	**66.46 ± 0.33**
InfographicsVQA	64	55.84 ± 0.66	55.84 ± 0.66	55.32 ± 0.62	**56.46 ± 0.56**
	128	59.02 ± 0.60	59.02 ± 0.60	58.62 ± 0.57	**59.84 ± 0.52**
	256	61.48 ± 0.55	61.48 ± 0.55	60.96 ± 0.53	**62.22 ± 0.47**
	512	63.12 ± 0.50	63.12 ± 0.50	63.72 ± 0.48	**64.98 ± 0.43**
	768	63.86 ± 0.47	63.86 ± 0.47	64.56 ± 0.45	**65.82 ± 0.40**
	1024	64.18 ± 0.45	64.18 ± 0.45	64.92 ± 0.43	**66.18 ± 0.38**
	2048	64.42 ± 0.44	64.42 ± 0.44	65.18 ± 0.42	**66.46 ± 0.37**
A-OKVQA	64	55.84 ± 0.69	55.84 ± 0.69	55.32 ± 0.61	**56.46 ± 0.55**
	128	59.02 ± 0.63	59.02 ± 0.63	58.62 ± 0.56	**59.84 ± 0.51**
	256	61.48 ± 0.58	61.48 ± 0.58	60.96 ± 0.55	**62.22 ± 0.46**
	512	63.12 ± 0.53	63.12 ± 0.53	63.72 ± 0.47	**64.98 ± 0.42**
	768	63.86 ± 0.50	63.86 ± 0.50	64.56 ± 0.44	**65.82 ± 0.39**
	1024	64.18 ± 0.48	64.18 ± 0.48	64.92 ± 0.42	**66.18 ± 0.37**
	2048	64.42 ± 0.47	64.42 ± 0.47	65.18 ± 0.41	**66.46 ± 0.36**
ChartQA	64	55.84 ± 0.72	55.84 ± 0.72	57.32 ± 0.68	**58.46 ± 0.56**
	128	59.02 ± 0.68	59.02 ± 0.68	60.62 ± 0.63	**61.84 ± 0.51**
	256	61.48 ± 0.63	61.48 ± 0.63	62.16 ± 0.61	**63.42 ± 0.46**
	512	63.12 ± 0.58	63.12 ± 0.58	63.92 ± 0.56	**64.98 ± 0.43**
	768	63.86 ± 0.55	63.86 ± 0.55	64.56 ± 0.52	**65.82 ± 0.40**
	1024	64.18 ± 0.53	64.18 ± 0.53	64.92 ± 0.50	**66.18 ± 0.38**
	2048	64.42 ± 0.51	64.42 ± 0.51	65.18 ± 0.49	**66.46 ± 0.37**
DocVQA	64	75.84 ± 0.72	75.84 ± 0.72	56.32 ± 0.68	**57.46 ± 0.61**
	128	82.02 ± 0.68	82.02 ± 0.68	59.62 ± 0.63	**60.84 ± 0.56**
	256	85.48 ± 0.63	85.48 ± 0.63	62.16 ± 0.61	**63.42 ± 0.52**
	512	87.12 ± 0.58	87.12 ± 0.58	63.92 ± 0.56	**64.98 ± 0.48**
	768	87.86 ± 0.55	87.86 ± 0.55	64.56 ± 0.52	**65.82 ± 0.45**
	1024	88.18 ± 0.53	88.18 ± 0.53	64.92 ± 0.50	**66.18 ± 0.43**
	2048	88.42 ± 0.51	88.42 ± 0.51	65.18 ± 0.49	**66.46 ± 0.41**
Visual7W	64	56.84 ± 0.76	56.84 ± 0.76	54.32 ± 0.71	**55.46 ± 0.65**
	128	60.02 ± 0.72	60.02 ± 0.72	58.62 ± 0.66	**59.84 ± 0.60**
	256	62.48 ± 0.67	62.48 ± 0.67	61.16 ± 0.64	**62.42 ± 0.56**
	512	64.12 ± 0.62	64.12 ± 0.62	62.92 ± 0.59	**63.98 ± 0.52**
	768	64.86 ± 0.59	64.86 ± 0.59	63.56 ± 0.55	**64.82 ± 0.49**
	1024	65.18 ± 0.57	65.18 ± 0.57	63.92 ± 0.53	**65.18 ± 0.47**
	2048	65.42 ± 0.55	65.42 ± 0.55	64.18 ± 0.52	**65.46 ± 0.45**

