LEARNING FROM THE PAST: HISTORICAL REPRESENTATION REUSE WITH NEGATIVE FEEDBACK FOR VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models typically rely on cross-attention to capture interactions between visual and textual representations, yet intermediate representations generated during training are rarely reused across training units. Moreover, fluctuations in these representations may reduce the stability of cross-modal learning. To address these issues, we propose Historical Intermediate Representation Reuse and a Negative-Feedback Adapter. HIRR reuses the cross-attention Key and Value representations from the previous training unit, allowing historical intermediate information to participate in subsequent cross-modal interactions. The Negative-Feedback Adapter further regulates the current representation according to its deviation from a historical reference, suppressing unnecessary representation fluctuations. We analyze these mechanisms from the perspectives of statistical variance reduction and local perturbation propagation, linking historical representation reuse to effective sample complexity and feedback stability. Experimental results across multiple vision-language models and image-text retrieval benchmarks show that combining historical representation reuse with negative-feedback regulation consistently improves cross-modal learning performance and data efficiency, while demonstrating applicability across different cross-attention-based architectures.

1 INTRODUCTION

In recent years, vision-language models (VLMs) have achieved substantial progress in tasks such as image-text retrieval, visual question answering, and cross-modal understanding. Cross-attention is widely used to enable fine-grained interactions between visual and textual representations, with one modality providing queries and the other providing keys and values Li et al. (2025); Lim et al. (2024); Yao et al. (2024). Throughout this paper, we treat each mini-batch as one training unit Zhang et al. (2025). Under conventional mini-batch training, cross-modal intermediate representations are recomputed at every training unit and are typically discarded after the corresponding parameter update He et al. (2020); Zheng et al. (2026); Xin et al. (2026). Although successive mini-batches generally contain different samples, they are processed by the same continuously updated model under a common objective and related data distributions Tyagi & Sharma (2025); Xu et al. (2024); Zhao et al. (2024). Their intermediate representations may therefore share statistical and cross-modal structural information even without sample-level correspondence Wang et al. (2026); Pu et al. (2025); Duan et al. (2025). This raises a natural question: can previously computed cross-modal representations be reused to improve training-information utilization without substantially increasing model size or computational cost?

To address this question, we propose Historical Intermediate Representation Reuse (HIRR), a general mechanism for vision-language models equipped with cross-attention. HIRR reuses cross-attention intermediate representations produced by preceding training units and reintroduces them into the current cross-attention computation. Consequently, the model does not rely solely on the current input to reconstruct visual-textual relationships from scratch, but can also exploit cross-modal interaction information acquired during previous training units Liu et al. (2024); Huang et al. (2024). Unlike methods that store raw historical samples or construct a large external memory bank, HIRR maintains only lightweight historical intermediate states, enabling information transfer across

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.

training units with limited additional storage and computation Li et al. (2024); Maragheh & Deldjoo (2025).

Directly incorporating historical representations, however, may also introduce redundancy, noise, or historical bias that is inconsistent with the current input Ye et al. (2026); Möller & Padó (2025). To further regulate the influence of historical information on the current representation, we introduce a Negative-Feedback Adapter after the cross-attention output. The adapter constructs a feedback signal from the discrepancy between the current cross-attention output and a historical reference representation, and uses a lightweight transformation to regulate the current representation accordingly. Rather than simply adding historical outputs to the current representation, this mechanism generates an adaptive correction according to the deviation of the current state from its historical reference. It is therefore designed to suppress unnecessary representation fluctuations arising from factors such as mini-batch sampling, limited training data, and stochastic optimization Li et al. (2026); Cheng et al. (2026).

HIRR and the Negative-Feedback Adapter operate on two complementary stages of cross-attention. HIRR focuses on the reuse of historical information during cross-attention computation, allowing previously computed cross-modal intermediate representations to continue participating in subsequent visual-textual interactions. The Negative-Feedback Adapter, in contrast, uses a historical reference to regulate the resulting cross-attention output. The overall approach does not require modification of the primary training objective of the underlying model and does not rely on an additional data-replay procedure Zhao et al. (2026); Zhu et al. (2026). It can therefore be incorporated as an auxiliary mechanism into different forms of cross-attention-based VLMs Gheini et al. (2021); Han et al. (2024).

Beyond the architectural design, we further analyze the relationship between HIRR and training-data utilization from a statistical estimation perspective. With finite training samples and stochastic optimization, cross-attention intermediate representations can fluctuate around their underlying cross-modal semantic relationships. When a historical representation is appropriately correlated with the current learning objective, it can provide additional statistical information for estimating the current cross-modal representation, thereby reducing the random variance associated with estimation based solely on the current training unit. Based on this observation, we develop a variance-reduction analysis for HIRR and further examine the relationship between representation variance and statistical estimation error under finite-sample conditions. Under the corresponding statistical assumptions, a lower representation-estimation variance leads to a more stable estimate for a fixed amount of training data and may consequently reduce the effective number of samples required to achieve a given level of statistical estimation error.

For the Negative-Feedback Adapter, we further analyze its effect from the perspective of representation-disturbance propagation. Cross-attention outputs may exhibit additional variations across successive training units because of sampling and optimization stochasticity. Negative-feedback regulation based on a historical reference can locally compensate for such variations according to the deviation of the current output from the reference state. Under appropriate conditions on the feedback strength and module gain, this mechanism can limit the accumulation and amplification of disturbances across successive training units, thereby improving the stability of cross-modal representation updates. We subsequently derive the corresponding theoretical conditions and analyze how negative-feedback regulation affects the propagation of representation perturbations.

We evaluate the proposed method on multiple vision-language models with explicit cross-attention mechanisms and across several image-text retrieval datasets. The selected models employ different forms of cross-attention for visual-textual interaction, including cross-modal attention pooling and multimodal fusion encoders. We further investigate performance under different fractions of training data to assess the proposed method under limited-data conditions.

