CRED: CROSS-LAYER PARTICIPATION REGULARIZATION AGAINST TOKEN COLLAPSE IN VISION-LANGUAGE EMBEDDING DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language embedding models map multimodal inputs into a shared semantic representation space. Embedding distillation transfers multimodal representation knowledge from larger teacher models to more compact students, but can lead to token representation collapse in the student. We find that the effective rank of image-token representations in the student undergoes a nearly order-of-magnitude reduction relative to its initial hidden state, whereas the teacher exhibits substantially milder changes across its depth. To address this discrepancy without requiring intermediate teacher activations, we introduce **CRED**, a teacher-independent regularizer based on the participation ratio (PR). CRED constrains the PR of selected student hidden layers to remain no lower than that of the initial hidden state, thereby preserving effective dimensionality throughout the network. Unlike layer-wise feature or covariance matching, CRED operates on a scalar spectral statistic that is invariant to representation rotations and requires no additional teacher inference. Moreover, PR can be efficiently approximated using random projections, avoiding eigendecomposition and scaling more favorably to inputs with large token counts. Combined with output-level distillation, CRED reduces layer-wise token collapse and improves performance on downstream multimodal tasks in MMEB over competitive baseline methods.

1 INTRODUCTION

Knowledge distillation (KD) (Hinton et al., 2015) is widely used to transfer knowledge from large teachers to compact students. Earlier vision-language distillation methods mainly focused on visual-linguistic pretraining models or CLIP-style dual encoders, transferring intermediate features, attention patterns, cross-modal affinities, or embedding similarities (Fang et al., 2021; Sun et al., 2023; Wu et al., 2023; Yang et al., 2024). More recently, VLM2Vec (Jiang et al., 2025) has established MLLMs as general-purpose multimodal embedding models, formulating diverse tasks as retrieval in a shared embedding space. This development has motivated several forms of teacher-guided supervision. UniME (Gu et al., 2025) transfers knowledge from a text embedding teacher before multimodal training, while MAMMa (Musacchio et al., 2026) uses self-distillation during multilingual adaptation. UniME-V2 (Gu et al., 2026), U-MARVEL (Li et al., 2026b), and Qwen3-VL-Embedding (Li et al., 2026a) use semantic-score or reranker supervision; HieRD (Le et al., 2026) instead directly distills a vision-language embedding teacher into a compact student through relational alignment. Despite this progress, the internal representation dynamics of compact VLM2Vec-style students remain less understood.

Existing VLM-based embedding distillation methods primarily supervise output embeddings, relational structures, or ranking signals, leaving the effective dimensionality of intermediate token representations largely unconstrained. Here, *effective dimensionality* refers to the effective number of directions along which representation variance is distributed, which can be much smaller than the nominal feature dimension. We find that compact students exhibit substantial layer-wise reductions in the effective dimensionality of image-token representations. This reduction occurs under contrastive fine-tuning and persists under a range of distillation objectives, although its severity and location vary across architectures and objectives. Related to *dimensional collapse* (Jing et al.,

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.

2022), this finding suggests that preserving output-level semantic structure alone may not prevent degeneration of the student's internal representations. Section 3.1 provides a detailed analysis.

To address this problem, we introduce **CRED**, a vision-language embedding distillation framework that transfers multimodal semantic structure while preserving the student's internal effective dimensionality. CRED combines two relational objectives: an embedding-level objective matches query-target similarity distributions, and a structural objective preserves joint intra-modal and cross-modal geometry. Because both operate in relational space, the teacher and student need not share the same embedding dimensionality. CRED further introduces a cross-layer regularizer based on the participation ratio (PR) (Gao et al., 2017), motivated by prior uses of participation-based statistics for multimodal diversity assessment (Xia et al., 2026). The PR of the student's input-level image-token representations serves as a detached reference. Selected deeper layers are penalized only when their PR falls below this reference, preserving effective dimensionality without intermediate teacher activations or explicit teacher–student layer matching. Finally, a randomized, matrix-free PR estimator avoids explicit covariance construction and eigendecomposition, enabling regularization across multiple layers.

Our main contributions are summarized as follows:

- We identify *layer-wise dimensional collapse* in compact vision-language embedding models. Across two architectures, we show that image-token effective dimensionality decreases under contrastive fine-tuning and that this reduction persists under a range of distillation objectives.
- We introduce **CRED**, which combines relational embedding distillation with a teacher-independent cross-layer PR regularizer. We also develop a randomized, matrix-free PR estimator that avoids explicit covariance or Gram-matrix construction and scales linearly with token count and embedding dimension for a fixed number of projections.
- We demonstrate improvements in MMEB classification and visual question answering across two compact student architectures, including evaluations under distribution shift. Ablations establish the complementary roles of relational distillation and cross-layer regularization and show that CRED mitigates layer-wise dimensional collapse.

2 Related Work and Background

Distillation for Language and Vision-Language Embedding Models Recent work has explored distillation for compact embedding models. In the text domain, EMO (Truong et al., 2025) transfers relational and representation-level knowledge from embedding teachers; TALAS (Dao et al., 2026) combines teacher-guided layer alignment with student-side self-distillation; and LEAF (Vujanic & Rückstieß, 2026) learns compact teacher-compatible spaces to preserve retrieval behavior. Their feature- and relation-based objectives can also inform multimodal distillation, although they were developed in different embedding settings.

For VLM2Vec-style multimodal embeddings, recent methods use distinct sources of supervision. HieRD (Le et al., 2026) directly distills a vision-language embedding teacher into a compact student through hierarchical intra-modal and cross-modal relational alignment, making it the closest to our compression setting. UniME (Gu et al., 2025) instead transfers knowledge from a text embedding teacher before multimodal training, while MAMMa (Musacchio et al., 2026) uses self-distillation for multilingual adaptation. UniME-V2 (Gu et al., 2026) uses MLLM-generated semantic scores, whereas U-MARVEL (Li et al., 2026b) and Qwen3-VL-Embedding (Li et al., 2026a) use reranker-derived relevance supervision. These approaches primarily preserve output features, similarities, or relational structure. CRED additionally constrains the effective dimensionality of intermediate image-token representations without requiring teacher activations.

Contrastive Embedding Distillation Vision-language embedding models are commonly optimized through contrastive learning, which encourages matched query-target pairs to have higher similarity than mismatched pairs (Radford et al., 2021; Jiang et al., 2025). Given a mini-batch of B positive query-target pairs $\{(q_i, p_i)\}_{i=1}^{B}$, let $z_{q_i}^{S}$ and $z_{p_i}^{S}$ denote the L2-normalized student embeddings. Using the remaining targets in the batch as negatives, the standard InfoNCE objective is

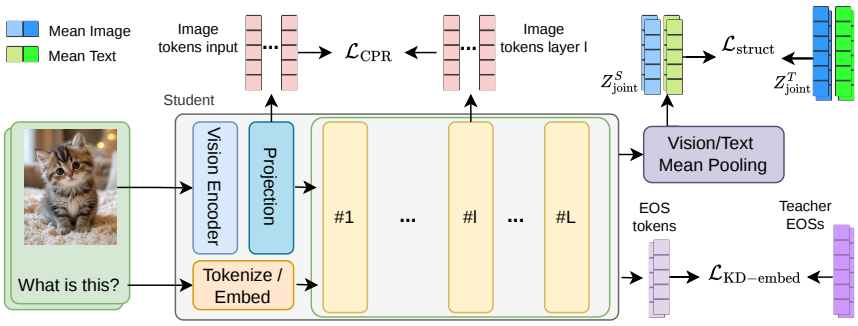

Figure 1: Overview of CRED. The student is trained with the standard contrastive objective together with teacher-driven multimodal distillation and cross-layer participation regularization. The distillation objectives transfer global query-target relations and pooled multimodal geometry, while the cross-layer objective uses the student's input-level image-token representation as a detached reference for preserving effective dimensionality at selected hidden layers.

defined as

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(\text{sim}(z^S_{q_i}, z^S_{p_i})/\tau\right)}{\sum_{j=1}^{B}\exp\left(\text{sim}(z^S_{q_i}, z^S_{p_j})/\tau\right)}, \tag{1}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a temperature parameter.

In teacher-student embedding distillation, the contrastive objective is augmented with an auxiliary distillation term:

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{InfoNCE}} + \mathcal{L}_{\text{distill}}, \tag{2}$$

where $\mathcal{L}_{\text{distill}}$ may align embeddings, similarity distributions, pairwise relations, or intermediate representations.

3 Methodology

Given a mini-batch of B query-target pairs $\{(q_i, p_i)\}_{i=1}^{B}$, let f^T and f^S denote the frozen teacher and compact student models, respectively. CRED augments the standard contrastive embedding objective with two complementary forms of supervision. First, cross-layer participation regularization (CPR) preserves the effective dimensionality of the student's image-token representations through a teacher-independent self-distillation signal. Second, unified multimodal distillation transfers the teacher's query-target similarity structure and the relational geometry of pooled visual and textual representations. Figure 1 provides an overview of the framework.

3.1 Motivation: Layer-Wise Image-Token Collapse

To quantify layer-wise representation collapse, we track the effective dimensionality of image-token representations across network depth. At layer l, we concatenate all valid image tokens into $H^{(l)} \in \mathbb{R}^{N_l \times D}$, where N_l is the total number of tokens and D is the hidden dimension. Let $\{\sigma_j^{(l)}\}_{j=1}^{r_l}$ be its singular values, with $r_l = \min(N_l, D)$, and define $p_j^{(l)} = (\sigma_j^{(l)})^2 / \sum_{k=1}^{r_l}(\sigma_k^{(l)})^2$. Following Shannon effective rank (Roy & Vetterli, 2007), we compute

$$\text{nERank}^{(l)} = \frac{1}{r_l}\exp\left(-\sum_{j=1}^{r_l} p_j^{(l)}\log p_j^{(l)}\right). \tag{3}$$

This normalization accounts for token count and hidden dimension; higher values indicate that spectral mass spans a larger fraction of the available dimensions.