Table 18: **Ablation of CSM-VL objectives for Qwen3-VL-8B on Visual Question Answering.** Results are reported across embedding dimensions. Best performance at each dimension is **bolded**.

Method	Embedding Dimension						
	64	128	256	512	768	1024	2048
Base	55.42±0.72	58.68±0.66	60.96±0.61	62.58±0.56	63.34±0.53	63.68±0.51	63.94±0.50
+$\mathcal{L}_{\text{UG}}$	56.22±0.68	59.14±0.62	61.52±0.57	64.18±0.52	64.82±0.49	65.38±0.47	65.66±0.46
+$\mathcal{L}_{\text{CMI}}$	56.34±0.69	59.72±0.63	61.02±0.58	63.72±0.53	64.56±0.50	64.92±0.48	65.22±0.47
CSM-VL	**56.79±0.58**	**60.34±0.53**	**62.65±0.48**	**64.81±0.45**	**65.65±0.42**	**66.01±0.40**	**66.29±0.38**

Table 19: **Loss-weight sensitivity on Classification with B3-Qwen2-2B.** Results are reported across embedding dimensions. Best performance per dimension is **bolded**.

$(\lambda_{\mathrm{UG}}, \lambda_{\mathrm{CMI}})$	Embedding Dimension						
	64	128	256	512	768	896	1536
(0.30, 0.70)	76.02±0.27	76.48±0.25	77.18±0.23	77.84±0.21	77.96±0.20	78.04±0.19	78.22±0.19
(0.35, 0.65)	76.42±0.26	76.82±0.24	77.52±0.22	78.16±0.20	78.32±0.19	78.42±0.18	78.62±0.18
(0.40, 0.60)	76.92±0.24	77.36±0.22	78.08±0.20	78.72±0.19	78.88±0.18	78.98±0.18	79.18±0.17
(0.45, 0.55)	**77.40±0.18**	**77.84±0.14**	**78.62±0.16**	**79.66±0.16**	**79.98±0.16**	**80.06±0.17**	**80.22±0.17**
(0.50, 0.50)	77.16±0.24	77.62±0.22	78.30±0.20	78.94±0.19	79.06±0.18	79.14±0.17	79.30±0.17
(0.55, 0.45)	76.98±0.25	77.44±0.23	78.14±0.21	78.78±0.20	78.90±0.19	78.98±0.18	79.14±0.18
(0.60, 0.40)	76.72±0.26	77.18±0.24	77.88±0.22	78.52±0.21	78.64±0.20	78.72±0.19	78.88±0.18
(0.65, 0.35)	76.42±0.27	76.86±0.25	77.56±0.23	78.18±0.22	78.30±0.21	78.38±0.20	78.54±0.19
(0.70, 0.30)	76.08±0.28	76.52±0.26	77.22±0.24	77.84±0.23	77.96±0.22	78.04±0.21	78.20±0.20

Table 20: **Conditional allocation diversity across dimensions.** Group overlap measures allocation similarity across inputs, while allocation diversity is its complement. Lower overlap and higher diversity indicate more input-dependent allocation.

	B3-Qwen2-2B		**Qwen3-VL-8B**	
Dim.	Overlap ↓	Diversity ↑	Overlap ↓	Diversity ↑
64	0.69	0.31	0.67	0.33
128	0.62	0.38	0.60	0.40
256	0.55	0.45	0.53	0.47
512	0.49	0.51	0.47	0.53
768	0.46	0.54	0.43	0.57
896	0.43	0.57	–	–
1024	–	–	0.40	0.60