Quota-Calibrated Fine-Grained Alignment with Context-Aware Marginals for Text-based Person Retrieval

Supplementary Material

We present additional implementation details and analysis of our proposed method QC-Align in this supplementary material.

A. Training Efficiency Analysis

Method	Pool Size	Graph N (img)	Efficiency		Acc.
			Time (ms)	Slowdown	R@1
CLIP	None	192	235	1.00×	68.71
+QC-Align	None	192	498	2.12×	**70.89**
	2×2	48	274	1.17×	70.52
	4×4	12	256	**1.09**×	69.63

Table 1. Efficiency-accuracy trade-off on CUHK-PEDES. Average text length: M=28.3 tokens.

We evaluate the trade-off between training efficiency and performance of QC-Align on CUHK-PEDES using a CLIP baseline with a batch size of 64, an average text length of 28.3 tokens, and 50 Sinkhorn iterations. All timing measurements are performed on a single NVIDIA RTX 4090 GPU and report the average per-iteration cost, including forward propagation and loss computation. To reduce the computational graph scale in optimal transport, we apply average pooling to image patches, yielding three configurations: no pooling (N=192 patches), 2×2 pooling (N=48), and 4×4 pooling (N=12).

As shown in Table 1, QC-Align on the full patch grid increases training time from 235 ms/iter to 498 ms/iter (2.12× overhead) while improving Rank-1 accuracy from 68.71% to 70.89%, validating the benefits of quota-calibrated fine-grained optimal transport alignment. With 2×2 pooling, training overhead significantly drops to 274 ms/iter (1.17×) with only a marginal accuracy decrease to 70.52%, demonstrating that a favorable efficiency-accuracy trade-off can be achieved by reducing transport graph size while preserving basic spatial resolution. Further applying 4×4 pooling lowers overhead to 256 ms/iter (1.09×), yet Rank-1 accuracy drops to 69.63%. This degradation arises because excessively coarse patch granularity weakens the local discriminability required by QC-Align, preventing CAME from generating accurate quotas and hindering the transport plan from capturing fine-grained word-region correspondences. Notably, QC-Align introduces additional computation *only during training* and is completely removed at inference time, incurring zero test-time overhead.

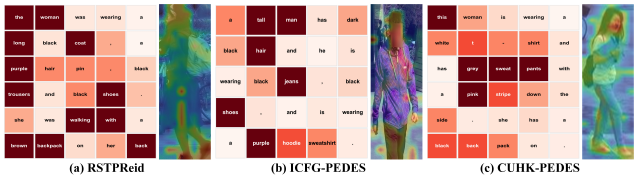

Figure 1. Visualization of learned marginal distributions (μ, ν) across three datasets. Left: text marginal ν as token heatmaps (darker indicates higher quota); Right: visual marginal μ overlaid on images.

B. Marginal Distribution Visualization

Figure 1 visualizes the learned marginal distributions (μ, ν) by QC-Align across three widely-used benchmarks. The text marginals ν (left matrices) consistently allocate substantially higher quotas to discriminative attribute tokens, including color terms (*e.g.*, "black", "purple", "white"), clothing types (*e.g.*, "coat", "jeans", "shirt"), and accessories (*e.g.*, "backpack", "stripe"). In contrast, function words and semantically weak tokens (*e.g.*, "a", "the", "is", "and") are systematically suppressed to near-background levels. This pattern of quota allocation effectively serves as an implicit semantic importance estimator and emerges consistently across all three datasets with distinct linguistic styles and domain characteristics, demonstrating that CAME exhibits robust cross-domain transferability in capturing semantic saliency without dataset-specific tuning.

From a cross-modal alignment perspective, the visual marginals μ (right heatmaps) spatially correspond to text-side semantic emphasis. Tokens with high ν values, such as "black coat", "purple hoodie", and "pink stripe", consistently trigger concentrated μ activations on corresponding body regions (*e.g.*, torso, upper body, garment details), indicating that quota-constrained optimal transport successfully guides many-to-many correspondences toward semantically meaningful alignments rather than uniform or noise-driven distributions. While occasional residual weights on non-discriminative tokens and sparse μ patterns in occluded regions suggest that quota estimation may still be affected by contextual ambiguity under extreme description redundancy or heavy occlusion, the overall visualization validates QC-Align's core claim: by dynamically allocating matching capacity through context-aware marginals, the framework explicitly enhances the interpretability and cross-domain consistency of local semantic alignment without requiring fine-grained annotations.