We compare the non-distilled SFT student with MSE (Romero et al., 2015), RKD (Park et al., 2019), CKD (Xu et al., 2025), and HieRD (Le et al., 2026), all using the same B3-Qwen2-2B (Thirukovalluru et al., 2025) teacher. Figure 2 shows that collapse persists under distillation, but its trajectory

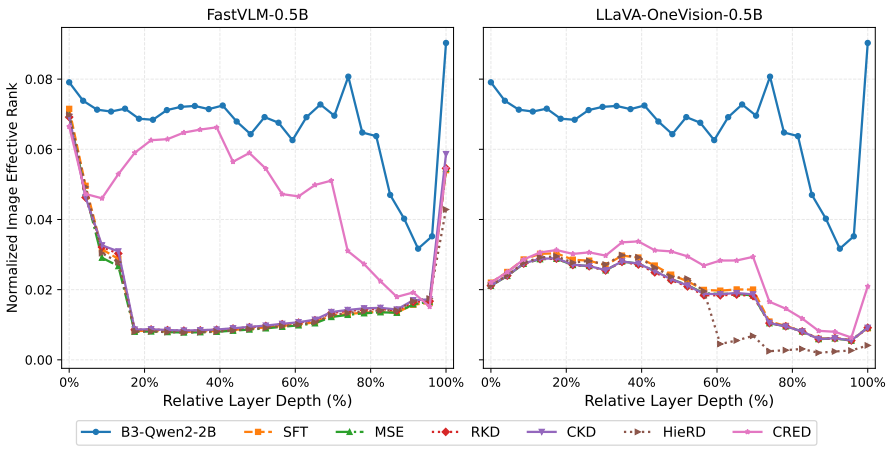

Figure 2: Layer-wise normalized effective rank of image-token representations for the B3-Qwen2-2B (Thirukovalluru et al., 2025) teacher and 0.5B students: FastVLM (Vasu et al., 2025) (**left**) and LLaVA-OneVision (Li et al., 2024) (**right**). Layers are indexed by relative depth. CRED substantially alleviates the early-to-intermediate rank collapse in FastVLM and better preserves dimensionality at several depths in LLaVA-OneVision.

depends on the student architecture. In FastVLM-0.5B (Vasu et al., 2025), all distillation baselines undergo a sharp reduction within the first 20% of the network and remain close to SFT thereafter. In LLaVA-OneVision-0.5B (Li et al., 2024), the decline is more gradual: most baselines stay near SFT, while HieRD reaches a markedly lower rank in deeper layers. The teacher maintains a higher normalized effective rank, suggesting that teacher supervision alone does not preserve the student's intermediate spectral spread. These observations motivate an explicit cross-layer dimensionality objective. Figure 2 also shows that CRED changes the FastVLM trajectory and preserves higher effective dimensionality at several middle-to-deep LLaVA-OneVision layers, without requiring the student to match the teacher's rank profile.

3.2 Cross-Layer Participation Regularization

Participation ratio. As shown in Section 3.1, compact students lose effective dimensionality across depth, a property not explicitly constrained by existing distillation objectives. Prior work has used participation-based statistics to study diversity in multimodal representations (Xia et al., 2026). We therefore use the *participation ratio* (PR) to regularize intermediate image-token representations. Unlike Shannon effective rank (Roy & Vetterli, 2007), which requires the full eigenvalue spectrum, PR (Gao et al., 2017) depends only on the first two spectral moments. For a non-zero positive semidefinite matrix Σ with eigenvalues $\{\lambda_j\}$,

$$\mathrm{PR}(\Sigma) = \frac{\left(\sum_j \lambda_j\right)^2}{\sum_j \lambda_j^2} = \frac{\mathrm{tr}(\Sigma)^2}{\mathrm{tr}(\Sigma^2)}. \tag{4}$$

PR is closely related to Shannon effective rank through Rényi entropy (Rényi, 1961): PR corresponds to exponentiated Rényi entropy of order 2, whereas Shannon effective rank corresponds to order 1. This yields

Proposition 1. *For any non-zero positive semidefinite matrix* Σ*,*

$$1 \le \mathrm{PR}(\Sigma) \le \mathrm{erank}_{\mathrm{Shannon}}(\Sigma) \le \mathrm{rank}(\Sigma). \tag{5}$$

The formal definitions and proof are provided in Appendix A.1. This inequality establishes PR as a conservative lower bound on Shannon effective rank: PR decreases as spectral mass becomes concentrated in a few dominant directions and increases as variance is distributed more evenly across dimensions. We therefore use PR as a computationally tractable surrogate for effective dimensionality. Crucially, PR depends only on $\mathrm{tr}(\Sigma)$ and $\mathrm{tr}(\Sigma^2)$, avoiding explicit eigendecomposition.

Let $\widetilde{H}^{(l)}_{i,\mathrm{img}} \in \mathbb{R}^{N_i^{\mathrm{img}} \times D}$ denote the centered image-token representation matrix for sample i at layer l, with token covariance $\Sigma^{(l)}_{i,\mathrm{img}}$. Although the participation ratio can be computed exactly from $\Sigma^{(l)}_{i,\mathrm{img}}$, exact evaluation of $\mathrm{tr}((\Sigma^{(l)}_{i,\mathrm{img}})^2)$ incurs a computational cost of $O(N_i^{\mathrm{img}} D \min\{N_i^{\mathrm{img}}, D\})$. This cost becomes prohibitive when the statistic must be evaluated at multiple layers for every sample during training. We therefore adopt a matrix-free stochastic approximation with M random projections, reducing the per-sample, per-layer complexity to $O(N_i^{\mathrm{img}} D M)$.

Matrix-free estimation. Let $a_1, \ldots, a_M$ be i.i.d. random directions uniformly distributed on the unit sphere $\mathbb{S}^{D-1}$, and let $A = [a_1, \ldots, a_M] \in \mathbb{R}^{D \times M}$. In practice, this is implemented as $a_s = g_s / \|g_s\|_2$ with standard Gaussian vectors $g_s \sim \mathcal{N}(0, I_D)$. By rotational invariance, this satisfies $\mathbb{E}[a_s a_s^\top] = I_D/D$ exactly. For clarity, we omit the sample, modality, and layer indices, and define $Y = \widetilde{H}A$. We estimate the first two spectral moments and approximate PR using the stochastic plug-in estimator

$$\widehat{\mathrm{PR}}_M(\Sigma) = \frac{\widehat{t}_1^2}{\widehat{t}_2}, \qquad \widehat{t}_1 = \frac{D}{M(N-1)} \|Y\|_F^2, \quad \widehat{t}_2 = \frac{D}{M} \left\| \frac{\widetilde{H}^\top Y}{N-1} \right\|_F^2. \tag{6}$$

Here, $\widehat{t}_1$ and $\widehat{t}_2$ are unbiased estimators of $\mathrm{tr}(\Sigma)$ and $\mathrm{tr}(\Sigma^2)$, respectively (Appendix A.2). While their ratio is not unbiased in general, the concentration properties of spherical random projections allow us to control the estimation error of $\widehat{\mathrm{PR}}_M$ non-asymptotically.

Proposition 2 (Finite-sample error of $\widehat{\mathrm{PR}}_M$)**.** *Let $\Sigma \neq 0$ be positive semidefinite and $a_1, \ldots, a_M \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\mathbb{S}^{D-1})$. Then, for any $\delta \in (0, \frac{1}{2})$,*

$$\mathbb{P}\Big[\big| \widehat{\mathrm{PR}}_M(\Sigma) - \mathrm{PR}(\Sigma) \big| \leq \delta \, \mathrm{PR}(\Sigma) \Big] \geq 1 - 4\exp\big(- cM\gamma(\delta) \big), \tag{7}$$

for some absolute constant $c > 0$, where

$$\gamma(\delta) = \min\left\{ \frac{\delta^2 \,\mathrm{tr}(\Sigma)^2}{\|\Sigma\|_F^2}, \frac{\delta \,\mathrm{tr}(\Sigma)}{\|\Sigma\|_{\mathrm{op}}}, \frac{\delta^2 \,\mathrm{tr}(\Sigma^2)^2}{\mathrm{tr}(\Sigma^4)}, \frac{\delta \,\mathrm{tr}(\Sigma^2)}{\|\Sigma\|_{\mathrm{op}}^2} \right\}. \tag{8}$$

This proposition quantifies the finite-sample error of the matrix-free PR estimator. Crucially, increasing the number of random projections M exponentially reduces the estimation risk while incurring only a linear computational cost. In practice, we treat M as a fixed hyperparameter throughout training, striking a highly effective trade-off between estimation accuracy and computational overhead. The full proof is detailed in Appendix A.2.

Teacher-independent cross-layer self-distillation. We use the student's input-level image-token representation $H^{(0)}_{i,\mathrm{img}}$, immediately after the visual input projection and before the language backbone's transformer blocks, as a dimensionality reference. This representation provides a sample-dependent baseline for the effective dimensionality available before backbone processing. A reference taken from a hidden layer may already reflect the dimensionality reduction that CPR aims to mitigate. This choice is motivated by Figure 2, where the effective-rank decline in FastVLM begins within the earliest transformer layers.

Let $\mathcal{S} \subseteq \{1, \ldots, L\}$ denote the selected student layers and $\mathcal{I}$ the valid samples in the current mini-batch. CRED applies the one-sided objective

$$\mathcal{L}_{\mathrm{CPR}} = \frac{1}{|\mathcal{S}|\,|\mathcal{I}|} \sum_{l \in \mathcal{S}} \sum_{i \in \mathcal{I}} \max\left(0, \, \mathrm{sg}\Big[\widehat{\mathrm{PR}}(\Sigma^{(0)}_{i,\mathrm{img}})\Big] - \widehat{\mathrm{PR}}(\Sigma^{(l)}_{i,\mathrm{img}})\right), \tag{9}$$

where $\mathrm{sg}[\cdot]$ denotes stop-gradient. CPR penalizes a layer only when its estimated PR falls below the detached input-level reference, leaving layers that meet or exceed this reference unpenalized. This encourages preservation of the dimensional diversity available at the student's input without requiring intermediate teacher activations or teacher–student layer correspondence.

