TempCLIP: Adaptive Temperature Control for Robust Multimodal Alignment

Anonymous Author 1
Institution removed for double-blind review
City, Country

Anonymous Author 2
Institution removed for double-blind review
City, Country

Anonymous Author 3
Institution removed for double-blind review
City, Country

Anonymous Author 4
Institution removed for double-blind review
City, Country

Anonymous Author 5
Institution removed for double-blind review
City, Country

Anonymous Author 6
Institution removed for double-blind review
City, Country

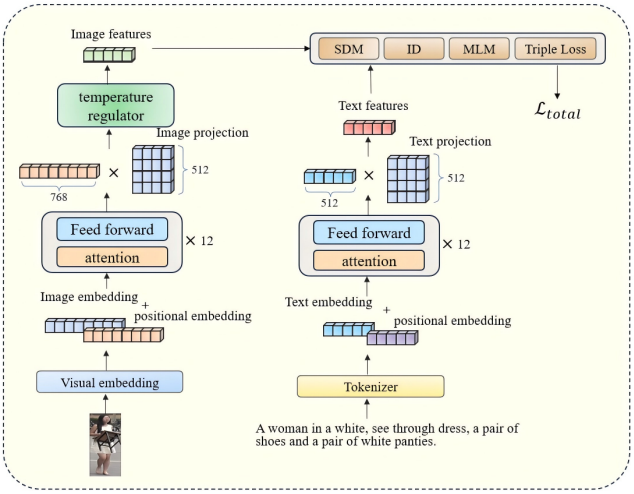

Figure 1: Overall Model Architecture

Abstract

Multimodal models such as CLIP align images and texts in a unified feature space, enabling cross-modal tasks like retrieval, captioning, and classification. Despite strong representation and zero-shot generalization, CLIP faces challenges in complex or few-shot scenarios, where occlusion, low light, or multiple objects reduce feature discrimination and semantic alignment. To address this, we introduce a learnable temperature controller in the image encoder to enhance feature separation, jointly optimize with ID, MLM, and SDM losses, and further propose a semantic similarity–weighted triplet loss to improve cross-modal understanding under challenging conditions.

Keywords

CLIP; multimodal learning; cross-modal alignment; temperature controller; few-shot learning; loss optimization

1 Introduction

Cross-modal understanding of images and text is a key research topic in computer vision and natural language processing. CLIP (Radford et al., ICML 2021)[1] projects images and texts into a shared feature space, enabling semantic alignment and supporting tasks such as retrieval, captioning, and classification. While CLIP demonstrates strong representation and zero-shot generalization, it remains limited in complex or few-shot scenarios due to insufficient fine-grained feature discrimination and suboptimal multi-loss optimization. To address these limitations, this study introduces a learnable temperature controller into the CLIP image encoder and employs a joint optimization framework combining ID loss, MLM loss, SDM loss, and a semantically weighted triplet loss[2]. The proposed method enhances feature separability and significantly improves cross-modal matching under challenging conditions, providing a robust enhancement to CLIP for vision–language understanding.

2 Related Work

Multimodal learning aims to bridge the semantic gap between images and text by learning a shared feature space, enabling tasks such as cross-modal retrieval, classification, and captioning. Early methods relied on CNNs for visual encoding and RNNs for textual modeling, optimizing cross-modal similarity with ranking or classification losses. With large-scale pretraining, Transformer-based architectures and contrastive objectives, exemplified by CLIP, have achieved strong zero-shot generalization and broad applicability. Nevertheless, CLIP's representations remain limited under complex or few-shot scenarios, where occlusion, low lighting, or multiple objects reduce feature discrimination. Recent studies introduce learnable temperature parameters in the Vision Transformer, allowing dynamic adjustment of feature smoothness and separability, which enhances cross-modal alignment and robustness. Loss design is also critical: ID Loss improves identity discriminability, MLM Loss strengthens textual representation, and triplet/contrastive losses optimize inter-sample distances but can struggle with semantically similar negatives. Multi-loss frameworks combining ID, MLM, Semantic Distance-preserving (SDM), and semantically weighted triplet losses have been proposed to preserve global alignment while capturing fine-grained local structures, improving retrieval accuracy and robustness. Our work builds upon these advances by integrating adaptive temperature control with a semantic similarity–weighted triplet loss to achieve more robust cross-modal feature alignment.

3 Method

3.1 Model Overview

To enhance the alignment and semantic representation of image and text features, this study proposes an improved multimodal representation learning framework based on the original CLIP model. The overall architecture consists of an image encoder, a text encoder, and a multi-loss optimization module, as illustrated in Figure 1.

