CoV-Align: Efficient Fine-grained Cross-Modal Alignment with Cohesive Visual Semantics Priority

Supplementary Material

Table 5. Ablation study on loss weight coefficients $\{\lambda_g, \lambda_v, \lambda_c\}$ for global contrastive loss, visual contrastive loss, and spatial concentration loss respectively. Results on Flickr30K with ViT-Base-224 and BERT-Base.

λ_v	λ_g	λ_c	RSUM	Variant
Varying λ_c (fix $\lambda_g = \lambda_v = 1$)				
1	1	1	515.0	
1	1	3	515.8	
1	1	5	**516.0**	**Default**
1	1	7	515.4	
1	1	10	515.5	
Varying λ_g (fix $\lambda_v = 1, \lambda_c = 5$)				
0.5	1	5	515.9	
2	1	5	513.3	$\downarrow$ 2.7
Varying λ_v (fix $\lambda_g = 1, \lambda_c = 5$)				
1	0.5	5	515.0	
1	2	5	467.1	$\downarrow$ 48.9

A. Hyperparameter Ablation Experiments

Loss Weight Robustness Analysis. Table 5 demonstrates the robustness of CoV-Align to loss weight hyperparameters λ_g, λ_v, λ_c. (1) The spatial concentration loss weight λ_c exhibits strong stability. Varying it from 1 to 10 (with λ_g = λ_v = 1) results in minimal performance fluctuation (<0.2% variance). (2) The visual contrastive loss weight λ_v also shows robustness across [0.5, 2] (with λ_v = 1, λ_c = 5), yielding only a 2.7-point variation (513.3-515.9 RSUM). (3) For the global contrastive loss weight λ_g (with λ_v = 1, λ_c = 5), performance remains stable within [0.5, 1], while extreme values (λ_g = 2) significantly degrade results (467.1 RSUM, $\downarrow$48.9 points) due to the fixed learning rate not being suitable for higher loss weights. Overall, CoV-Align maintains consistent performance across a wide range of hyperparameter settings, demonstrating strong robustness and ease of deployment without extensive tuning.

Analysis of Region Number Settings. To investigate the optimal number of semantic regions, we conduct ablation studies on the Flickr30K dataset with different N_q values. As shown in Table 3, when N_q = 4, the model achieves suboptimal performance due to insufficient semantic granularity, limiting its ability to capture diverse visual concepts. The optimal performance is achieved at Nq = 8, which strikes an effective balance between semantic diversity and representation coherence. Further increasing to N_q = 12 or 16 causes performance degradation, as excessive region fragmentation leads to redundant features and increased difficulty in learning discriminative representations. These results demonstrate that moderate region partitioning (N_q = 8) effectively captures semantic diversity while maintaining computational efficiency. Therefore, we adopt Nq = 8 as the default configuration in our framework.

B. Visualization of Shared Projection Matrix

In this section, we provide detailed quantitative and qualitative analysis to validate the effectiveness of using a shared projection matrix W in Eq. 2. We compare it against the conventional approach that uses two separate projection matrices (W_q for queries and W_k for keys).

Quantitative Results. As shown in Table 6, the shared matrix achieves superior performance, validating that it effectively enforces consistency across attention distributions to generate more semantically coherent coarse-grained features.

Method	TR		IR	
	R@1	R@10	R@1	R@10
Impact of Shared Projection Matrix				
w/ Two Separate Matrices (W_q, W_k)	76.8	96.5	62.5	93.2
w/ Shared Matrix (W)	**78.5**	**96.9**	**63.3**	**93.7**

Table 6. Ablation studies for our CoV-Align on Flickr30K dataset. TR denotes Text-to-Image Retrieval, and IR denotes Image-to-Text Retrieval.

Qualitative Visualization. To provide intuitive understanding of why the shared projection matrix achieves superior performance, we visualize the attention distributions in Figure 5. For each example, we show the raw image (left), attention map with shared matrix W (middle), and attention map with separate matrices W_q, W_k (right), where red highlights indicate attended regions. As illustrated, the shared matrix produces spatially concentrated attention patterns that accurately focus on semantically relevant objects. For instance, in case (a) with the phrase "a blue stroller", the shared matrix precisely attends to the stroller region. In stark contrast, separate matrices generate highly scattered attention distributions that extensively cover irrelevant background regions (indicated by widespread red areas). This scattered pattern introduces substantial noise

and weakens the intra-semantic coherence within visual regions. These visualizations provide intuitive evidence that the shared projection matrix effectively enforces consistency across attention distributions, thereby facilitating more accurate fine-grained cross-modal alignment.

(a) a blue stroller

(b) shovels

(c) a shovel

Figure 5. Visualization comparison of attention distributions between shared projection matrix (W) and separate matrices (W_q, W_k) on three representative examples.