Our main contributions are summarized as follows:

1. We propose HIRR, which reuses historical cross-attention key/value representations across training units without storing raw samples or a large external memory.
2. We introduce a Negative-Feedback Adapter and provide theoretical analyses linking HIRR to variance reduction and data utilization, and feedback regulation to local perturbation attenuation.

3. We validate the method across multiple VLMs, datasets, and training-data regimes, with additional analyses of data utilization and representation variance.

2 METHOD

2.1 OVERVIEW

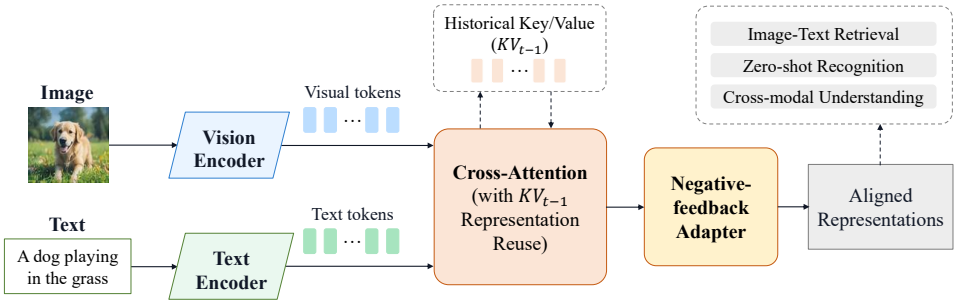

Figure 1: Overall concept of the proposed HIRR method. The framework contains two components: cross-attention representation reuse with historical key/value features and a negative-feedback adapter that refines the current multimodal representation.

We propose HIRR and a Negative-Feedback Adapter for vision–language models with cross-attention. As shown in Figure 1, our method consists of two complementary components. HIRR reuses key and value representations from the preceding training unit inside the current cross-attention computation, while the Negative-Feedback Adapter regulates the resulting cross-attention output using its deviation from a historical reference. Both mechanisms modify only the information flow around cross-attention and leave the original task objective unchanged.

Let X_t^q denote the modality representation used to generate queries at training unit t, and let X_t^c denote the representation providing cross-modal context. Standard cross-attention computes

$$Q_t = X_t^q W_Q, \qquad K_t = X_t^c W_K, \qquad V_t = X_t^c W_V, \tag{1}$$

followed by

$$Z_t = \mathrm{Softmax}\left(\frac{Q_t K_t^\top}{\sqrt{d_k}}\right) V_t, \tag{2}$$

where d_k is the key dimension. HIRR modifies the key/value representations used in Eq. 2, whereas the Negative-Feedback Adapter acts on the resulting cross-attention output.

2.2 HISTORICAL INTERMEDIATE REPRESENTATION REUSE

Let R_{t-1}^K and R_{t-1}^V denote the key and value representations stored from training unit $t-1$. Although consecutive training units generally contain different samples, HIRR does not assume sample-level correspondence between the current and historical representations. Instead, the historical key/value states are treated as a lightweight statistical reference produced by the same model under the same training objective and data distribution. During a local stage of training, such representations may preserve shared cross-modal structure despite stochastic variation caused by mini-batch sampling and optimization. HIRR therefore reuses the preceding states as an auxiliary representation prior rather than as semantically matched features of the current samples. HIRR forms history-enhanced representations as

$$\bar{K}_t = \frac{K_t + \alpha_K P_K\left(R_{t-1}^K\right)}{1+\alpha_K}, \qquad \bar{V}_t = \frac{V_t + \alpha_V P_V\left(R_{t-1}^V\right)}{1+\alpha_V}, \tag{3}$$

where $\alpha_K, \alpha_V \geq 0$ control the historical-reuse strengths, and $P_K(\cdot)$ and $P_V(\cdot)$ align the historical states with the current representation spaces. When the structures are identical, these mappings reduce to the identity.

The corresponding cross-attention output becomes

$$Z_t = \mathrm{Softmax}\left(\frac{Q_t \bar{K}_t^\top}{\sqrt{d_k}}\right)\bar{V}_t. \tag{4}$$

The normalized combination in Eq. 3 avoids systematically increasing the representation scale and permits the current and historical states to be interpreted as weighted estimates of an underlying cross-modal representation. Setting $\alpha_K = \alpha_V = 0$ recovers standard cross-attention.

After training unit t, the historical states are updated using the directly computed representations,

$$R_t^K = \mathrm{sg}(K_t), \qquad R_t^V = \mathrm{sg}(V_t), \tag{5}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. We store K_t and V_t rather than $\bar{K}_t$ and $\bar{V}_t$, preventing historical information from recursively accumulating across successive training units. HIRR therefore requires only the immediately preceding key/value states and does not store raw historical samples or a large external memory.

2.3 Negative-Feedback Adapter

Although historical representation reuse introduces information from preceding training units, the reused representation may not always be fully consistent with the current cross-modal state. We therefore introduce a Negative-Feedback Adapter (NFA) to regulate the current cross-attention output with respect to a historical reference.

Let Z_t denote the current cross-attention output and S_{t-1} the historical reference retained from the preceding training unit. We first define their representation discrepancy as

$$E_t = Z_t - S_{t-1}. \tag{6}$$

The discrepancy is processed by a lightweight bottleneck adapter,

$$A(E_t) = W_{\mathrm{up}}\,\sigma(W_{\mathrm{down}}E_t)\,, \tag{7}$$

where W_{down} and W_{up} denote the down and up projection layers, respectively, and $\sigma(\cdot)$ is a nonlinear activation function.

The regulated representation is then computed as

$$\widetilde{Z}_t = Z_t - \lambda A(E_t), \tag{8}$$

where $\lambda > 0$ controls the feedback strength. The adapter therefore learns a correction from the discrepancy between the current and historical representations, and the correction is subtracted from the current representation. When the learned correction is positively aligned with the current discrepancy, this update acts as negative feedback and attenuates the local representation deviation.