3.3 Overall Training Objective

While Section 3.2 introduces $\mathcal{L}_{\mathrm{CPR}}$ to mitigate internal dimensional collapse, the student must simultaneously acquire robust multimodal semantic representations. CRED therefore combines a standard contrastive learning objective on the raw data with two complementary distillation objectives from a frozen teacher.

Global semantic alignment. Rather than directly matching teacher and student embeddings, which may be restrictive across different representation spaces, we align their relative retrieval behavior. For each query-target pair (q_i, p_i) and model $\star \in \{S, T\}$, let $z^{\star}_{q_i}$ and $z^{\star}_{p_i}$ denote the L2-normalized embeddings. Let $\mathcal{N}_i = \{1, \ldots, B\} \setminus \{i\}$ index the in-batch negative targets. Using cosine similarity, we define the distillation objective and the distributions over this shared support as

$$\mathcal{L}_{\text{KD-embed}} = \frac{1}{B}\sum_{i=1}^{B} D_{\mathrm{KL}}\big(\mathcal{P}_i^T \,\|\, \mathcal{P}_i^S\big), \quad \text{where } \mathcal{P}^{\star}_{i,j} = \frac{\exp\Big(\mathrm{sim}(z^{\star}_{q_i}, z^{\star}_{p_j})/\tau\Big)}{\sum_{k\in\mathcal{N}_i}\exp\big(\mathrm{sim}(z^{\star}_{q_i}, z^{\star}_{p_k})/\tau\big)},\ j \in \mathcal{N}_i. \tag{10}$$

Positive pairs are excluded from these distillation distributions and are supervised through $\mathcal{L}_{\mathrm{InfoNCE}}$. This objective transfers the teacher's relative similarities among negative targets without requiring the student to reproduce the teacher's absolute embedding coordinates.

Unified cross-modal structural preservation. Global embedding alignment alone does not explicitly preserve the relational geometry among visual and textual representations. We therefore transfer this structure by matching their pairwise similarities in a shared relational space. We collect the mean-pooled visual and textual representations into $Z^{\star}_{\mathrm{joint}}$, $\star \in \{S, T\}$, and L2-normalize each representation. The relational matrices and structural objective are defined as

$$\mathcal{L}_{\mathrm{struct}} = \mathrm{MSE}(G^S, G^T), \qquad \text{where } G^{\star} = Z^{\star}_{\mathrm{joint}}\left(Z^{\star}_{\mathrm{joint}}\right)^{\top}. \tag{11}$$

This objective preserves both intra-modal and cross-modal relational structure without directly matching student and teacher feature dimensions.

Final objective. The overall training objective of CRED integrates the task-specific contrastive loss with the distillation terms and the cross-layer dimensionality regularizer:

$$\mathcal{L}_{\mathrm{CRED}} = \mathcal{L}_{\mathrm{InfoNCE}} + \mathcal{L}_{\text{KD-embed}} + \mathcal{L}_{\mathrm{struct}} + \lambda_{\mathrm{CPR}}\mathcal{L}_{\mathrm{CPR}}, \tag{12}$$

where λ_{CPR} controls the strength of the dimensionality regularizer. The InfoNCE and distillation objectives provide complementary semantic and relational learning signals, while the teacher-independent $\mathcal{L}_{\mathrm{CPR}}$ regularizes the internal representation geometry across network depth. Together, these objectives promote multimodal relational transfer while mitigating intermediate representation collapse.

4 Experiment

4.1 Experimental Setup

Experimental setup. Following the MMEB protocol (Jiang et al., 2025), we formulate all tasks as multimodal embedding-based matching, rank candidates by cosine similarity, and report Precision@1 (P@1). We evaluate CRED on 11 in-distribution MMEB benchmarks, comprising five image classification (CLS) and six visual question answering (VQA) datasets, and further assess out-of-distribution (OOD) generalization on held-out CLS and VQA benchmarks. We use **B3-Qwen2-2B** (Thirukovalluru et al., 2025) as the default teacher and distill it into two compact vision-language models: **FastVLM-0.5B** (Vasu et al., 2025) and **LLaVA-OneVision-0.5B** (Li et al., 2024).

Baselines. We compare CRED with the non-distilled **Student** baseline and representative distillation methods, including **MSE** (Romero et al., 2015), **RKD** (Park et al., 2019), **CKD** (Xu et al., 2025), **EMO** (Truong et al., 2025), **EM-KD** (Feng et al., 2026), and **HieRD** (Le et al., 2026). Teacher performance is reported as a reference. Dataset statistics, training configuration, and detailed CRED implementation are provided in Appendices B, C.1 and C.2. Additional experimental results are presented in Appendix E.1, E.2, and E.3.

Table 1: Image classification results on MMEB. We report Precision@1 (%) across five classification benchmarks. Bold and underline indicate the best and second-best student results within each teacher-student setting, respectively.

Method	ImgNet-1K	N24News	Hateful	VOC2007	SUN397	Avg.
B3-Qwen2-2B → FastVLM-0.5B						
Teacher	82.9	79.2	56.2	88.0	80.9	77.4
Student	52.57 ±0.29	68.90 ±0.50	59.93 ±0.06	78.23 ±0.50	65.00 ±0.26	64.93
RKD	53.73 ±0.32	70.40 ±0.17	60.97 ±0.21	79.80 ±0.36	65.97 ±0.21	66.17
MSE	52.90 ±0.26	70.17 ±0.40	58.33 ±0.12	78.03 ±0.15	65.17 ±0.25	64.92
CKD	53.13 ±0.21	70.47 ±0.23	61.00 ±0.20	78.47 ±0.32	66.97 ±0.25	66.01
EMO	52.60 ±0.20	68.27 ±0.12	59.10 ±0.20	80.03 ±0.55	59.20 ±0.17	63.84
EM-KD	53.20 ±0.17	67.23 ±0.12	59.20 ±0.17	78.23 ±0.21	63.33 ±0.15	64.24
HieRD	55.93 ±0.21	71.27 ±0.38	60.83 ±0.47	80.37 ±0.25	67.13 ±0.12	67.11
CRED (Ours)	**60.60** ±0.17	**74.57** ±0.06	**61.73** ±0.40	**86.87** ±0.46	**73.63** ±0.06	**71.48**
B3-Qwen2-2B → LLaVA-OneVision-0.5B						
Teacher	82.9	79.2	56.2	88.0	80.9	77.4
Student	55.37 ±0.25	66.90 ±0.20	57.17 ±0.15	83.97 ±0.06	65.73 ±0.25	65.83
RKD	55.20 ±0.17	67.03 ±0.25	57.57 ±0.45	83.60 ±0.26	66.97 ±0.15	66.07
MSE	55.07 ±0.21	66.77 ±0.55	56.10 ±0.17	84.30 ±0.26	66.97 ±0.29	65.84
CKD	56.67 ±0.42	66.97 ±0.31	58.47 ±0.15	84.47 ±0.47	68.17 ±0.35	66.95
EMO	55.97 ±0.15	66.20 ±0.17	56.63 ±0.21	83.33 ±0.31	66.10 ±0.26	65.65
EM-KD	53.43 ±0.21	63.37 ±0.06	53.33 ±0.23	82.63 ±0.15	65.17 ±0.15	62.59
HieRD	63.27 ±0.06	73.63 ±0.15	**59.27** ±0.06	85.73 ±0.06	68.50 ±0.17	70.08
CRED (Ours)	**66.57** ±0.21	**75.03** ±0.29	53.10 ±0.72	**90.63** ±0.21	**75.70** ±0.44	**72.21**

Table 2: Visual question answering results on MMEB. We report Precision@1 (%) across six VQA benchmarks. Bold and underline indicate the best and second-best student results within each teacher-student setting, respectively.

Method	OK-VQA	A-OK	DocVQA	I-VQA	ChartQA	Vis7W	Avg.
B3-Qwen2-2B → FastVLM-0.5B							
Teacher	63.0	53.5	92.3	58.9	53.1	53.1	62.3
Student	50.70 ±0.10	50.30 ±0.36	74.63 ±0.55	34.53 ±0.40	49.67 ±0.25	46.37 ±0.38	51.03
RKD	51.13 ±0.25	51.13 ±0.31	73.93 ±0.29	35.10 ±0.17	50.10 ±0.36	46.53 ±0.25	51.32
MSE	52.63 ±0.25	50.13 ±0.31	74.20 ±0.17	34.03 ±0.25	50.93 ±0.15	46.77 ±0.35	51.45
CKD	51.87 ±0.31	50.73 ±0.38	74.83 ±0.23	35.00 ±0.40	49.97 ±0.32	45.90 ±0.36	51.38
EMO	51.03 ±0.15	49.60 ±0.46	73.30 ±0.20	34.03 ±0.32	48.50 ±0.50	47.47 ±0.35	50.66
EM-KD	51.17 ±0.06	49.43 ±0.23	77.17 ±0.06	36.80 ±0.10	47.47 ±0.23	47.33 ±0.15	51.56
HieRD	52.60 ±0.26	51.70 ±0.10	77.57 ±0.06	38.07 ±0.21	50.07 ±0.31	46.53 ±0.31	52.76
CRED (Ours)	**54.53** ±0.29	**52.03** ±0.06	**81.07** ±0.12	**39.63** ±0.49	**54.37** ±0.12	**52.43** ±0.23	**55.68**
B3-Qwen2-2B → LLaVA-OneVision-0.5B							
Teacher	63.0	53.5	92.3	58.9	53.1	53.1	62.3
Student	47.93 ±0.31	43.37 ±0.60	38.43 ±0.15	21.33 ±0.45	25.17 ±0.40	43.80 ±0.30	36.67
RKD	47.83 ±0.25	43.60 ±0.17	39.07 ±0.42	21.57 ±0.55	25.50 ±0.20	43.80 ±0.26	36.89
MSE	48.07 ±0.21	42.20 ±0.30	37.90 ±0.30	21.17 ±0.21	24.73 ±0.15	42.97 ±0.40	36.17
CKD	48.05 ±0.40	42.75 ±0.21	36.30 ±0.25	19.35 ±0.21	25.35 ±0.32	43.70 ±0.25	35.92
EMO	46.17 ±0.21	43.53 ±0.31	41.20 ±0.17	20.83 ±0.15	22.67 ±0.40	42.97 ±0.15	36.23
EM-KD	44.67 ±0.06	38.07 ±0.21	36.77 ±0.47	21.30 ±0.72	22.43 ±0.38	39.53 ±0.23	33.79
HieRD	52.43 ±0.21	46.93 ±0.15	49.23 ±0.31	23.40 ±0.20	34.60 ±0.26	46.90 ±0.20	42.25
CRED (Ours)	**53.10** ±0.20	**47.13** ±0.12	**61.40** ±0.20	**27.27** ±0.06	**37.23** ±0.06	**50.07** ±0.15	**46.03**

4.2 Main Results

We first evaluate CRED on the in-distribution CLS and VQA benchmarks used for training, and then assess its generalization to held-out OOD CLS and VQA datasets without additional fine-tuning. We compare against the SFT student and representative feature-, relational-, and embedding-distillation baselines under the same teacher-student settings.

