Joint Global-Local Feature Alignment With Fine-Tuned Pretrained Transformers for Text-Based Person Search

Thi Thanh Thuy Pham[1], Huong-Giang Doan[2,*]
[1]Faculty of Cybersecurity and High Tech Crime Prevention - Academy of People Security, No. 125, Tran Phu, Van Quan, Ha Noi, Vietnam
[2]Faculty of Control and Automation - Electric Power University, No. 235, Hoang Quoc Viet, Nghia Do, Ha Noi, Vietnam
E-mail: thanh-thuy.pham@mica.edu.vn, giangdth@epu.edu.vn
*Corresponding author

Keywords: Text-based person search, transformer, contrastive image-text pretraining, global-local feature alignment

Received: January 2, 2025

Text-based person search (TBPS) aims to retrieve person images from a database using natural language description. Although significant progress has been made, TBPS remains challenging due to the complexities of cross-modal understanding. This work proposes a novel framework named ***GLAlign*** *that jointly aligns global and local features from both vision and text modalities using large-scale, pretrained, fine-tuned transformers. Specifically, we utilize ViT-B/32 for visual encoding and GPT-2 (English) or PhoBERT (Vietnamese) for textual encoding. To enhance alignment, we perform human part parsing and noun phrase extraction, enabling fine-grained local feature correspondence between body regions and descriptive attributes. The proposed method is evaluated on four benchmark datasets: CUHK-PEDES, CUHK-PEDES-VN, 3000VnPerson-Search, and 3000Vn-V2E. In the CUHK-PEDES dataset, our model achieves a Rank-1 accuracy of 80.75%, outperforming state-of-the-art methods such as PLOT (75.28%) and RaSa (76.51%). On the 3000VnPerson-Search dataset, our model reaches a Rank-1 accuracy of 85.72% for Vietnamese descriptions, indicating its robustness across both high-resource and low-resource languages. These results demonstrate the effectiveness of combining global-local alignment with fine-tuned vision language transformers for the TBPS task. The source codes are available at: https://github.com/TextBasedPersonSearch/PersionSearch*

Povzetek: Predstavljeno je TBPS ogrodje, ki združuje globalno in lokalno poravnavo slik in besedil z velikimi predtreniranimi transformatorji. Novost je v ujemanju delov telesa z NP frazami, kar omogoča bolj kvalitetno iskanje kot metoda ViTAA.

1 Introduction

Text-based person search (TBPS) is the task of retrieving person images from a large-scale gallery based on free-form natural language descriptions. This problem arises in practical scenarios such as security surveillance and social media moderation, where users or analysts may describe a target individual verbally rather than having access to an exemplar image. TBPS plays a critical role in enabling human-centered retrieval systems under such real-world constraints.

Compared to conventional image-based person re-identification, TBPS poses additional challenges due to the modality gap between textual and visual inputs. Natural language descriptions are often ambiguous, context-dependent, or linguistically diverse, and visual appearances can vary significantly across different camera views or environments. Effective retrieval thus requires fine-grained semantic reasoning that captures both global contextual meaning and local attribute-level correspondences.

Several deep neural networks (DNNs) have been proposed for the TBPS problem. For visual feature extraction, common models are CNN-based networks, such as VGG16 [1], ResNet50 [2], ResNet101 [3]. For extracting textual features, RNN-based networks such as LSTM or BiLSTM [4], [5] are popularly utilized. Recent advances in multimodal learning have led to powerful vision language models such as CLIP [6] and BLIP [7], which enable strong global alignment between images and texts through large-scale pre-training. However, these models often overlook fine-grained visual-textual associations - for example, matching detailed phrases like *'red sneakers'* with corresponding regions in the image. Moreover, the performance of such models under multilingual conditions, particularly in low-resource languages such as Vietnamese, remains underexplored.

This work is motivated by the need to enhance both the accuracy and interpretability of TBPS systems by integrating global and local features, and by leveraging pretrained transformer models in multilingual settings. To that end, we pose the following research questions:

– **RQ1**: Can large-scale pretrained transformer models

be effectively fine-tuned for joint global and local feature alignment in TBPS?

- **RQ2**: How does the incorporation of noun phrase–to–body-part alignment affect retrieval performance?
- **RQ3**: How does the use of multilingual textual descriptions (e.g., Vietnamese vs. English) influence retrieval accuracy, and what are the implications of using machine-translated versus native corpora?

To address these questions, we propose a unified framework named **GLAlign** that fine-tunes pretrained transformers for global-local alignment in TBPS. The image features are extracted using ViT-B/32, while the text is encoded using GPT-2 for English and PhoBERT for Vietnamese. Human parsing is used to segment person images into semantic body regions (e.g., head, upper body, lower body), which are then associated with corresponding noun phrases identified in the textual descriptions.