3 Theoretical Analysis of Feedback Representation Reuse

We analyze the proposed method from two complementary perspectives. First, we interpret HIRR as a weighted estimator that combines current and historical cross-attention representations, and derive conditions under which historical reuse reduces representation-estimation variance. We then connect this variance reduction to effective data utilization under a finite-sample model. Second, we analyze the Negative-Feedback Adapter through local perturbation propagation and establish conditions under which feedback regulation contracts the deviation of the current representation from its historical reference. Detailed proofs are provided in Appendix A.1.

3.1 Variance Reduction by Historical Representation Reuse

Consider an intermediate cross-attention representation H_t, corresponding to either K_t or V_t, and its historical counterpart R_{t-1}. During a local stage of training, we model them as

$$H_t = H^* + \varepsilon_t, \qquad R_{t-1} = H^* + \varepsilon_{t-1}, \tag{9}$$

where H^* denotes a locally shared latent cross-modal representation rather than the representation of an identical training sample. The model assumes that consecutive training units share a common statistical component while exhibiting unit-specific stochastic perturbations. We assume locally unbiased perturbations, $\mathbb{E}[\varepsilon_t] = \mathbb{E}[\varepsilon_{t-1}] = 0$, with

$$\mathbb{E}\|\varepsilon_t\|^2 = \sigma_t^2, \qquad \mathbb{E}\|\varepsilon_{t-1}\|^2 = \sigma_h^2, \qquad \rho = \frac{\mathbb{E}\langle \varepsilon_t, \varepsilon_{t-1} \rangle}{\sigma_t \sigma_h}. \tag{10}$$

The normalized HIRR representation is

$$\widehat{H}_t = \frac{H_t + \alpha R_{t-1}}{1+\alpha}, \qquad \alpha \geq 0. \tag{11}$$

Theorem 1 (Variance Reduction by Historical Intermediate Representation Reuse). Under the assumptions above,

$$\mathbb{E}\left\|\widehat{H}_t - H^*\right\|^2 = \frac{\sigma_t^2 + \alpha^2 \sigma_h^2 + 2\alpha\rho\sigma_t\sigma_h}{(1+\alpha)^2}. \tag{12}$$

For $\alpha > 0$, HIRR reduces the estimation error relative to the current representation whenever

$$\alpha\left(\sigma_h^2 - \sigma_t^2\right) < 2\left(\sigma_t^2 - \rho\sigma_t\sigma_h\right). \tag{13}$$

Hence, under Eq. 13,

$$\mathbb{E}\left\|\widehat{H}_t - H^*\right\|^2 < \sigma_t^2. \tag{14}$$

In the approximately equal-variance case, $\sigma_t^2 \approx \sigma_h^2 = \sigma^2$, any finite $\alpha > 0$ yields variance reduction whenever $\rho < 1$. Thus, the historical and current representations need not be independent; it is sufficient that their stochastic perturbations are not perfectly identical. The complete proof and the equal-variance derivation are given in Appendix A.2.

3.2 Implication for Effective Data Utilization

The variance result above also provides a statistical interpretation of data utilization. Define the HIRR variance ratio as

$$\kappa = \frac{\mathbb{E}\left\|\widehat{H}_t - H^*\right\|^2}{\sigma_t^2}. \tag{15}$$

When Theorem 1 holds, $0 < \kappa < 1$.

Consider a finite-sample regime in which the statistical estimation error satisfies

$$\mathcal{E}(n) \leq \frac{C\sigma_t^2}{n}, \tag{16}$$

where n denotes the effective number of training samples and $C > 0$ is a task- and model-dependent constant. Under HIRR, the corresponding bound becomes

$$\mathcal{E}_{\mathrm{HIRR}}(n) \leq \frac{C\kappa\sigma_t^2}{n}. \tag{17}$$

Therefore, for the same target estimation error, the sufficient effective sample requirement satisfies

$$n_{\mathrm{HIRR}}^{\mathrm{req}} = \kappa n_{\mathrm{base}}^{\mathrm{req}}, \qquad 0 < \kappa < 1. \tag{18}$$

This relation should be interpreted as an implication of the finite-sample statistical model rather than a guarantee of a fixed reduction in the amount of real training data. We therefore evaluate the predicted trend empirically using different training-data fractions. The detailed derivation is provided in Appendix A.3.

3.3 Local Perturbation Analysis of Negative Feedback

We next analyze when the Negative-Feedback Adapter attenuates the discrepancy between the current representation and its historical reference. Unlike a global contraction assumption over the entire feature space, our analysis focuses on the local perturbations encountered along the learned representation trajectory.

Let

$$E_t = Z_t - S_{t-1} \tag{19}$$

denote the discrepancy before feedback. According to Eq. equation 8, the discrepancy after feedback is

$$\widetilde{E}_t = \widetilde{Z}_t - S_{t-1} = E_t - \lambda A(E_t). \tag{20}$$

Theorem 2 (Local Perturbation Attenuation). Consider the feedback update in Eq. equation 20. For perturbations E in a local set Ω encountered along the learned representation trajectory, suppose that the adapter satisfies

$$\langle E, A(E)\rangle \geq \mu \|E\|_2^2, \qquad \mu > 0, \tag{21}$$

and

$$\|A(E)\|_2 \leq L_A \|E\|_2, \qquad L_A > 0. \tag{22}$$

If the feedback strength satisfies

$$0 < \lambda < \frac{2\mu}{L_A^2}, \tag{23}$$

then

$$\|\widetilde{E}\|_2 \leq \rho \|E\|_2, \qquad \rho = \sqrt{1 - 2\lambda\mu + \lambda^2 L_A^2} < 1. \tag{24}$$

Hence, under the local alignment condition, the feedback update attenuates the representation discrepancy with respect to the historical reference.