Table 3: Training runtime and GPU memory usage.

Method	Time (s/step)	Avg. Mem (GB)	Peak Mem (GB)
RKD	1.54	14.64	29.18
MSE	1.33	14.05	28.65
CKD	1.87	14.11	28.65
EMO	4.23	18.34	63.40
EM-KD	4.16	19.72	118.14
HieRD	5.50	18.36	63.48
CRED	1.41	14.18	30.67
CRED (exact PR)	1.74	14.19	30.97

Table 4: Effect of CPR across distillation methods. IND classification P@1 (%); bold indicates the best student result.

Method	**ImgNet**	**N24**	**Hateful**	**VOC**	**SUN**	**Avg.**
Teacher	82.90	79.20	56.20	88.00	80.90	77.40
SFT	52.57	68.90	59.93	78.23	65.00	64.93
CKD	53.13	70.47	61.00	78.47	66.97	66.01
CKD + CPR	58.30	72.20	61.40	81.10	67.20	68.04
HieRD	55.93	71.27	60.83	80.37	67.13	67.11
HieRD + CPR	59.00	73.00	61.90	80.40	66.50	68.16
CRED w/o CPR	**60.74**	72.92	**64.10**	85.56	67.38	70.14
CRED	60.60	**74.57**	61.73	**86.87**	**73.63**	**71.48**

In-distribution performance. Tables 1 and 2 show that CRED achieves the highest average CLS and VQA performance across both student architectures. Its gains over HieRD across both task families suggest that preserving internal dimensionality complements teacher-guided relational transfer. CRED achieves the best student results on all six VQA benchmarks for both architectures, with a particularly large gain of 12.17 points over HieRD on DocVQA for LLaVA-OneVision. Although performance declines on HatefulMemes with LLaVA-OneVision, CRED improves over SFT on every remaining benchmark across both architectures, demonstrating broad benefits. Together with the layer-wise analysis, these results support combining output-level knowledge transfer with intermediate dimensionality preservation to improve performance across diverse multimodal tasks.

Out-of-distribution generalization. Figure 3 shows consistent OOD gains across held-out CLS and VQA benchmarks for both student architectures. These results support the motivation for preserving effective dimensionality: discouraging spectral concentration can retain more diverse representation directions that remain useful beyond the training distribution. Specifically, CRED improves OOD VQA over SFT by 7.04 and 12.33 points for FastVLM and LLaVA-OneVision, respectively. Together with the higher layer-wise effective ranks in Figure 2, the OOD improvements suggest that CRED preserves transferable representational diversity. Further benchmark-level analysis is provided in Section E.1.

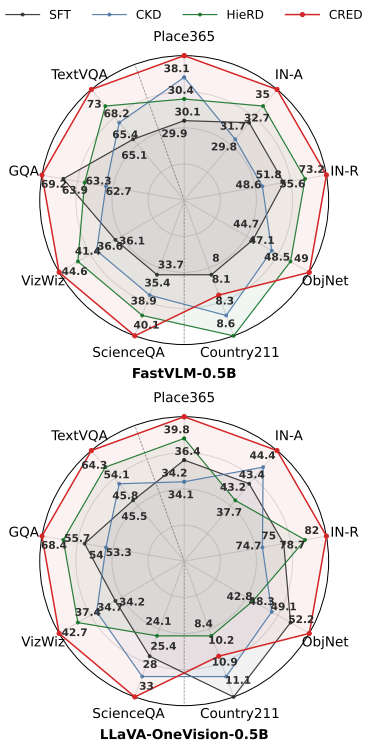

Figure 3: Out-of-distribution performance of representative distillation methods. Each axis reports P@1 (%) on a held-out classification or VQA benchmark.

4.3 Ablation Study

Unless otherwise specified, all ablation experiments follow our default experimental setting. We analyze the contribution of individual components and the sensitivity of CRED to key design choices, with additional results provided in Appendix D.

Computational efficiency. Table 3 shows that CRED combines dimensionality preservation with practical training costs, requiring 1.41 s/step and 30.67 GB peak memory. The matched exact-PR variant isolates the estimator's computational benefit: randomized estimation reduces total step time by 18.97% with similar memory usage. Together with the comparable downstream results in Table 9, this supports matrix-free estimation as an efficient substitute for exact PR computation.

Effect of loss components. To assess the contribution and complementarity of each objective in CRED, we conduct an ablation study on the embedding distillation, structural preservation, and CPR losses as shown in

Table 5: Loss-component ablation. Average P@1 (%) on IND and OOD classification benchmarks.

$\mathcal{L}_{\text{InfoNCE}}$	$\mathcal{L}_{\text{KD-embed}}$	$\mathcal{L}_{\text{struct}}$	$\mathcal{L}_{\text{CPR}}$	**IND**	**OOD**	**Avg.**
✓				64.93	33.22	49.08
✓	✓			70.06	38.12	54.09
✓		✓		68.94	37.38	53.16
✓			✓	68.92	37.54	53.23
✓	✓	✓		70.14	38.21	54.18
✓	✓		✓	**71.94**	39.14	55.54
✓		✓	✓	68.92	37.50	53.21
✓	✓	✓	✓	71.48	**40.70**	**56.09**

Table 5. Each auxiliary loss improves over SFT individually, while combining embedding distillation with CPR achieves the strongest IND performance. Adding structural preservation to this combination further improves OOD performance, yielding the best overall average and suggesting a complementary role under distribution shift. Crucially, adding CPR to the two teacher-driven losses raises OOD performance from 38.21% to 40.70%, directly supporting its contribution beyond output-level transfer. The benefits are nevertheless configuration-dependent: combining structural preservation with CPR alone provides no improvement over either component individually. These results favor the full objective for balanced IND and OOD performance, rather than implying that all loss combinations are uniformly beneficial.

Generalization of CPR Across Distillation Methods Figure 4 shows that intermediate-layer collapse is already present in pretrained FastVLM and persists under SFT and several distillation objectives. Adding CPR raises normalized effective rank across intermediate layers for CKD, HieRD, and CRED, indicating that its effect does not depend on a particular teacher-transfer loss. This directly supports the intended action of the PR penalty: counteracting spectral concentration where output-level supervision alone leaves it largely unchanged. The agreement across objectives strengthens the interpretation that CPR addresses a shared limitation of compact student representations. Table 4 connects these geometric changes to downstream utility: CPR improves average classification performance for CKD, HieRD, and CRED by 2.03, 1.05, and 1.34 points, respectively. Thus, the increased effective rank accompanies measurable task gains across all three objectives. Individual benchmarks do not improve uniformly, but the consistent average gains support dimensionality preservation as complementary supervision rather than a benefit specific to CRED's distillation formulation.

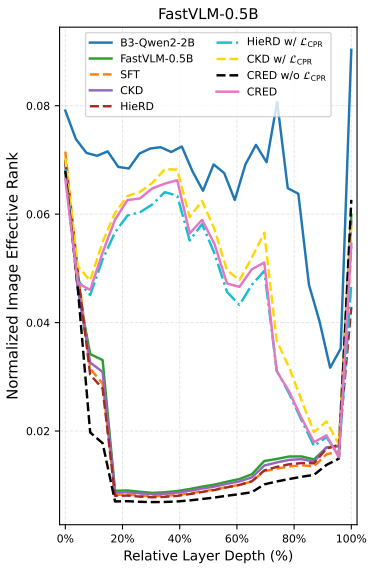

Figure 4: Layer-wise normalized ERank with and without CPR.

5 Conclusion

We introduced **CRED**, a vision-language embedding distillation framework that combines semantic and relational knowledge transfer with cross-layer dimensionality preservation. To address image-token dimensional collapse in compact students, CRED uses a one-sided participation-ratio penalty with the student's input-level representations as a detached reference. The regularizer requires no intermediate teacher activations or teacher–student layer matching, and a randomized matrix-free estimator enables efficient computation. Across two compact architectures, CRED improves average MMEB classification and visual question answering performance, as well as results on held-out classification benchmarks. Ablations and layer-wise analyses show that dimensionality preservation complements distillation objectives and mitigates internal collapse. Together, these findings support supervising internal effective dimensionality alongside semantic and relational alignment.

REPRODUCIBILITY STATEMENT

Training Implementation. We implement our training framework in PyTorch using bfloat16 precision. Detailed implementation, dataset statistics, and hyperparameters are provided in Appendices B and C.

Inference Implementation. At inference time, we encode queries and candidates with the trained student and rank candidates by cosine similarity. Evaluation requires neither the teacher nor the CPR regularizer, following the standard MMEB evaluation protocol.

Evaluation Datasets. We evaluate the distilled models on the MMEB benchmark, covering its classification and visual question answering (VQA) tasks. Detailed dataset statistics are provided in Appendix B.

Source Code. We include the source code in the supplementary materials to support reproducibility and will release it publicly upon acceptance.

Computational Resources. All experiments were conducted on a single NVIDIA B200 GPU. Detailed hardware configurations and training hyperparameters are provided in Appendix C.

AI USE STATEMENT

In this work, generative AI tools were employed to assist with identifying and synthesizing relevant literature to assess potential overlap with prior work, and improving the clarity and readability of the manuscript. These tools were also used to help resolve minor coding issues.