The model comprises an image encoder and a text encoder. In the image branch, inputs are processed through visual embeddings, positional encoding, and a multi-layer Transformer. A learnable temperature controller is applied before feature projection to dynamically scale and normalize features, enhancing feature disentanglement and cross-modal alignment. In the text branch, tokenized inputs are embedded and passed through Transformer layers. Text features are then aligned with the temperature-adjusted image features in the projection space and jointly optimized using SDM Loss, ID Loss, MLM Loss, and a semantically weighted modified triplet loss.

3.2 Image Encoder Improvements

3.2.1 Image Encoder. In the original CLIP model, the image encoder primarily adopts a Vision Transformer (ViT) architecture. Its processing pipeline is as follows:

Patch Partitioning and Linear Projection Given an input image $I \in \mathbb{R}^{H\times W\times C}$, it is first divided into N non-overlapping fixed-size patches $\{x_p^i\}_{i=1}^N$, each of size $P \times P$. These patches are then mapped to vector representations through a linear projection:

$$X' = \text{Softmax}\left(\frac{Q_w K_w^\top + B_w}{\sqrt{d_h}}\right) V_w \tag{1}$$

Here, B_w denotes the relative positional bias, and d_h represents the dimension of each attention head. Subsequently, X' is further transformed through a multi-layer perceptron (MLP) with residual connections:

$$X' = X' + \text{MLP}(\text{LN}(X')) \tag{2}$$

Layer Normalization (LN) is applied to standardize the features, and the MLP typically consists of two fully connected layers with a nonlinear activation (GELU) in between:

$$\text{MLP}(X) = XW_1 + b_1 \xrightarrow{\text{GELU}} (XW_2 + b_2) \tag{3}$$

After passing through layers of the Transformer encoder, where each layer consists of Multi-Head Self-Attention (MHSA) and a Feed-Forward Network (FFN):

$$z_0^i = x_p^i W_e + b_e, \quad i = 1, 2, \ldots, N \tag{4}$$

3.2.2 Temperature Controller. In the original CLIP model, contrastive learning employs a fixed temperature parameter to scale the distribution of image–text feature similarities. However, a fixed temperature can cause features to become overly concentrated or excessively flattened in complex or few-shot scenarios, reducing the model's ability to distinguish similar pedestrians. To address this issue, this study introduces a learnable temperature controller as a core enhancement to the image encoder, as illustrated in Figure 2. The specific design is as follows: a learnable temperature parameter τ and a bias term b_τ are introduced into the feature outputs of the Vision Transformer, with a function *Softplus* applied to ensure non-negativity of the temperature, i.e.,

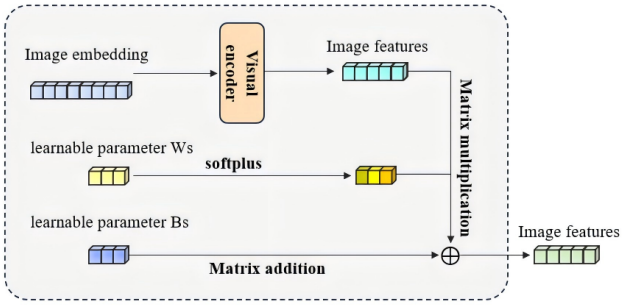

Figure 2: Temperature Controller

$$\tau_{\text{learned}} = \text{Softplus}(\tau), \quad \tau_{\text{learned}} > 0$$

$$X_{\text{total}} = X' \times \tau_{\text{learned}} + b_\tau$$

Here, b_τ denotes the bias value of the temperature controller. After passing through L layers of encoding, the output $X_{\text{cls}}^{(L)}$ of [CLS] token serves as the global feature representation of the image:

$$f_I = X_{\text{cls}}^{(L)} \in \mathbb{R}^D$$

When computing image–text similarity, the temperature parameter dynamically scales the dot product of feature vectors, thereby adaptively adjusting the "sharpness" of the similarity distribution. This dynamic mechanism offers several advantages: **Enhanced feature disentanglement:** Features of similar pedestrians are more effectively separated in the feature space[3], reducing false positives and false negatives. **Adaptation to few-shot learning:** With limited data, the model can automatically adjust the temperature to optimize the similarity distribution, improving fine-grained feature learning. **Improved robustness in complex environments:** The mechanism enhances adaptability to variations in illumination, occlusion, and background interference, thereby increasing pedestrian detection and recognition accuracy.