Theorem 2 gives a sufficient condition rather than a global guarantee on the adapter. In particular, the condition in Eq. equation 21 is imposed only on perturbations encountered along the learned representation trajectory and does not require $A(\cdot)$ to be coercive or injective over the entire ambient feature space. This distinction is important for the bottleneck adapter in Eq. equation 7, whose down-projection may contain directions that are not preserved by the bottleneck mapping. The attenuation behavior predicted by Theorem 2 is empirically examined in Appendix A.4 and Appendix A.6.

4 Experiments

4.1 Experimental Setup

We evaluate the proposed method on three vision–language models with explicit cross-attention mechanisms: β-CLIP, ALBEF, and BLIP. β-CLIP uses text-conditioned cross-attention to aggregate visual patch representations, whereas ALBEF and BLIP employ multimodal fusion encoders with cross-attention between visual and textual features Radford et al. (2021); Li et al. (2021; 2022). Because the exact cross-attention locations differ across backbones, the insertion points of the proposed modules are adapted to each architecture while following the same formulation in Section 2. Unless otherwise specified, HIRR denotes the complete method combining Historical Intermediate Representation Reuse and the Negative-Feedback Adapter.

Experiments are conducted on Flickr8k, Flickr30k, and MS COCO Plummer et al. (2015); Lin et al. (2014). For each backbone, the baseline and HIRR variants use identical data splits, preprocessing procedures, and primary training settings. We evaluate both image-to-text (I2T) and text-to-image (T2I) retrieval using Recall@1, Recall@5, and Recall@10. In repeated experiments, training random seeds are varied while the subset seed is fixed so that baseline and HIRR models use exactly the same training subsets. Unless otherwise specified, all quantitative results in the tables are reported in percentage form.

4.2 Main Retrieval Results

We first evaluate β-CLIP and β-CLIP-based HIRR on Flickr8k, Flickr30k, and MS COCO. Both variants use identical data splits and training settings. Table 1 reports the complete I2T and T2I retrieval results.

Table 1: Retrieval performance of β-CLIP and β-CLIP-based HIRR on three datasets.

Dataset	Method	I2T R@1	I2T R@5	I2T R@10	T2I R@1	T2I R@5	T2I R@10
Flickr8k	β-CLIP	77.5 ± 0.8	95.8 ± 0.2	98.7 ± 0.1	69.6 ± 0.1	91.8 ± 0.3	96.3 ± 0.2
	HIRR	80.2 ± 0.4	96.2 ± 0.2	99.0 ± 0.1	70.7 ± 0.1	93.1 ± 0.2	96.8 ± 0.1
Flickr30k	β-CLIP	87.7 ± 0.8	98.9 ± 0.1	99.6 ± 0.1	76.5 ± 0.1	94.3 ± 0.1	97.2 ± 0.1
	HIRR	89.2 ± 0.8	99.0 ± 0.1	99.6 ± 0.1	77.4 ± 0.3	94.9 ± 0.1	97.3 ± 0.1
MS COCO	β-CLIP	61.9 ± 0.2	85.5 ± 0.2	92.0 ± 0.1	48.6 ± 0.1	75.2 ± 0.1	84.0 ± 0.1
	HIRR	63.1 ± 0.2	86.3 ± 0.3	92.2 ± 0.1	49.2 ± 0.1	76.0 ± 0.1	84.3 ± 0.0

These experiments establish the overall retrieval effect of HIRR across datasets of different scales. The following analyses further examine the data-utilization effect and the representation-variance behavior associated with HIRR. To isolate the contribution of each component, we further conduct a component-wise ablation study on MS COCO using the β-CLIP backbone. The results in Appendix A.5 separately evaluate historical Key/Value reuse and the Negative-Feedback Adapter, showing their individual effects as well as their complementary contribution when combined.

4.3 Data Utilization and Theory-Oriented Analysis

We next study HIRR under limited-data conditions using MS COCO and β-CLIP. Its larger training set enables controlled construction of subsets at multiple data ratios.

4.3.1 Retrieval Performance under Different Training-Data Ratios

We train the baseline and HIRR models using 25%, 50%, 75%, 90%, and 100% of the full MS COCO training set. For each ratio, both methods use exactly the same training subset, while the validation and test sets remain unchanged. Table 2 reports all six retrieval metrics.

Table 2: Retrieval performance on MS COCO under different training-data ratios.

Data	Method	I2T R@1	I2T R@5	I2T R@10	T2I R@1	T2I R@5	T2I R@10
25%	β-CLIP	58.5 ± 0.2	83.5 ± 0.2	90.4 ± 0.2	46.2 ± 0.1	73.0 ± 0.1	82.2 ± 0.1
	HIRR	59.9 ± 0.3	83.8 ± 0.4	90.5 ± 0.1	46.9 ± 0.1	73.3 ± 0.1	82.2 ± 0.0
50%	β-CLIP	60.4 ± 0.2	84.8 ± 0.3	91.2 ± 0.2	47.4 ± 0.1	74.2 ± 0.1	83.1 ± 0.1
	HIRR	62.5 ± 0.4	85.0 ± 0.3	91.4 ± 0.2	47.6 ± 0.1	74.8 ± 0.1	83.0 ± 0.1
75%	β-CLIP	61.6 ± 0.3	85.4 ± 0.2	91.8 ± 0.2	48.0 ± 0.1	74.7 ± 0.1	83.6 ± 0.1
	HIRR	62.8 ± 0.3	85.9 ± 0.2	91.9 ± 0.2	48.2 ± 0.1	75.4 ± 0.1	83.7 ± 0.1
90%	β-CLIP	61.4 ± 0.2	85.4 ± 0.2	91.9 ± 0.1	48.4 ± 0.1	75.0 ± 0.1	84.0 ± 0.1
	HIRR	63.0 ± 0.1	86.0 ± 0.2	92.2 ± 0.2	48.7 ± 0.1	75.9 ± 0.2	84.3 ± 0.1
100%	β-CLIP	61.9 ± 0.2	85.5 ± 0.2	92.0 ± 0.1	48.6 ± 0.1	75.2 ± 0.1	84.0 ± 0.1
	HIRR	63.1 ± 0.2	86.3 ± 0.3	92.2 ± 0.1	49.2 ± 0.1	76.0 ± 0.1	84.3 ± 0.0

As shown in Table 2, this experiment answers two questions: HIRR improves retrieval performance at the same training-data ratio, and HIRR trained with less data can match or exceed the full-data baseline.