Generative AI was not used to develop theoretical models, formulate mathematical claims, establish key mathematical arguments, or construct proofs. Nor was it used to design research methodologies or experiments, implement the proposed method, translate materials, clean or reformat datasets, perform qualitative analyses, or interpret experimental results. All other disclosure categories were not applicable to this work.

All AI-assisted outputs were reviewed and verified by the authors, who retain full responsibility for the research methodology, theoretical contributions, experimental findings, and final manuscript.

ETHICS STATEMENT

This work investigates knowledge distillation for vision-language embedding models, with a focus on preserving the effective dimensionality of intermediate representations. Our experiments are conducted using publicly available research datasets and pretrained models. We do not collect new personal data or conduct experiments involving human participants.

Although our method aims to improve the efficiency of multimodal representation learning, distilled models may inherit biases and limitations from their teacher models and training data. These issues should be carefully considered when deploying such models in real-world applications.

REFERENCES

Quoc Phong Dao, Hoang Son Nguyen, Pham Khanh Chi, Linh Ngo Van, Nguyen Thi Ngoc Diep, Thien Huu Nguyen, and Trung Le. TALAS: Teacher-anchored layer alignment with adaptive sharpness-aware minimization for embedding distillation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32705–32723, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1509. URL https://aclanthology.org/2026.acl-long.1509/.

Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Compressing visual-linguistic model via knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1428–1438, October 2021.

Ze Feng, Sen Yang, Boqiang Duan, Wankou Yang, and Jingdong Wang. Em-kd: Distilling efficient multimodal large language model with unbalanced vision tokens. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21111–21119, 2026.

Peiran Gao, Eric Trautmann, Byron Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv*, 2017. doi: 10.1101/214262. URL https://www.biorxiv.org/content/early/2017/11/12/214262.

Tiancheng Gu, Kaicheng Yang, Ziyong Feng, Xingjun Wang, Yanzhao Zhang, Dingkun Long, Yingda Chen, Weidong Cai, and Jiankang Deng. Breaking the modality barrier: Universal embedding learning with multimodal llms. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 2860–2869, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3754845. URL https://doi.org/10.1145/3746027.3754845.

Tiancheng Gu, Kaicheng Yang, Kaichen Zhang, Xiang An, Ziyong Feng, Yueyi Zhang, Weidong Cai, Jiankang Deng, and Lidong Bing. Unime-v2: Mllm-as-a-judge for universal multimodal embedding learning. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i26.39284. URL https://doi.org/10.1609/aaai.v40i26.39284.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015. URL https://api.semanticscholar.org/CorpusID:7200347.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 1255–1279, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/04261fce1705c4f02f062866717d592a-Paper-Conference.pdf.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=YevsQ05DEN7.

Vinh Le, Nguyen Hong Dang, Tu Vu, Linh Ngo Van, Duc Anh Nguyen, and Trung Le. HieRD: Hierarchical relational distillation for vision-language embedding models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=ZyqAd4xpT5.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026a.

Xiaojie Li, Chu Li, Shi-Zhe Chen, and Xi Chen. U-MARVEL: Unveiling key factors for universal multimodal retrieval via embedding learning with MLLMs. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=CUFJluq5O8.

Elio Musacchio, Lucia Siciliani, Pierpaolo Basile, and Giovanni Semeraro. Mamma: adaptation of multimodal llm embeddings to multilinguality. *Neural Computing and Applications*, 38(15), August 2026. ISSN 1433-3058. doi: 10.1007/s00521-026-12339-x. URL http://dx.doi.org/10.1007/s00521-026-12339-x.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3962–3971, 2019. URL https://api.semanticscholar.org/CorpusID:131765296.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pp. 547–562. University of California Press, 1961.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6550.

Olivier Roy and M. Vetterli. The effective rank: a measure of effective dimensionality, September 2007. URL https://doi.org/10.5281/zenodo.40328.

Ximeng Sun, Pengchuan Zhang, Peizhao Zhang, Hardik Shah, Kate Saenko, and Xide Xia. Dime-fm : Distilling multimodal and efficient foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15521–15533, October 2023.

Raghuveer Thirukovalluru, Rui Meng, Ye Liu, Karthikeyan K, 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 D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 23218–23241. Curran Associates, Inc., 2025. doi: 10.52202/085713-0784. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/21aa6840fcaa0cc3a1d16bafede47a2f-Paper-Conference.pdf.

Minh-Phuc Truong, Hai An Vu, Tu Vu, Nguyen Thi Ngoc Diep, Linh Ngo Van, Thien Huu Nguyen, and Trung Le. Emo: Embedding model distillation via intra-model relation and optimal transport alignments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. URL https://aclanthology.org/2025.emnlp-main.385/.

Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokul 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)*, June 2025.

Robin Vujanic and Thomas Rückstieß. LEAF: Knowledge distillation of text embedding models with teacher-aligned representations. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 43362–43383, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.2008. URL https://aclanthology.org/2026.acl-long.2008/.

Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi (Stephen) Chen, Xinggang Wang, Hongyang Chao, and Han Hu. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21970–21980, October 2023.

Zixuan Xia, Hao Wang, Pengcheng Weng, Yanyu Qian, Yangxin Xu, William Dan, and Fei Wang. Diverse via bounded agreement: Geometric regularization for multimodal fusion. *arXiv preprint arXiv:2601.21670*, 2026.

Alex Tianyi Xu, Alex Wilf, Paul Pu Liang, Alexander Obolenskiv, Daniel Fried, and Louis-Philippe Morency. Comparative knowledge distillation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7701–7710, 2025. doi: 10.1109/WACV61041.2025.00748.

Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. Clip-kd: An empirical study of clip model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15952–15962, June 2024.

A Theoretical Foundations of CRED

A.1 Participation Ratio and Shannon Effective Rank

Let $\Sigma \in \mathbb{R}^{D\times D}$ be a non-zero positive semidefinite matrix with rank r and non-zero eigenvalues $\lambda_1, \ldots, \lambda_r > 0$. We define the normalized eigenvalue spectrum as

$$p_i = \frac{\lambda_i}{\sum_{j=1}^{r} \lambda_j}, \qquad \sum_{i=1}^{r} p_i = 1. \tag{13}$$

The Shannon effective rank (Roy & Vetterli, 2007) is defined as

$$\mathrm{erank}_{\mathrm{Shannon}}(\Sigma) = \exp\left(-\sum_{i=1}^{r} p_i \log p_i\right). \tag{14}$$

The participation ratio (Gao et al., 2017) is

$$\mathrm{PR}(\Sigma) = \frac{1}{\sum_{i=1}^{r} p_i^2} = \frac{\left(\sum_{i=1}^{r} \lambda_i\right)^2}{\sum_{i=1}^{r} \lambda_i^2} = \frac{\mathrm{tr}(\Sigma)^2}{\mathrm{tr}(\Sigma^2)}. \tag{15}$$

Both quantities can be interpreted through Rényi entropy (Rényi, 1961). For $\alpha > 0$ and $\alpha \neq 1$,

$$H_\alpha(p) = \frac{1}{1-\alpha} \log\left(\sum_{i=1}^{r} p_i^\alpha\right). \tag{16}$$

The Shannon entropy is recovered by continuous extension at $\alpha = 1$:

$$H_1(p) = \lim_{\alpha\to 1} H_\alpha(p) = -\sum_{i=1}^{r} p_i \log p_i. \tag{17}$$

Therefore,

$$\mathrm{erank}_{\mathrm{Shannon}}(\Sigma) = \exp(H_1(p)), \qquad \mathrm{PR}(\Sigma) = \exp(H_2(p)). \tag{18}$$

The following proposition formalizes the relationship between PR and Shannon effective rank.

Proposition 3. *For any non-zero positive semidefinite matrix Σ,*

$$1 \le \mathrm{PR}(\Sigma) \le \mathrm{erank}_{\mathrm{Shannon}}(\Sigma) \le \mathrm{rank}(\Sigma). \tag{19}$$

Moreover,

$$\mathrm{PR}(\Sigma) = \mathrm{erank}_{\mathrm{Shannon}}(\Sigma) \tag{20}$$

if and only if all non-zero eigenvalues of Σ are equal.

Proof. Let $s_2 = \sum_{i=1}^{r} p_i^2$. Since $\log(\cdot)$ is concave, Jensen's inequality gives

$$\sum_{i=1}^{r} p_i \log p_i \le \log\left(\sum_{i=1}^{r} p_i^2\right) = \log s_2. \tag{21}$$

Therefore,

$$H_1(p) = -\sum_{i=1}^{r} p_i \log p_i \geq -\log s_2 = H_2(p). \tag{22}$$

Exponentiating both sides and using equation 18 yields

$$\mathrm{PR}(\Sigma) \leq \mathrm{erank}_{\mathrm{Shannon}}(\Sigma). \tag{23}$$

For the lower bound, since $0 \leq p_i \leq 1$ and $\sum_{i=1}^{r} p_i = 1$,

$$\sum_{i=1}^{r} p_i^2 \leq \sum_{i=1}^{r} p_i = 1, \tag{24}$$

which implies

$$\mathrm{PR}(\Sigma) = \frac{1}{\sum_{i=1}^{r} p_i^2} \geq 1. \tag{25}$$

For the upper bound, Jensen's inequality gives

$$H_1(p) = \sum_{i=1}^{r} p_i \log \frac{1}{p_i} \leq \log \left(\sum_{i=1}^{r} p_i \frac{1}{p_i} \right) = \log r. \tag{26}$$

Therefore,

$$\mathrm{erank}_{\mathrm{Shannon}}(\Sigma) \leq r = \mathrm{rank}(\Sigma). \tag{27}$$

Finally, equality in Jensen's inequality holds if and only if p_i is constant over its support. Since $\sum_{i=1}^{r} p_i = 1$, this requires $p_i = 1/r$ for all $i = 1, \ldots, r$, which by equation 13 is equivalent to all non-zero eigenvalues of Σ being equal. □

Thus, when computed from the same normalized spectrum, PR provides a conservative measure of effective dimensionality relative to Shannon effective rank. Unlike Shannon effective rank, PR depends only on the first two spectral moments through $\mathrm{tr}(\Sigma)$ and $\mathrm{tr}(\Sigma^2)$, enabling the efficient matrix-free estimator introduced in Appendix A.2.