We conduct experiments on four datasets, including both English and Vietnamese corpora: CUHK-PEDES, CUHK-PEDES-VN, 3000VnPerson-Search, and 3000Vn-V2E. Results show that our model achieves state-of-the-art performance across multiple settings, and is robust to both cross-lingual and cross-dataset variations.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 presents the proposed framework. Section 4 details the experiments and results. Section 5 provides in-depth discussion, and Section 6 concludes the paper.

2 Related work

Text-based person search (TBPS) is a cross-modal retrieval task that involves retrieving images of a target person from a gallery based on a natural language description. Existing approaches to TBPS can be broadly grouped into three categories: (i) global visual-textual alignment, (ii) global-local feature alignment, and (iii) zero-shot or fine-tuned retrieval using large-scale pretrained vision-language models.

Global alignment

Early TBPS methods emphasized the global alignment between the overall image and the accompanying textual description. For example, GNA-RNN [8] employed VGG16 for visual encoding and LSTM for sentence representation, with joint embeddings learned through attention mechanisms. Zheng et al. [15] proposed the dual-path CNN, which uses two ResNet-50 networks, one for images and one for text, and introduces instance-level and ranking losses to enhance discriminability. However, these CNN-RNN-based methods are limited in their ability to capture long-range semantic dependencies and complex compositional language structures.

Global-local alignment

To enable more fine-grained semantic matching, subsequent studies focused on combining global and local features. Wang et al. [4] introduced the ViTAA (Visual-Textual Attributes Alignment) model, which associates image regions with attribute-level noun phrases using attention-based matching, improving both interpretability and retrieval accuracy. SCRF [9] proposed a correlation filtering mechanism to enhance word-to-region alignment. Although effective, these methods still rely heavily on convolutional and recurrent backbones, which limit their capacity for generalized multimodal reasoning. Traditional techniques in automatic text analysis [16] also emphasized the importance of structured semantic representations, which remain relevant for improving alignment in multimodal settings.

Pretrained vision-language models

Recent advances in large-scale multimodal pretraining have led to powerful transformer-based vision-language models, such as CLIP [6], BLIP [7], Florence-2 [17], and ALBEF [18]. These models learn global alignment from massive image-text corpora and can be directly adapted to TBPS in zero-shot or fine-tuned settings. More recently, IRRA [10] and RaSa [12] leveraged transformer-based backbones to achieve state-of-the-art performance on CUHK-PEDES.

Despite their strong global representations, these models generally lack explicit local alignment mechanisms. Most of them operate on full-image and full-sentence embeddings without decomposing into semantic parts or aligning textual attributes with specific body regions. Furthermore, there remains limited exploration of low-resource languages such as Vietnamese, where syntactic structures and cultural expressions can differ significantly from English, posing additional challenges to cross-lingual generalization [19]. Recent research in multilingual sentiment analysis and aspect-based opinion mining [20] highlights the necessity of semantic-aware modeling when handling language diversity.

Our contribution

To address these limitations, we propose a novel framework named **GLAlign** that performs joint global-local alignment using fine-tuned pre-trained transformers. We use ViT-B/32 [21] for visual encoding and GPT-2 [22] or PhoBERT [23] for textual encoding. Our method leverages human parsing and noun phrase extraction to associate specific body parts with descriptive phrases, enabling fine-grained matching. Furthermore, we evaluate our model not only on CUHK-PEDES (English), but also on Vietnamese datasets including CUHK-PEDES-VN and 3000VnPerson-Search.

Table 1 summarizes representative methods in terms of feature types, backbones, datasets, languages, and Rank-

Table 1: Comparison of TBPS methods in terms of feature types, backbones, datasets, and Rank-1 performance

Method	Feature Type	Backbone (Image/Text)	Dataset	Language	Rank-1 (%)
GNA-RNN [8]	Global	VGG16 / LSTM	CUHK-PEDES	English	19.71
ViTAA [4]	Global + Local	ResNet-50 / Bi-LSTM	CUHK-PEDES	English	55.97
SCRF [9]	Global + Local	ResNet-50 / BERT	CUHK-PEDES	English	64.04
IRRA [10]	Global	CLIP-ViT-B/16 / CLIP-Xformer	CUHK-PEDES	English	73.38
PLOT [11]	Global + Part Attention	CLIP-ViT-B/16 / CLIP-Xformer	CUHK-PEDES	English	75.28
RaSa [12]	Global + Relational Emb.	ALBEF / ALBEF	CUHK-PEDES	English	76.51
GLAlign (Ours)	**Global + Local**	**ViT-B/32 / GPT-2**	CUHK-PEDES	English	**80.75**
Method in [13]	Global + Local	ResNet-50 / Bi-LSTM	CUHK-PEDES-VN	Vietnamese	52.40
Method in [14]	Global	ResNet-50 / ResNet-50	3000VnPerson-Search	Vietnamese	33.07
GLAlign (Ours)	**Global + Local**	**ViT-B/32 / PhoBERT**	CUHK-PEDES-VN	Vietnamese	**77.72**
GLAlign (Ours)	**Global + Local**	**ViT-B/32 / PhoBERT**	3000VnPerson-Search	Vietnamese	**85.72**