4.3.2 Empirical Effective Data Requirement

To summarize the data-utilization effect, we use the performance of β-CLIP trained on 100% of MS COCO as a reference. For metric m, we define the empirical equivalent data ratio as

$$r_m^* = \min\left\{r : M^m_{\text{HIRR}}(r) \geq M^m_{\beta\text{-CLIP}}(1.0)\right\}, \tag{25}$$

where $M^m(r)$ denotes the performance on metric m at training-data ratio r.

Table 3: Empirical data requirement of HIRR relative to full-data β-CLIP on MS COCO.

Metric	β-CLIP at 100%	Minimum HIRR Data Ratio	Relative Data Reduction
I2T R@1	61.9	50%	50%
I2T R@5	85.5	75%	25%
I2T R@10	92.0	90%	10%
T2I R@1	48.6	90%	10%
T2I R@5	75.2	75%	25%
T2I R@10	84.0	90%	10%

Because the evaluated ratios are discrete, Table 3 should be interpreted as an empirical data requirement rather than an exact continuous sample-complexity estimate. The results show that HIRR can reach the full-data β-CLIP performance with reduced training data, indicating improved data utilization efficiency. This provides an empirical counterpart to the effective-data interpretation.

4.3.3 Representation Variance Analysis

Theorem 1 and 2 predicts that HIRR can reduce stochastic representation-estimation variance under the stated conditions. We test this prediction by extracting key, value, and cross-attention output representations from a fixed set of MS COCO validation samples using independently trained models.

For sample i and run j, let $H_i^{(j)}$ denote the extracted representation and $\bar{H}_i$ its mean across M runs. We define the normalized empirical variance as

$$\mathcal{V}(H) = \frac{1}{MNd} \sum_{i=1}^{N} \sum_{j=1}^{M} \left\| H_i^{(j)} - \bar{H}_i \right\|_2^2, \tag{26}$$

where N is the number of fixed validation samples and d is the representation dimension.

Table 4: Empirical representation variance on MS COCO.

Training Data	Method	Var(K)	Var(V)	Var(Z)
25%	β-CLIP	0.0161	0.0198	0.0215
	HIRR	0.0134	0.0160	0.0188
50%	β-CLIP	0.0182	0.0201	0.0245
	HIRR	0.0147	0.0165	0.0196
100%	β-CLIP	0.0229	0.0257	0.0271
	HIRR	0.0221	0.0233	0.0252

Table 4 shows that HIRR consistently reduces the empirical variance of K, V, and Z across all training-data ratios, supporting the predicted representation-stabilization effect.

4.4 Generalization across Cross-Attention Models

Finally, we evaluate ALBEF and BLIP to test whether the proposed method generalizes beyond the cross-attention structure of β-CLIP. For each backbone, baseline and HIRR variants use identical data splits and training settings.

Table 5: Retrieval performance of ALBEF and ALBEF-based HIRR on three datasets.

Dataset	Method	I2T R@1	I2T R@5	I2T R@10	T2I R@1	T2I R@5	T2I R@10
Flickr8k	ALBEF	73.7 ± 0.6	92.9 ± 0.2	97.6 ± 0.1	63.8 ± 0.3	85.2 ± 0.5	90.4 ± 0.2
	HIRR	75.0 ± 0.5	93.7 ± 0.3	97.8 ± 0.1	64.4 ± 0.3	85.8 ± 0.4	91.2 ± 0.3
Flickr30k	ALBEF	84.0 ± 0.5	95.1 ± 0.1	98.1 ± 0.1	70.3 ± 0.3	92.8 ± 0.2	96.7 ± 0.1
	HIRR	85.2 ± 0.2	95.5 ± 0.2	98.2 ± 0.1	70.7 ± 0.3	93.2 ± 0.1	96.7 ± 0.1
MS COCO	ALBEF	60.6 ± 0.2	80.8 ± 0.1	91.4 ± 0.1	47.3 ± 0.2	73.6 ± 0.2	83.5 ± 0.1
	HIRR	62.1 ± 0.4	81.9 ± 0.3	91.8 ± 0.2	47.5 ± 0.1	74.3 ± 0.2	84.8 ± 0.1

The insertion locations are adapted to the corresponding cross-attention modules of ALBEF and BLIP, while the representation-reuse and negative-feedback principles remain unchanged. As shown

Table 6: Retrieval performance of BLIP and BLIP-based HIRR on three datasets.

Dataset	Method	I2T R@1	I2T R@5	I2T R@10	T2I R@1	T2I R@5	T2I R@10
Flickr8k	BLIP	75.9 ± 0.3	94.6 ± 0.3	97.6 ± 0.1	65.9 ± 0.3	90.1 ± 0.2	96.0 ± 0.2
	HIRR	77.1 ± 0.4	95.0 ± 0.3	97.9 ± 0.2	66.4 ± 0.1	90.6 ± 0.2	96.2 ± 0.1
Flickr30k	BLIP	85.8 ± 0.1	97.2 ± 0.2	99.3 ± 0.1	73.8 ± 0.3	93.5 ± 0.1	96.5 ± 0.1
	HIRR	86.7 ± 0.3	97.8 ± 0.3	99.4 ± 0.1	75.8 ± 0.2	94.1 ± 0.2	96.8 ± 0.1
MS COCO	BLIP	57.2 ± 0.3	80.5 ± 0.3	89.9 ± 0.1	47.0 ± 0.6	72.7 ± 0.3	83.6 ± 0.3
	HIRR	60.2 ± 0.2	81.1 ± 0.1	90.3 ± 0.1	49.9 ± 0.2	73.1 ± 0.2	84.3 ± 0.1

in Table 5 and Table 6, consistent improvements across the three backbones would indicate that HIRR is not specific to a single cross-attention implementation.