A.2 Proof details for the matrix-free PR estimator

This section establishes the finite-sample error bound for $\widehat{\mathrm{PR}}_M(\Sigma)$ stated in Proposition 2. For clarity, we consider a single sample and omit the sample and layer indices.

Estimator construction and unbiasedness. Recall that

$$\Sigma = \frac{\widetilde{H}^\top \widetilde{H}}{N-1}, \qquad \widetilde{H} \in \mathbb{R}^{N \times D}, \qquad N \geq 2, \tag{28}$$

where $\widetilde{H}$ is the centered image-token representation matrix defined in Section 3.2. The estimator is applied independently to each valid sample in the mini-batch.

Let $a_1, \ldots, a_M \in \mathbb{R}^D$ be independent random directions satisfying

$$\mathbb{E}[a_s a_s^\top] = \frac{I_D}{D}, \qquad s = 1, \ldots, M. \tag{29}$$

In our implementation,

$$g_s \sim \mathcal{N}(0, I_D), \qquad a_s = \frac{g_s}{\|g_s\|_2}, \tag{30}$$

which satisfies the above identity by rotational invariance. Hence, for any symmetric matrix B,

$$\mathbb{E}[a^\top B a] = \frac{\mathrm{tr}(B)}{D}, \qquad \mathbb{E}[\|Ba\|_2^2] = \mathbb{E}[a^\top B^2 a] = \frac{\mathrm{tr}(B^2)}{D}. \tag{31}$$

Let

$$A = [a_1, \ldots, a_M] \in \mathbb{R}^{D \times M}, \qquad Y = \widetilde{H} A \in \mathbb{R}^{N \times M}. \tag{32}$$

Then

$$\widehat{t}_1 = \frac{D}{M}\sum_{s=1}^{M} a_s^\top \Sigma a_s = \frac{D}{M(N-1)}\|Y\|_F^2, \tag{33}$$

and

$$\widehat{t}_2 = \frac{D}{M}\sum_{s=1}^{M} \|\Sigma a_s\|_2^2 = \frac{D}{M}\left\|\frac{\widetilde{H}^\top Y}{N-1}\right\|_F^2. \tag{34}$$

Therefore,

$$\mathbb{E}[\widehat{t}_1] = \mathrm{tr}(\Sigma), \qquad \mathbb{E}[\widehat{t}_2] = \mathrm{tr}(\Sigma^2), \tag{35}$$

so both spectral-moment estimators are unbiased. We then define the stochastic plug-in estimator

$$\widehat{\mathrm{PR}}_M(\Sigma) = \frac{\widehat{t}_1^2}{\widehat{t}_2}. \tag{36}$$

Concentration of the spectral-moment estimators. We first study a single random direction. Let $B \succeq 0$ be nonzero, let $a \sim \mathrm{Unif}(S^{D-1})$, and define $X_B = Da^\top Ba$. By rotational symmetry, $\mathbb{E}[aa^\top] = I_D/D$, so $\mathbb{E}[X_B] = \mathrm{tr}(B)$.

To bound fluctuations of X_B, write a standard Gaussian vector as $g = Ra$, where $g \sim \mathcal{N}(0, I_D)$, $R = \|g\|_2$ is independent of a, and $\mathbb{E}[R^2] = D$. Conditional on a, Jensen's inequality gives

$$\begin{aligned} e^{\theta D a^\top Ba} &= e^{\theta \mathbb{E}[R^2 a^\top Ba | a]} \\ &\le \mathbb{E}\left[e^{\theta R^2 a^\top Ba} \mid a\right] = \mathbb{E}\left[e^{\theta g^\top Bg} \mid a\right]. \end{aligned} \tag{37}$$

Taking expectations over a yields

$$\mathbb{E}\left[e^{\theta X_B}\right] \le \mathbb{E}\left[e^{\theta g^\top Bg}\right]. \tag{38}$$

The inequality holds for either sign of θ whenever the Gaussian expectation is finite.

Let $\lambda_1, \dots, \lambda_D$ be the eigenvalues of B. Here, $\|B\|_{\mathrm{op}} = \sup_{\|v\|_2=1}\|Bv\|_2$ denotes the operator norm; for $B \succeq 0$, it equals the largest eigenvalue of B. For $|\theta| \le 1/(4\|B\|_{\mathrm{op}})$, the Gaussian quadratic-form moment generating function and the bound $-\log(1-u) - u \le u^2$ for $|u| \le 1/2$ give

$$\begin{aligned} \log \mathbb{E}\left[e^{\theta(X_B - \mathrm{tr}(B))}\right] &\le \sum_{i=1}^{D}\left[-\frac{1}{2}\log(1-2\theta\lambda_i) - \theta\lambda_i\right] \\ &\le 2\theta^2 \sum_{i=1}^{D}\lambda_i^2 = 2\theta^2\|B\|_F^2. \end{aligned} \tag{39}$$

Now let $a_1, \dots, a_M$ be independent random directions, and set $\widehat{t}_B = M^{-1}\sum_{s=1}^{M} Da_s^\top Ba_s$. Independence allows us to add the preceding logarithmic bounds:

$$\log \mathbb{E}\left[e^{\theta M(\widehat{t}_B - \mathrm{tr}(B))}\right] \le 2M\theta^2\|B\|_F^2, \qquad |\theta| \le \frac{1}{4\|B\|_{\mathrm{op}}}. \tag{40}$$

Applying the exponential Markov inequality to the upper tail with

$$\theta = \min\left\{\frac{\varepsilon \mathrm{tr}(B)}{4\|B\|_F^2}, \frac{1}{4\|B\|_{\mathrm{op}}}\right\},$$

and to the lower tail with the corresponding negative value, gives

$$\mathbb{P}\left(|\widehat{t}_B - \mathrm{tr}(B)| > \varepsilon \mathrm{tr}(B)\right) \le 2\exp\left[-\frac{M}{8}\min\left\{\frac{\varepsilon^2 \mathrm{tr}(B)^2}{\|B\|_F^2}, \frac{\varepsilon \mathrm{tr}(B)}{\|B\|_{\mathrm{op}}}\right\}\right]. \tag{41}$$

Applying this result with $B = \Sigma$ and $B = \Sigma^2$, and using a union bound, we obtain

$$\mathbb{P}\left(\begin{array}{c} |\widehat{t}_1 - \mathrm{tr}(\Sigma)| \le \varepsilon \mathrm{tr}(\Sigma), \\ |\widehat{t}_2 - \mathrm{tr}(\Sigma^2)| \le \varepsilon \mathrm{tr}(\Sigma^2) \end{array}\right) \ge 1 - 4\exp\left[-\frac{M}{8}\gamma(\varepsilon)\right], \tag{42}$$

where

$$\gamma(\varepsilon) = \min\left\{\frac{\varepsilon^2 \mathrm{tr}(\Sigma)^2}{\|\Sigma\|_F^2}, \frac{\varepsilon \mathrm{tr}(\Sigma)}{\|\Sigma\|_{\mathrm{op}}}, \frac{\varepsilon^2 \mathrm{tr}(\Sigma^2)^2}{\mathrm{tr}(\Sigma^4)}, \frac{\varepsilon \mathrm{tr}(\Sigma^2)}{\|\Sigma\|_{\mathrm{op}}^2}\right\}. \tag{43}$$

Here we used $\|\Sigma^2\|_F^2 = \mathrm{tr}(\Sigma^4)$ and $\|\Sigma^2\|_{\mathrm{op}} = \|\Sigma\|_{\mathrm{op}}^2$.

Error of the PR estimator. On the event in Eq. equation 42, write

$$\hat{t}_1 = (1+e_1)\,\mathrm{tr}(\Sigma), \qquad \hat{t}_2 = (1+e_2)\,\mathrm{tr}(\Sigma^2), \qquad |e_1|, |e_2| \le \varepsilon. \tag{44}$$

Since $\widehat{\mathrm{PR}}_M = \hat{t}_1^2/\hat{t}_2$,

$$\frac{\widehat{\mathrm{PR}}_M(\Sigma)}{\mathrm{PR}(\Sigma)} = \frac{(1+e_1)^2}{1+e_2}. \tag{45}$$

For $0 < \varepsilon \le 1/2$,

$$\left| \frac{(1+e_1)^2}{1+e_2} - 1 \right| \le \frac{(1+\varepsilon)^2}{1-\varepsilon} - 1 = \frac{3\varepsilon + \varepsilon^2}{1-\varepsilon} \le 7\varepsilon. \tag{46}$$

Set $\varepsilon = \delta/7$. The preceding inequality gives

$$\left| \widehat{\mathrm{PR}}_M(\Sigma) - \mathrm{PR}(\Sigma) \right| \le \delta\,\mathrm{PR}(\Sigma). \tag{47}$$

Since every term in γ is linear or quadratic in its argument, $\gamma(\delta/7) \ge \gamma(\delta)/49$. Consequently, the bound holds with probability at least

$$1 - 4\exp\left[-\frac{M}{392}\gamma(\delta) \right]. \tag{48}$$

This proves Proposition 2 with $c = 1/392$.

Computational remark. The estimator requires only the matrix-matrix products $Y = \widetilde{H}A$ and $\widetilde{H}^\top Y$, with total computational cost $O(NDM)$. Thus, when $M \ll \min(N, D)$, it avoids explicitly materializing either the $D \times D$ covariance matrix or the $N \times N$ Gram matrix. Their formation costs $O(ND^2)$ and $O(N^2D)$, respectively.

B DATASETS

We evaluate CRED on tasks from the MMEB (Jiang et al., 2025). Training uses five classification datasets (ImageNet-1K, N24News, HatefulMemes, VOC2007, and SUN397) and six VQA datasets (OK-VQA, A-OKVQA, DocVQA, InfographicsVQA, ChartQA, and Visual7W), with evaluation on their corresponding MMEB splits. OOD generalization is assessed on held-out datasets not used for training: Place365, ImageNet-A, ImageNet-R, ObjectNet, and Country-211 for classification, and ScienceQA, VizWiz, GQA, and TextVQA for VQA. We also evaluate visual grounding, using MSCOCO for training and RefCOCO, RefCOCO-Matching, and Visual7W-Pointing for additional evaluation. Following MMEB, all tasks are formulated as matching between independently encoded multimodal queries and candidate targets. Table 6 summarizes the datasets and evaluation protocol.