3.3 Text Encoder

To ensure effective modeling of textual inputs, this study adopts a standardized text preprocessing and encoding pipeline. Given a text sequence $x = \{x_1, x_2, \ldots, x_n\}$, it is first tokenized into subword units to mitigate out-of-vocabulary issues (e.g., "A person is walking" → {A, person, is, walk, ing}). Each token is then mapped to a vector via a trainable embedding matrix E and augmented with positional encodings P to retain sequential information. Special tokens [CLS] and [SEP] are added at the beginning and end of the sequence, where [CLS] is used for global semantic aggregation and [SEP] serves as sequence boundary indicators.

$$H_0 = [h_{\text{CLS}}, E(x_1) + P_1, \ldots, E(x_n) + P_n, h_{\text{SEP}}]$$

The sequence H_0 is fed into a stack of L encoding layers for contextual modeling:

$$H_l = \text{TransformerLayer}_l(H_{l-1}), \quad l = 1, 2, \ldots, L$$

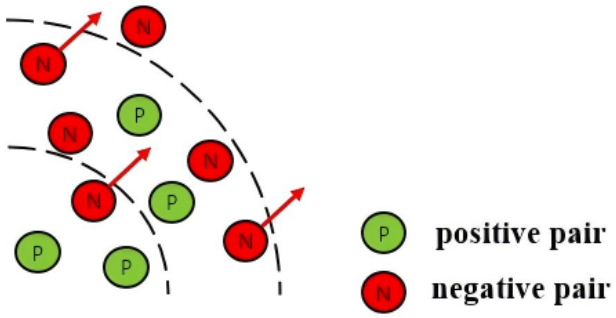

Figure 3: Illustration of the Modified Triplet Loss

Here, TransformerLayer$(\cdot)$ comprises multi-head self-attention and feed-forward networks. Finally, the global textual representation is obtained via the special token [CLS]:

$$h_{\text{global}} = H_L[\text{CLS}] \in \mathbb{R}^d$$

This representation captures both contextual dependencies and semantic discriminability, enabling alignment with the image modality in a shared semantic space and providing high-quality features for subsequent multi-loss optimization.

3.4 Loss Function Design

To enhance the discriminability and semantic consistency of cross-modal feature learning, this study employs a multi-loss framework during training, including Identity Discrimination Loss (ID Loss), Masked Language Modeling Loss (MLM Loss), Semantic Distance-preserving Loss (SDM Loss), and a modified Triplet Loss. ID Loss strengthens the model's ability to differentiate categories in image or text modalities, MLM Loss enhances textual semantic understanding[4], and SDM Loss preserves the semantic structure between samples within the cross-modal feature space.

3.4.1 Modified Triplet Loss. The traditional Triplet Loss is commonly used in cross-modal retrieval to reduce the distance between positive pairs while increasing the distance between negative pairs. However, it does not fully leverage semantic similarity information. In this study, we propose a semantically weighted modified triplet loss:

$$L_{\text{ASTLoss}} = \max\left(0,\ \alpha + \left\|z_a - z_p\right\|_2^2 - w_{\text{an}} \cdot \left\|z_a - z_n\right\|_2^2\right)$$

Here, z_a, z_p, and z_n denote the feature vectors of the anchor, positive, and negative samples, respectively; α is the margin parameter; and w_{an} represents the semantic similarity weight between the anchor and negative samples. Higher semantic similarity relaxes the negative sample constraint, while lower similarity strengthens it, enabling more rational cross-modal alignment, as illustrated in Figure 3.

3.4.2 Combined Loss Function. To fully exploit the complementary strengths of each loss term, this study adopts a weighted fusion strategy to form the final training objective:

$$L = \lambda_{\text{ID}} L_{\text{ID}} + \lambda_{\text{MLM}} L_{\text{IMLM}} + \lambda_{\text{SDM}} L_{\text{SDM}} + \lambda_{\text{AST Loss}} L_{\text{AST Loss}}$$

Here, λ_{ID}, λ_{MLM}, λ_{SDM}, and $\lambda_{\text{AST Loss}}$ are the weighting hyperparameters for each loss component.

Adjusting these values allows balancing the contribution of different optimization objectives to the overall model performance.