5 Conclusion

This paper proposes Historical Intermediate Representation Reuse and a Negative-Feedback Adapter for vision–language models equipped with cross-attention. HIRR reuses the cross-attention key/value representations generated by the preceding training unit, allowing previously acquired cross-modal interaction information to continue contributing to subsequent learning. The Negative-Feedback Adapter further regulates the current representation according to the discrepancy between the current output and its historical reference, thereby suppressing unnecessary representation fluctuations. We further analyze these mechanisms from the perspectives of statistical variance reduction and local perturbation propagation, establishing theoretical connections among historical representation reuse, effective sample requirements, and feedback regulation.

Experiments on Flickr8k, Flickr30k, and MS COCO show that HIRR consistently improves the image–text retrieval performance of β-CLIP and exhibits more efficient data utilization under limited-data conditions. Further experiments on ALBEF and BLIP demonstrate consistent performance improvements, indicating that the proposed method can be applied to different forms of cross-attention-based VLMs. Overall, the results suggest that reusing historical cross-attention representations across training units, together with negative-feedback regulation, provides an effective approach to improving the efficiency of cross-modal representation learning.

AI use statement

In this work, we used generative AI tools solely for language editing to improve the grammar, wording, and readability of the manuscript. Generative AI tools were not used to generate other content. All AI-assisted edits were carefully reviewed and verified by the authors to ensure that they did not alter the technical meaning or scientific claims of the manuscript. We take full responsibility for the final content of this work.

Reproducibility statement

The source code used in this work is provided in the supplementary materials to facilitate reproduction of the reported results.

References

Zhiyong Cheng, Yike Jin, Zhijie Zhang, Huilin Chen, Zhangling Duan, and Meng Wang. Modeling stage-wise evolution of user interests for news recommendation. In *Proceedings of the ACM Web Conference 2026*, pp. 6755–6764, 2026.

Siyuan Duan, Yuan Sun, Dezhong Peng, Zheng Liu, Xiaomin Song, and Peng Hu. Fuzzy multimodal learning for trusted cross-modal retrieval. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20747–20756. IEEE, 2025.

Mozhdeh Gheini, Xiang Ren, and Jonathan May. Cross-attention is all you need: Adapting pre-trained transformers for machine translation. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 1754–1765, 2021.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9726–9735. IEEE, 2020.

Xiaoshuang Huang, Hongxiang Li, Meng Cao, Long Chen, Chenyu You, and Dong An. Cross-modal conditioned reconstruction for language-guided medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(4):1821–1835, 2024.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Kun Li, Xin Jing, and Chengang Jing. Vector storage based long-term memory research on llm. *International Journal of Advanced Network, Monitoring and Controls*, 9(3):69–79, 2024.

Xiping Li, Aier Yang, Jianghong Ma, Kangzhe Liu, Shanshan Feng, Haijun Zhang, and Yi Zhao. Cpgrec+: A balance-oriented framework for personalized video game recommendations. *ACM Transactions on Information Systems*, 44(3):1–44, 2026.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Qi Zhi Lim, Chin Poo Lee, Kian Ming Lim, and Ahmad Kamsani Samingan. Unirag: Unification, retrieval, and generation for multimodal question answering with pre-trained language models. *IEEE Access*, 12:71505–71519, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Bo Liu, Zexin Lu, and Yan Wang. Towards medical vision-language contrastive pre-training via study-oriented semantic exploration. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 4861–4870, 2024.

Reza Yousefi Maragheh and Yashar Deldjoo. The future is agentic: Definitions, perspectives, and open challenges of multi-agent recommender systems. *arXiv preprint arXiv:2507.02097*, 2025.

Lucas Möller and Sebastian Padó. Explaining neural news recommendation with attributions onto reading histories. *ACM Transactions on Intelligent Systems and Technology*, 16(1):1–25, 2025.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

Ruitao Pu, Yang Qin, Dezhong Peng, Xiaomin Song, and Huiming Zheng. Deep reversible consistency learning for cross-modal retrieval. *IEEE Transactions on Multimedia*, 27:4095–4106, 2025.

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.

Sahil Tyagi and Prateek Sharma. Omnilearn: A framework for distributed deep learning over heterogeneous clusters. *IEEE Transactions on Parallel and Distributed Systems*, 36(6):1253–1267, 2025.

Zening Wang, Haoyu Guo, Yongxiang Yao, Yongjun Zhang, Peihao Wu, and Yi Wan. Ml-mifd: Multi-level multimodal invariant feature descriptor. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 49:79–85, 2026.

Yi Xin, Jianjiang Yang, Siqi Luo, Yuntao Du, Qi Qin, Haoxing Chen, Kangrui Cen, Yangfan He, Bin Fu, Yuewen Cao, et al. Parameter-efficient fine-tuning for pre-trained vision models: A survey and benchmark. *International Journal of Computer Vision*, 134(6):304, 2026.

Mengying Xu, Linyin Luo, Hanjiang Lai, and Jian Yin. Category-level contrastive learning for unsupervised hashing in cross-modal retrieval: M. xu et al. *Data Science and Engineering*, 9(3): 251–263, 2024.

Tao Yao, Shouyong Peng, Lili Wang, Ying Li, and Yujuan Sun. Cross-modality interaction reasoning for enhancing vision-language pre-training in image-text retrieval: T. yao et al. *Applied Intelligence*, 54(23):12230–12245, 2024.

Kailun Ye, Xiangjie Kong, Yuchao Zhang, Xuan Wang, Linan Zhu, Jiaxin Du, Guojiang Shen, and Jianxin Li. Dual history enhancement with hybrid hypergraph-graph networks for temporal knowledge graph reasoning. In *Proceedings of the ACM Web Conference 2026*, pp. 3800–3810, 2026.