Table 6: Dataset statistics used in our experiments. "I" and "T" denote image and text, respectively. Datasets without training instances are used only for additional or OOD evaluation.

Task	Dataset	Query	Target	Train	Eval	Cand.
Classification	ImageNet-1K	I	T	100K	1K	1000
	N24News	I+T	I	49K	1K	24
	HatefulMemes	I	T	8K	1K	2
	VOC2007	I	T	8K	1K	20
	SUN397	I	T	20K	1K	397
	Place365	I	T	–	1K	365
	ImageNet-A	I	T	–	1K	1000
	ImageNet-R	I	T	–	1K	200
	ObjectNet	I	T	–	1K	313
	Country-211	I	T	–	1K	211
VQA	OK-VQA	I+T	T	9K	1K	1000
	A-OKVQA	I+T	T	17K	1K	1000
	DocVQA	I+T	T	40K	1K	1000
	InfographicsVQA	I+T	T	24K	1K	1000
	ChartQA	I+T	T	28K	1K	1000
	Visual7W	I+T	T	70K	1K	1000
	ScienceQA	I+T	T	–	1K	1000
	VizWiz	I+T	T	–	1K	1000
	GQA	I+T	T	–	1K	1000
	TextVQA	I+T	T	–	1K	1000
Visual Grounding	MSCOCO	I+T	I	100K	1K	1000
	RefCOCO	I+T	I	–	1K	1000
	RefCOCO-Matching	I+T	I+T	–	1K	1000
	Visual7W-Pointing	I+T	I	–	1K	1000

Classification. The classification suite spans object, scene, news, and multimodal-content recognition. ImageNet-1K, VOC2007, and SUN397 evaluate object and scene recognition through image-to-label matching, while N24News and HatefulMemes require multimodal understanding. We additionally use Place365, ImageNet-A, ImageNet-R, ObjectNet, and Country-211 to evaluate generalization beyond the training distributions. These datasets cover distribution shifts arising from scene changes, natural adversarial examples, artistic renditions, unusual object configurations, and geographic content.

Visual question answering. The VQA benchmarks span diverse forms of multimodal understanding and reasoning. OK-VQA and A-OKVQA emphasize external and commonsense knowledge, DocVQA and InfographicsVQA focus on document and text-rich visual understanding, ChartQA requires chart interpretation and numerical reasoning, and Visual7W evaluates general visual question answering. For OOD evaluation, we additionally consider ScienceQA, VizWiz, GQA, and TextVQA, covering scientific reasoning, visually impaired user queries, compositional scene understanding, and scene-text reasoning. Following MMEB, each image-question pair is encoded as a query, while candidate answers are independently encoded as textual targets and ranked by embedding similarity.

Visual grounding. We further evaluate whether the learned embedding space transfers to localization-oriented matching tasks. MSCOCO and RefCOCO require matching an image-text query to a target image region, Visual7W-Pointing retrieves the region corresponding to a question, and RefCOCO-Matching compares multimodal region-expression pairs. These experiments provide an additional test of generalization beyond the primary classification and VQA settings.

C Implementation Details

C.1 Training configuration

We implement all models in PyTorch and train with bfloat16 precision on a single NVIDIA B200 GPU. We use B3-Qwen2-2B (Thirukovalluru et al., 2025), a Qwen2-VL-2B-Instruct-based multimodal embedding model, as the teacher, and distill it into FastVLM-0.5B (Vasu et al., 2025) and LLaVA-OneVision-0.5B (Li et al., 2024). Unless otherwise specified, all methods follow the same architecture-specific training setup for fair comparison.

For CRED, we set the random projection budget to $M = 256$, regularize layers at 50%–80% of the network depth, and use $\lambda_{\mathrm{CPR}} = 0.5$ for FastVLM-0.5B and 1.0 for LLaVA-OneVision-0.5B. The full training configuration is summarized in Table 7.

Table 7: Training configuration for the two student architectures.

Configuration	**FastVLM-0.5B**	**LLaVA-OneVision-0.5B**
Epochs	1	1
Batch size	16	16
Optimizer	AdamW	AdamW
Student learning rate	1×10^{-4}	1×10^{-4}
Projector learning rate	5×10^{-4}	5×10^{-4}
Learning-rate schedule	Constant	Constant
Warmup ratio	0.05	0.05
Weight decay	0.01	0.01
LoRA rank r	64	64
Image resolution	448×448	336×336
Precision	BF16	BF16
M	256	256
λ_{CPR}	0.5	1.0
Regularized layer depth	50%–80%	50%–80%

C.2 Detailed Implementation of CRED

We provide additional implementation details for CRED. Since the teacher f^T is frozen throughout training, we avoid redundant teacher forward passes by pre-computing and caching its output embeddings once, prior to training, and reusing this cache at every subsequent training step. Each training step then requires only a single forward pass through the student f^S: the resulting representations are used to compute the contrastive and distillation objectives against the cached teacher outputs, as well as the cross-layer participation regularizer $\mathcal{L}_{\mathrm{CPR}}$, whose matrix-free computation is detailed in Algorithm 1.

$\mathcal{L}_{\mathrm{struct}}$ pools both text- and image-token representations, while $\mathcal{L}_{\mathrm{CPR}}$ operates only on image tokens; in both cases, the relevant tokens must first be separated from the packed input sequence. For each sample i and layer l, the student's hidden state $H_i^{(l)}$ contains text and image tokens interleaved according to the input template; we recover the image-token block $H_{i,\mathrm{img}}^{(l)}$ (and, analogously, the text-token block) by gathering the corresponding token positions. This extraction is applied identically at every layer $l \in \{0, \ldots, L\}$, so that $H_{i,\mathrm{img}}^{(0)}$ and $H_{i,\mathrm{img}}^{(l)}$ refer to the same set of image tokens observed at different depths of the network.

Once extracted, the image tokens are treated differently by the two objectives. For $\mathcal{L}_{\mathrm{struct}}$, the final-layer image tokens $H_{i,\mathrm{img}}^{(L)}$ are reduced to a single vector per image by averaging over the token dimension and rescaling to unit norm, the same mean-pooling and L2-normalization applied to the text tokens: this collapses the token dimension and discards their internal spread, since $\mathcal{L}_{\mathrm{struct}}$ only needs one representative embedding per image (and, analogously, per text sequence). For $\mathcal{L}_{\mathrm{CPR}}$, in contrast, the image tokens are kept as a full $N_i^{\mathrm{img}} \times D$ matrix rather than pooled, because the participation ratio is precisely a measure of how variance is spread across this set of tokens; mean-

pooling before computing $\widehat{\mathrm{PR}}$ would collapse the token-level spectral structure that CPR is designed to regularize. Algorithm 1 therefore takes the unpooled matrices $H^{(0)}_{i,\mathrm{img}}$ and $H^{(l)}_{i,\mathrm{img}}$ as input.

Algorithm 1 Cross-layer participation regularization (CPR)

Require: $\{H^{(0)}_{i,\mathrm{img}}, H^{(l)}_{i,\mathrm{img}}\}_{l\in\mathcal{S}}$, number of random projections M, minimum valid tokens $\tau_{\min} = 4$
Ensure: $\mathcal{L}_{\mathrm{CPR}}$

$\mathcal{I} = \{i : N_i^{\mathrm{img}} \geq \tau_{\min}\}$
for $l \in \mathcal{S}$ **do**
  Draw $g_s \sim \mathcal{N}(0, I_D)$, $s = 1, \ldots, M$; set $A = [g_1/\|g_1\|_2, \ldots, g_M/\|g_M\|_2] \in \mathbb{R}^{D\times M}$
  for $i \in \mathcal{I}$ **do**
    for $r \in \{0, l\}$ **do**
      Center $H = H^{(r)}_{i,\mathrm{img}}$: $\widetilde{H} = H - \mathbf{1}\mu^\top$, $\mu = \frac{1}{N_i^{\mathrm{img}}} H^\top \mathbf{1}$
      $Y = \widetilde{H}A$
      $\hat{t}_1 = \dfrac{D}{M(N_i^{\mathrm{img}} - 1)} \|Y\|_F^2, \quad \hat{t}_2 = \dfrac{D}{M} \left\| \dfrac{\widetilde{H}^\top Y}{N_i^{\mathrm{img}} - 1} \right\|_F^2$
      $\widehat{\mathrm{PR}}_i^{(r)} = \hat{t}_1^2/\hat{t}_2$
    end for
  end for
end for
$\mathcal{L}_{\mathrm{CPR}} = \dfrac{1}{|\mathcal{S}||\mathcal{I}|} \displaystyle\sum_{l\in\mathcal{S}} \sum_{i\in\mathcal{I}} \max\big(0,\ \mathrm{sg}[\widehat{\mathrm{PR}}_i^{(0)}] - \widehat{\mathrm{PR}}_i^{(l)}\big)$
return $\mathcal{L}_{\mathrm{CPR}}$

D Additional ablations

D.1 Hyperparameter Sensitivity Analysis.

Table 8: Sensitivity to the CPR weight λ_{CPR}.

λ_{CPR}	IND	OOD	Avg.
0.0	70.06	38.12	54.09
0.5	**71.48**	40.70	**56.09**
1.0	70.64	**41.02**	55.83
2.0	71.00	39.38	55.19
3.0	70.70	40.12	55.41

Table 9: Sensitivity to the random projection budget M.

M	IND	OOD	Avg.
64	71.48	40.00	55.74
128	71.38	40.06	55.72
256	71.48	**40.70**	**56.09**
512	**71.72**	40.14	55.93
Exact PR	71.42	40.12	55.77

We further examine the sensitivity of CRED to the CPR weight λ_{CPR} and the number of random projections M. Tables 8 and 9 summarize the results under the B3-Qwen2-2B $\rightarrow$ FastVLM-0.5B setting.