4 Experiments

4.1 Experimental Setup and Model Training

Experiments were conducted on Ubuntu with an Intel Xeon Gold 6262 CPU, 512 GB RAM, and an NVIDIA RTX 3090 GPU. The model was implemented in Python 3.10 with PyTorch 2.0, using the CUHK-PEDES dataset (~40,000 images, ~80,000 textual descriptions). Images were resized to 224×224 and normalized; texts were tokenized with *SimpleTokenizer* (fixed length 77). Data augmentation included random cropping and horizontal flipping. The backbone adopts CLIP ViT-B/16 for vision and a Transformer-based encoder for text, with a learnable temperature controller added to the final image layer (approximately 195 M parameters). Training used batch size 64, Adam optimizer (momentum 0.9, weight decay 4×10^{-5}), initial learning rate 1×10^{-5}, cosine annealing, linear warmup for the first 5 epochs, and a maximum of 120 epochs with learning rate reduced by a factor of 0.1 at epochs 20 and 50. The training objective combines SDM, MLM, ID Loss, and the modified triplet loss in a joint optimization scheme.

4.2 Comparison Methods and Evaluation Metrics

To validate the effectiveness of the proposed approach, multiple comparative experiments were conducted. The baseline model is the original CLIP (ViT-B/32), while the IRRA model is included as a representative method employing a temperature control mechanism. Finally, the proposed improved model (Ours, CLIP + Temperature Controller) was evaluated with a learnable temperature controller in the final layer of the image encoder. Evaluation metrics include Recall@K (K=1, 5, 10), Top-K Accuracy, and mean Average Precision (mAP) to assess retrieval accuracy and recall in text-to-image tasks. Additionally, mean Inverse Negative Penalty (mINP) was adopted as an auxiliary metric to evaluate robustness on long-tail samples.

Table 1: Comparison of state-of-the-art methods on CUHK-PEDES.

Methods	Type	Ref	ImageEnc.	TextEnc.	Rank-1	Rank-5	Rank-10	mAP	mINP
ISANet [5]	L	arXiv22	RN50	LSTM	63.92	82.15	87.69	-	-
LBUL [6]	L	MM22	RN50	BERT	64.04	82.66	87.22	-	-
SAF [7]	L	ICASSP22	ViT-Base	BERT	64.13	82.62	88.42	-	-
TIPCB [8]	L	Neuro22	RN50	BERT	64.26	83.19	89.12	-	-
CAIBC [9]	L	MM22	RN50	BERT	64.43	82.87	88.37	-	-
AXM-Net [10]	L	MM22	RN50	BERT	64.44	80.52	86.77	58.73	-
LGUR [11]	L	MM22	DeiT-Small	BERT	65.25	83.12	89.01	-	-
IVT [12]	G	ECCVW22	ViT-Base	BERT	65.59	83.11	89.21	-	-
CFine [13]	L	arXiv22	CLIP-ViT	BERT	69.57	85.93	91.15	-	-
IRRA	G	CVPR23	CLIP-ViT	CLIP-Xformer	72.385	87.995	93.064	65.277	49.963
OURS	G	–	CLIP-ViT	CLIP-Xformer	73.002	88.905	93.34	67.129	52.854

4.3 Experimental Results and Analysis

4.3.1 Comparison Experiments. From the table, early ResNet50 + LSTM/BERT approaches (e.g., CMPM/C, ViTAA, NAFS, DSSL)

Figure 4: First-Rank Visualization

generally achieve Rank-1 in the 50Methods leveraging CLIP pretraining (e.g., Han et al., CFine, IRRA) further improve performance. For instance, CFine using CLIP-ViT + BERT achieves Rank-1=69.57 and Rank-10=91.15, demonstrating the inherent advantage of CLIP in cross-modal retrieval tasks. As a strong baseline, IRRA achieves Rank-1/5/10 of 72.385/87.995/93.064, with mAP=65.277 and mINP=49.963. Building upon this, the proposed OURS method incorporates a multimodal triplet loss and a learnable temperature controller, resulting in Rank-1=73.002, Rank-5/10=88.905/93.34, mAP=67.129 (+2.84The improvements indicate that optimizing the feature space with a triplet loss and adaptively adjusting similarity distributions via the temperature controller effectively mitigates the interference of sparse and hard samples, enhancing generalization and robustness in complex retrieval scenarioscitezhu2021. To provide a more intuitive view of retrieval performance, the First-Rank Visualization is presented in Figure 4.

Table 2: Ablation study results on CUHK-PEDES.