Zeyu Zhang, Tianqi Cheng, and Yuki Todo. Arxsa: A general negative feedback control theory in vision-language models. In *EMNLP (Findings)*, pp. 11100–11110, 2025.

Chu Zhao, Shengjie Dong, Yuhai Zhao, and Yuan Li. Neutroncache: an efficient cache-enhanced distributed graph neural network training system. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 3310–3319, 2024.

Chuangxin Zhao, Canran Xiao, Siyuan Ma, Mengyao Lyu, Yanbiao Ma, Jun Xia, Guiguang Ding, and Yang Liu. Attention-spectrum regularization for replay-free continual multimodal llms. *arXiv preprint arXiv:2606.23063*, 2026.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, volume 2026, pp. 35791–35820, 2026.

Wenjie Zhu, Yabin Zhang, Wenjun Zeng, and Lei Zhang. Mmooc: A comprehensive benchmark for out-of-context evaluation in multimodal large language models. *arXiv preprint arXiv:2607.27637*, 2026.

A Appendix

A.1 Detailed Theoretical Derivations

This appendix provides the detailed derivations for the theoretical results presented in Section 3.

A.2 Proof of Theorem 1

Recall the normalized HIRR representation

$$\widehat{H}_t = \frac{H_t + \alpha R_{t-1}}{1+\alpha}, \tag{27}$$

with

$$H_t = H^* + \varepsilon_t, \qquad R_{t-1} = H^* + \varepsilon_{t-1}. \tag{28}$$

Subtracting the underlying representation H^* gives

$$\widehat{H}_t - H^* = \frac{\varepsilon_t + \alpha \varepsilon_{t-1}}{1+\alpha}. \tag{29}$$

Taking the squared norm and expectation,

$$\mathbb{E}\left\|\widehat{H}_t - H^*\right\|^2 = \frac{1}{(1+\alpha)^2}\mathbb{E}\left\|\varepsilon_t + \alpha\varepsilon_{t-1}\right\|^2$$
$$= \frac{1}{(1+\alpha)^2}\left(\mathbb{E}\|\varepsilon_t\|^2 + \alpha^2\mathbb{E}\|\varepsilon_{t-1}\|^2 + 2\alpha\mathbb{E}\langle\varepsilon_t, \varepsilon_{t-1}\rangle\right). \tag{30}$$

Using

$$\mathbb{E}\|\varepsilon_t\|^2 = \sigma_t^2, \qquad \mathbb{E}\|\varepsilon_{t-1}\|^2 = \sigma_h^2, \tag{31}$$

and

$$\mathbb{E}\langle\varepsilon_t, \varepsilon_{t-1}\rangle = \rho\sigma_t\sigma_h, \tag{32}$$

we obtain

$$\mathbb{E}\left\|\widehat{H}_t - H^*\right\|^2 = \frac{\sigma_t^2 + \alpha^2\sigma_h^2 + 2\alpha\rho\sigma_t\sigma_h}{(1+\alpha)^2}, \tag{33}$$

which gives Eq. 12.

To determine when HIRR improves over the current-unit estimate, we require

$$\frac{\sigma_t^2 + \alpha^2\sigma_h^2 + 2\alpha\rho\sigma_t\sigma_h}{(1+\alpha)^2} < \sigma_t^2. \tag{34}$$

Multiplying both sides by $(1+\alpha)^2$ yields

$$\sigma_t^2 + \alpha^2\sigma_h^2 + 2\alpha\rho\sigma_t\sigma_h < \sigma_t^2 + 2\alpha\sigma_t^2 + \alpha^2\sigma_t^2. \tag{35}$$

For $\alpha > 0$, rearranging gives

$$\alpha\left(\sigma_h^2 - \sigma_t^2\right) < 2\left(\sigma_t^2 - \rho\sigma_t\sigma_h\right), \tag{36}$$

which is Eq. 13. Therefore,

$$\mathbb{E}\left\|\widehat{H}_t - H^*\right\|^2 < \sigma_t^2 \tag{37}$$

whenever this condition holds. □

Equal-variance case. A particularly interpretable case arises when successive training units have approximately equal perturbation scales,

$$\sigma_t^2 \approx \sigma_h^2 = \sigma^2. \tag{38}$$

Substituting Eq. 38 into Eq. 12 gives

$$\mathbb{E}\left\|\widehat{H}_t - H^*\right\|^2 = \sigma^2\frac{1 + \alpha^2 + 2\alpha\rho}{(1+\alpha)^2}. \tag{39}$$

For variance reduction, we require

$$1 + \alpha^2 + 2\alpha\rho < (1+\alpha)^2 = 1 + \alpha^2 + 2\alpha. \tag{40}$$

For any finite $\alpha > 0$, this reduces to

$$\rho < 1. \tag{41}$$

Thus, HIRR does not require the current and historical perturbations to be statistically independent. As long as the two training units preserve approximately the same underlying cross-modal representation while their stochastic perturbations are not perfectly identical, the historical representation can contribute additional estimation information. As $\rho \to 1$, the additional variance-reduction effect diminishes.

The result does not imply that arbitrary historical information is beneficial. If the historical representation substantially deviates from the current underlying semantics, or if its stochastic error becomes excessively large, Eq. 13 may no longer hold.

A.3 Derivation of the Effective Sample Requirement

Consider the finite-sample statistical-error model introduced in Section 3.2,

$$\mathcal{E}(n) \le \frac{C\sigma_t^2}{n}, \tag{42}$$

where $C > 0$ is a task- and model-dependent constant.