CPR weight. Table 8 suggests that CPR benefits from a moderate strength rather than increasingly strong regularization. Among the tested non-zero weights, $\lambda_{\mathrm{CPR}} = 0.5$ gives the best overall average, whereas 1.0 favors OOD performance. Larger weights provide no further improvement, consistent with balancing dimensionality preservation against the semantic learning objectives. The average varies by less than one point across these settings, indicating that performance does not depend on a narrowly tuned non-zero weight.

Number of random projections. Performance is also stable with respect to the number of random projections M. Varying M from 64 to 512 changes the average by only 0.37%, with $M = 256$ giving the highest average of 56.09%, which we use as the default. Exact PR achieves a comparable

Table 10: Effect of the layer range used for cross-layer participation regularization.

Depth	IND	OOD	Avg.
$[50\%, 80\%]$	71.48	40.70	56.09
$[80\%, 90\%]$	71.68	40.04	55.86
$[80\%, 100\%]$	71.14	40.36	55.75

Table 11: Out-of-distribution CLS and VQA results in Precision@1 (%) for FastVLM-0.5B and LLaVA-OneVision-0.5B distilled from B3-Qwen2-2B. Bold and underline indicate the best and second-best student results within each backbone and evaluation column, respectively.

Method	OOD CLS						OOD VQA				
	Place365	IN-A	IN-R	ObjNet	Country	Avg.	SciQA	VizWiz	GQA	TextVQA	Avg.
B3-Qwen2-2B → FastVLM-0.5B											
Teacher	44.5	50.7	90.4	72.0	26.2	56.8	39.70	49.20	67.60	80.20	59.18
SFT	29.9 ±0.18	31.7 ±0.21	51.8 ±0.25	44.7 ±0.19	8.0 ±0.12	33.2	33.67 ±0.15	36.10 ±0.17	63.87 ±0.12	65.10 ±0.20	49.68
MSE	28.9 ±0.15	30.8 ±0.20	48.9 ±0.22	43.2 ±0.18	7.8 ±0.10	31.9	33.50 ±0.20	36.53 ±0.06	64.90 ±0.27	65.10 ±0.30	50.01
RKD	29.5 ±0.20	23.3 ±0.17	52.6 ±0.24	**49.4** ±0.21	7.2 ±0.11	32.4	33.43 ±0.40	36.70 ±0.27	62.90 ±0.30	65.30 ±0.20	49.58
CKD	30.4 ±0.19	29.8 ±0.18	48.6 ±0.20	47.1 ±0.23	8.3 ±0.13	32.8	35.37 ±0.25	36.60 ±0.27	62.70 ±0.10	65.43 ±0.12	50.03
EM-KD	30.5 ±0.17	22.3 ±0.16	58.4 ±0.28	46.5 ±0.22	7.1 ±0.11	33.0	34.97 ±0.21	36.37 ±0.25	63.80 ±0.30	64.70 ±0.27	49.96
HieRD	30.1 ±0.16	32.7 ±0.18	55.6 ±0.22	48.5 ±0.20	**8.6** ±0.14	35.1	38.90 ±0.61	41.37 ±0.12	63.33 ±1.27	68.87 ±0.68	53.12
CRED	**38.1** ±0.46	**35.0** ±0.35	**73.23** ±0.58	49.0 ±0.35	8.1 ±0.06	**40.7**	**40.07** ±0.23	**44.60** ±0.00	**69.17** ±0.06	**73.03** ±0.06	**56.72**
B3-Qwen2-2B → LLaVA-OneVision-0.5B											
Teacher	44.5	50.7	90.4	72.0	26.2	56.8	39.70	49.20	67.60	80.20	59.18
SFT	34.20 ±1.14	43.20 ±0.17	74.97 ±0.15	49.13 ±0.23	**11.10** ±0.17	42.52	25.37 ±0.06	34.17 ±0.25	54.00 ±0.40	45.47 ±0.12	39.75
MSE	35.30 ±0.20	43.93 ±0.12	76.33 ±0.29	48.07 ±0.12	9.87 ±0.06	42.70	26.43 ±0.38	33.63 ±0.15	53.80 ±0.35	45.37 ±0.06	39.81
RKD	34.63 ±0.31	43.93 ±0.23	74.40 ±0.26	46.03 ±0.12	10.23 ±0.15	41.85	27.67 ±0.21	34.27 ±0.15	52.87 ±0.12	45.37 ±0.38	40.04
CKD	34.07 ±0.25	43.40 ±0.00	74.73 ±0.38	48.33 ±0.15	10.93 ±0.21	42.29	27.97 ±0.25	34.73 ±0.12	53.33 ±0.21	45.80 ±0.35	40.46
EM-KD	34.70 ±0.17	34.43 ±0.23	65.90 ±0.35	42.13 ±0.21	8.37 ±0.06	37.11	25.70 ±0.10	31.03 ±0.23	56.87 ±0.21	41.70 ±0.30	38.83
HieRD	36.43 ±0.29	37.73 ±0.06	78.73 ±0.12	42.80 ±0.17	8.37 ±0.15	40.81	24.10 ±0.00	37.35 ±0.35	55.70 ±0.14	54.15 ±0.49	42.83
CRED	**39.77** ±0.12	**44.37** ±0.12	**82.00** ±0.00	**52.20** ±0.00	10.23 ±0.06	**45.71**	**32.95** ±0.10	**42.68** ±0.05	**68.40** ±0.00	**64.30** ±0.08	**52.08**

average of 55.77%, indicating that the randomized estimator preserves the effectiveness of exact computation while reducing its runtime (Table 3).

D.2 Effect of the Regularized Layer Range.

We compare three regularized layer ranges while keeping all other training hyperparameters fixed (Table 10). The default 50%–80% range achieves the best OOD performance and overall average. Regularizing 80%–90% slightly improves IND performance but reduces OOD performance, while extending the range to 80%–100% also yields a lower overall average, supporting our default choice.

E Additional Results

E.1 Out-of-distribution Results

Table 11 provides a detailed comparison on held-out CLS and VQA benchmarks. For FastVLM-0.5B, CRED achieves an OOD CLS average of 40.70%, improving over SFT by 7.50% and HieRD by 5.60%. The gain is particularly large on ImageNet-R, where CRED reaches 73.23%, compared with 51.80% for SFT and 58.40% for the strongest competing baseline. CRED also obtains the best results on Places365 and ImageNet-A, while RKD and HieRD remain slightly stronger on ObjectNet and Country211, respectively. On OOD VQA, CRED achieves the best result on all four benchmarks and improves the average from 49.68% with SFT and 53.12% with HieRD to 56.72%.

For LLaVA-OneVision-0.5B, CRED similarly improves the OOD CLS average to 45.71%, compared with 42.52% for SFT and 42.70% for the strongest competing baseline. It achieves the best student performance on four of the five classification benchmarks, with Country211 being the only

Table 12: Visual grounding results on MMEB. We report Precision@1 (P@1, %) on four grounding benchmarks using the B3-Qwen2-2B → FastVLM-0.5B setting. Bold indicates the best student performance.

Method	MSCOCO	RefCOCO	RefCOCO-Matching	Visual7W-Pointing	Average
Teacher	70.5	87.9	87.1	74.7	80.1
Student	72.1	82.8	**77.2**	82.3	78.6
RKD	72.9	82.9	76.7	81.3	78.4
CKD	73.1	79.0	74.6	82.3	77.2
EMO	72.1	81.7	76.5	80.5	77.7
EM-KD	70.6	80.5	75.4	79.9	76.6
HieRD	73.5	83.0	76.4	82.7	78.9
CRED (Ours)	**73.8**	**83.8**	77.0	**83.5**	**79.5**

exception. The gains are more pronounced on OOD VQA, where CRED reaches an average of 52.08%, exceeding SFT by 12.33% and HieRD by 9.25%. In particular, CRED substantially improves GQA and TextVQA to 68.40% and 64.30%, respectively. Overall, the results show that CRED transfers consistently across architectures and task families, while its largest improvements emerge under more challenging distribution shifts.

E.2 Visual Grounding Results

Table 12 evaluates CRED on visual grounding tasks requiring fine-grained correspondence between language and localized visual content, where it achieves the highest student average of 79.5%, exceeding HieRD and the non-distilled student by 0.6 and 0.9 points, respectively. CRED obtains the best student performance on MSCOCO, RefCOCO, and Visual7W-Pointing, indicating that its benefits extend beyond classification and VQA to region-level multimodal matching. On RefCOCO-Matching, CRED improves over all compared distillation methods but remains slightly below the non-distilled student (77.0% vs. 77.2%), showing that the gains are not uniform across all grounding benchmarks.

Table 13: Image classification results with a larger teacher. We report Precision@1 (P@1, %) on five MMEB classification benchmarks using the B3-Qwen2-7B → FastVLM-0.5B setting.

Method	ImageNet-1K	N24News	HatefulMemes	VOC2007	SUN397	Average
Teacher	84.4	81.3	64.3	89.9	82.6	80.5
Student	52.6	68.9	59.9	78.2	65.0	64.9
RKD	52.9	67.2	58.9	79.4	63.3	64.3
MSE	52.4	67.0	57.7	79.9	62.8	64.0
CKD	52.2	67.8	58.2	79.0	63.5	64.1
EMO	52.9	67.3	54.8	80.1	59.3	62.9
EM-KD	48.5	65.0	53.6	77.7	61.6	61.3
HieRD	55.2	71.7	61.5	80.6	66.0	67.0
CRED (Ours)	**62.2**	**74.6**	**61.9**	**86.5**	**72.3**	**71.5**

E.3 Scaling to a Larger Teacher

Table 13 shows that CRED achieves the best student performance on all five classification benchmarks with the B3-Qwen2-7B teacher, reaching an average of 71.5%, compared with 64.9% for the

non-distilled student and 67.0% for HieRD. The improvements over HieRD are particularly pronounced on ImageNet-1K (62.2% vs. 55.2%), VOC2007 (86.5% vs. 80.6%), and SUN397 (72.3% vs. 66.0%). These results indicate that CRED remains effective under a larger teacher-student capacity gap, extending its applicability beyond the default 2B teacher setting.