1 performance. Our approach achieves superior results across both English and Vietnamese benchmarks, highlighting its effectiveness and generalizability across languages.

3 Proposed framework

The proposed framework of **GLAlign** method is shown in Fig. 1. The pre-trained transformer models are fine-tuned on image and text pairs of the TBPS datasets. The global and local features of the images and descriptions are extracted and aligned from the fine-tuned models.

3.1 Global feature embeddings

Global feature encoding is implemented for both the image and text. For the vision encoder, we closely follow the implementation of the ViT-B/32 architecture in [6] as a visual backbone network. This image encoder is derived from the original in [21] with minor modifications in the addition of an additional layer normalization to the combined patch and position embeddings before the transformer and uses a slightly different initialization scheme. The textual backbone encoder for English descriptions is the GPT-2, a pre-trained text transformer introduced in the OpenAI technical [22], and for Vietnamese descriptions it is the PhoBERT pre-trained model [23].

- **Global visual feature embedding:** The input image $\mathbf{I} \in \mathbb{R}^{H\times W\times C}$ represents an image with height H, width W, and C color channels (e.g. RGB), which is resized to 224×224 and then divided into a sequence of N patches, where $N = \frac{H}{P} \times \frac{W}{P}$ and P is the patch size (32×32 pixels). Each patch is flattened into a vector $\mathbf{p}_i$ of size $P^2 \times C$. Each flattened patch $\mathbf{p}_i \in \mathbb{R}^{P^2\times C}$ is then passed through a linear projection (a fully connected layer) to map it to a lower-dimensional embedding space. This yields a sequence of patch embeddings. Let $\mathbf{E}$ be the sequence of patch embeddings: $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, .., \mathbf{e}_N\}$, where each embedding $\mathbf{e}_i \in \mathbb{R}^D$ and D is the dimensionality of the embedding space. The linear projection can be written as Eq. (1) as follows:

$$\mathbf{e}_i = \mathbf{W}_p \cdot \mathbf{p}_i + \mathbf{b} \tag{1}$$

where $\mathbf{W}_p \in \mathbb{R}^{D\times(P^2\times C)}$, $\mathbf{b} \in \mathbb{R}^D$.

The next step is adding learnable positional encodings to the patch embeddings. This allows the model to understand the relative or absolute positions of the patches within the image. The positional encoding $\mathbf{PE}_i \in \mathbb{R}^D$ for each patch is added element-wise as shown in the Eq. (2) as follows:

$$\mathbf{e}_i^{\text{pos}} = \mathbf{e}_i + \mathbf{PE}_i \tag{2}$$

The positional encodings $\mathbf{PE}_i$ are learned or fixed sinusoidal functions that provide a unique position for each patch.

Before feeding the combined patch and positional embeddings into the transformer, we apply Layer Normalization (LN) on $\mathbf{e}_i^{\text{pos}}$ as follows:

$$\mathbf{e}_i^{\text{norm}} = \text{LN}\left(\mathbf{e}_i^{\text{pos}}\right) \tag{3}$$

A special [**CLS**] token is then added at the beginning of the patch embedding sequence $\mathbf{E}_{\text{norm}} = \{\mathbf{e}_1^{\text{norm}}, \mathbf{e}_2^{\text{norm}}, .., \mathbf{e}_N^{\text{norm}}\}$ resulting in the final transformer input:

$$\mathbf{X}_{\text{input}} = \text{concat}\left([\mathbf{CLS}\,], \mathbf{E}_{\text{norm}}\right) \tag{4}$$

where [CLS] token is a learnable parameter and serves as the representative feature of the entire image, aggregating information from all patches; $\mathbf{X}_{\text{input}} \in \mathbb{R}^{(N+1)\times D}$ is the final sequence of patch embeddings, including the [**CLS**] token.