No.	Methods	AST Loss	TC	Rank-1	Rank-5	Rank-10	mAP	mINP
0	Baseline	×	×	72.385	87.995	93.064	65.277	49.963
1	+AST Loss	✓	×	72.823	88.791	93.275	66.687	52.213
2	+TC	×	✓	72.953	88.645	93.291	66.918	52.538
3	OURS	✓	✓	73.002	88.905	93.340	67.129	52.854

4.3.2 Ablation Experiments. The ablation study results on the CUHK-PEDES dataset are presented in Table 2. The baseline model achieves Rank-1, Rank-5, and Rank-10 of 72.385%, 87.995%, and 93.064%, with mAP of 65.277% and mINP of 49.963%, serving as the reference. Incorporating AST Loss improves Rank-1 to 72.823%, and increases mAP and mINP to 66.687% and 52.213%, indicating that AST Loss enhances feature discriminability. When only the Temperature Controller (TC) is applied, Rank-1 reaches 72.953%, with mAP and mINP of 66.918% and 52.538%, demonstrating that TC optimizes the feature distribution. Combining AST Loss and TC (OURS) yields the best performance across all metrics: Rank-1 = 73.002%, Rank-5/Rank-10 = 88.905%/93.340%, and mAP/mINP = 67.129%/52.854%, showing significant improvements over the baseline. These results indicate that AST Loss enhances discriminability, TC refines the feature distribution, and their combination produces a synergistic effect that substantially improves retrieval performance and robustness.

5 Discussion and Conclusion

The experimental results demonstrate the effectiveness of the proposed method[15]. The triplet loss enhances feature discriminability and retrieval ranking quality by enlarging the distance between semantically negative samples, while the Temperature Controller further improves adaptability by dynamically scaling and shifting Transformer features, optimizing similarity distribution. Although Rank-1 and mAP do not always peak simultaneously, the combined approach achieves superior performance across all metrics, confirming its robustness and effectiveness in complex semantic retrieval tasks.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021. [Online]. Available: https://arxiv.org/abs/2103.00020
[2] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," in *Proc. European Conf. Computer Vision (ECCV)*, pp. 686–701, 2018.
[3] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Adversarial representation learning for text-to-image matching," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, pp. 5814–5824, 2019.
[4] Z. Wang, Z. Fang, J. Wang, and Y. Yang, "ViTAA: Visual-textual attributes alignment in person search by natural language," in *European Conf. Computer Vision (ECCV)*, pp. 402–420, Springer, 2020.
[5] S. Yan, H. Tang, L. Zhang, and J. Tang, "Image-specific information suppression and implicit local alignment for text-based person search," *arXiv preprint arXiv:2208.14365*, 2022.
[6] Z. Wang, A. Zhu, J. Xue, X. Wan, C. Liu, T. Wang, and Y. Li, "Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 1984–1992.
[7] S. Li, M. Cao, and M. Zhang, "Learning semantic aligned feature representation for text-based person search," in *ICASSP 2022*, 2022, pp. 2724–2728.
[8] Y. Chen, G. Zhang, Y. Lu, Z. Wang, and Y. Zheng, "TIPCB: A simple but effective part-based convolutional baseline for text-based person search," *Neurocomputing*, vol. 494, pp. 171–181, 2022.
[9] Z. Wang, A. Zhu, J. Xue, X. Wan, C. Liu, T. Wang, and Y. Li, "CAIBC: Capturing all-round information beyond color for text-based person retrieval," *arXiv preprint arXiv:2209.05773*, 2022.
[10] A. Farooq, M. Awais, J. Kittler, and S. S. Khalid, "AXM-Net: Implicit cross-modal feature alignment for person re-identification," *Pattern Recognition*, vol. 36, no. 4, pp. 4477–4485, 2022.
[11] Z. Shao, X. Zhang, M. Fang, Z. Lin, J. Wang, and C. Ding, "Learning granularity-unified representations for text-to-image person re-identification," *arXiv preprint arXiv:2207.07802*, 2022.
[12] X. Shu, W. Wen, H. Wu, K. Chen, Y. Song, R. Qiao, B. Ren, and X. Wang, "See finer, see more: Implicit modality alignment for text-based person retrieval," *arXiv preprint arXiv:2208.08608*, 2022.
[13] S. Yan, N. Dong, L. Zhang, and J. Tang, "CLIP-driven fine-grained text-image person re-identification," *arXiv preprint arXiv:2210.10276*, 2022.
[14] C. Gao, G. Cai, X. Jiang, F. Zheng, J. Zhang, Y. Gong, P. Peng, X. Guo, and X. Sun, "Contextual non-local alignment over full-scale representation for text-based person search," *arXiv preprint arXiv:2101.03036*, 2021.
[15] A. Zhu, Z. Wang, Y. Li, X. Wan, J. Jin, T. Wang, F. Hu, and G. Hua, "DSSL: Deep surroundings-person separation learning for text-based person retrieval," in *Proc. 29th ACM Int. Conf. Multimedia (MM)*, pp. 209–217, 2021.