For a target estimation error $\epsilon > 0$, a sufficient baseline sample requirement is

$$n_{\text{base}}^{\text{req}} = \frac{C\sigma_t^2}{\epsilon}. \tag{43}$$

Define

$$\kappa = \frac{\mathbb{E}\left\|\widehat{H}_t - H^*\right\|^2}{\sigma_t^2}. \tag{44}$$

When the condition of Theorem 1 holds,

$$0 < \kappa < 1. \tag{45}$$

The corresponding HIRR statistical-error bound becomes

$$\mathcal{E}_{\text{HIRR}}(n) \le \frac{C\kappa\sigma_t^2}{n}. \tag{46}$$

Hence, a sufficient sample requirement for the same target error ϵ is

$$n_{\text{HIRR}}^{\text{req}} = \frac{C\kappa\sigma_t^2}{\epsilon} = \kappa n_{\text{base}}^{\text{req}}. \tag{47}$$

Since $0 < \kappa < 1$, the finite-sample model predicts a smaller effective sample requirement after variance reduction.

This relationship concerns effective statistical sample requirements under the assumed error model. It does not imply that the amount of real training data in a deep vision–language model can always be reduced by exactly the factor κ, because practical performance also depends on optimization dynamics, model capacity, data distribution, and task-specific objectives. The data-fraction experiments in Section 4 are therefore used to test whether the predicted trend is reflected in empirical retrieval performance.

A.4 Proof of Theorem 2

From Eq. equation 20, the discrepancy after feedback is

$$\widetilde{E} = E - \lambda A(E). \tag{48}$$

Taking the squared Euclidean norm gives

$$\begin{aligned} \|\widetilde{E}\|_2^2 &= \|E - \lambda A(E)\|_2^2 \\ &= \|E\|_2^2 - 2\lambda\langle E, A(E)\rangle + \lambda^2\|A(E)\|_2^2. \end{aligned} \tag{49}$$

For $E \in \Omega$, using Eqs. equation 21 and equation 22, we obtain

$$\begin{aligned} \|\widetilde{E}\|_2^2 &\le \|E\|_2^2 - 2\lambda\mu\|E\|_2^2 + \lambda^2 L_A^2\|E\|_2^2 \\ &= \left(1 - 2\lambda\mu + \lambda^2 L_A^2\right)\|E\|_2^2. \end{aligned} \tag{50}$$

Define

$$\rho^2 = 1 - 2\lambda\mu + \lambda^2 L_A^2. \tag{51}$$

Under

$$0 < \lambda < \frac{2\mu}{L_A^2}, \tag{52}$$

we have

$$\rho^2 < 1. \tag{53}$$

Therefore,

$$\|\widetilde{E}\|_2 \le \rho \|E\|_2, \qquad \rho < 1, \tag{54}$$

which proves that the feedback operation attenuates the local representation discrepancy under the stated conditions. □

Point-wise interpretation. For an individual nonzero perturbation E_t with $A(E_t) \neq 0$, the exact expansion in Eq. equation 49 also shows that attenuation holds whenever

$$\langle E_t, A(E_t) \rangle > 0 \tag{55}$$

and

$$0 < \lambda < \frac{2\langle E_t, A(E_t) \rangle}{\|A(E_t)\|_2^2}. \tag{56}$$

Thus, contraction depends on the learned alignment between the correction and the encountered discrepancy rather than on a global positivity property of the bottleneck adapter.

A.5 Component Ablation

We conduct a component-wise ablation study on MS COCO with the β-CLIP backbone to isolate the contributions of the two components of HIRR. Specifically, we compare the original β-CLIP baseline with historical Key/Value reuse only, the Negative-Feedback Adapter only, and the complete HIRR model containing both components. All variants follow the same training and evaluation settings as the main experiments.

As shown in Table 7, the two components are evaluated independently and jointly, allowing their respective contributions to be distinguished from the improvement of the complete model.

Table 7: Component ablation of HIRR on MS COCO using the β-CLIP backbone. All values are reported in percentage (%).

Historical KV	NFA	I2T R@1	I2T R@5	I2T R@10	T2I R@1	T2I R@5	T2I R@10
×	×	61.9 ± 0.2	85.5 ± 0.2	92.0 ± 0.1	48.6 ± 0.1	75.2 ± 0.1	84.0 ± 0.1
✓	×	62.3 ± 0.2	85.7 ± 0.1	92.1 ± 0.1	48.9 ± 0.1	75.6 ± 0.0	84.1 ± 0.1
×	✓	62.1 ± 0.2	85.5 ± 0.2	92.2 ± 0.1	49.0 ± 0.0	75.8 ± 0.1	84.1 ± 0.1
✓	✓	63.1 ± 0.2	86.3 ± 0.3	92.2 ± 0.1	49.2 ± 0.1	76.0 ± 0.1	84.3 ± 0.0

A.6 Negative-Feedback Perturbation Analysis

Theorem 2 provides a sufficient condition under which the Negative-Feedback Adapter attenuates the local discrepancy between the current representation and its historical reference. We empirically examine whether this attenuation behavior is observed on representations encountered during inference.

For each evaluated representation, we measure the discrepancy before and after feedback:

$$D_{\text{before}} = \|Z_t - S_{t-1}\|_2, \qquad D_{\text{after}} = \|\widetilde{Z}_t - S_{t-1}\|_2. \tag{57}$$

We further define the attenuation ratio

$$\gamma = \frac{D_{\text{after}}}{D_{\text{before}}}. \tag{58}$$

A value of $\gamma < 1$ indicates that the feedback operation reduces the local representation discrepancy, corresponding to the attenuation behavior described in Theorem 2.

The observed $\gamma < 1$ indicates that, for the representations encountered in the evaluated model, the learned feedback reduces the discrepancy from the historical reference. As shown in Table 8

Table 8: Perturbation attenuation of β-CLIP-based HIRR on MS COCO. D_{before} and D_{after} denote the representation discrepancies before and after negative-feedback regulation, respectively.

Method	D_{before}	D_{after}	γ
HIRR	27.1	24.4	90.0%

This empirical behavior is consistent with the local attenuation condition analyzed in Theorem 2. Importantly, the theoretical result is conditional and does not claim that the bottleneck adapter is contractive over the entire representation space.