$\mathbf{X}_{\text{input}}$ is then passed through the transformer encoder layers. Each transformer encoder layer consists of multi-head self-attention (MHSA) and feed-forward network (FFN). MHSA mechanism allows each patch to take care of all other patches in the image, capturing long-range dependencies. FFN is a position-wise feed-forward network applied to each token independently. The output of each transformer encoder layer is passed to the next encoder layer. After passing through the Transformer encoder layers, the output sequence $\mathbf{X}_{\text{output}}$ contains information for all patches as well as the [**CLS**] token. The output of the [CLS] token from the final encoder layer is used as the final global image embedding. This output vector is considered to be the encoded feature of the entire image as shown in Eq. (5) as follows:

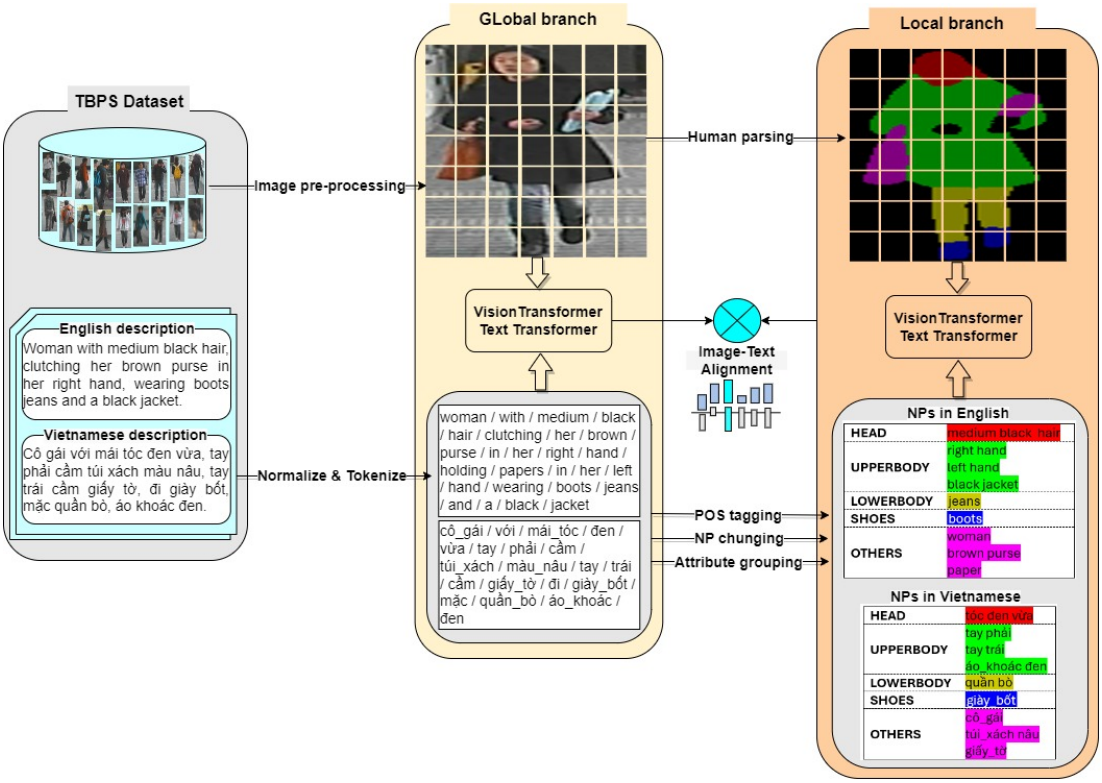

Figure 1: The overall framework for global and local feature learning using pre-trained CLIP model and fine-tuning on text-based person search dataset of CUHK-PEDES dataset.

$$\mathbf{z}_{\text{img}}^{(glob)} = \mathbf{X}_{\text{output}}[0] \tag{5}$$

where $\mathbf{X}_{\text{output}}[0] \in \mathbb{R}^{(N+1)\times D}$ corresponds to the feature vector of the [**CLS**] token; $\mathbf{z}_{\text{img}}^{(glob)}$ is the global image embedding representing the entire image after it was processed by the Vision Transformer. The global visual embedding $\mathbf{z}_{\text{img}}^{(glob)}$ is normalized using the normalization $L2$, to make it easier to compare with the global textual embedding.

- Global textual feature embedding: The input text **T** is first tokenized in a sequence of n tokens. Each token t_i is assigned to an embedding vector $\mathbf{e}_i$. The embedding for a token is obtained by multiplying the token index with a learned embedding matrix $\mathbf{W}_{txt}$ as shown in Eq. (6) follows:

$$\mathbf{e}_i = \mathbf{W}_t \cdot t_i \tag{6}$$

where $\mathbf{W}_t$ is the token embedding matrix of size $V \times D$, where V denotes the vocabulary size and D is the embedding dimension. The result of this step is a sequence of token embeddings $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, .., \mathbf{e}_n]$. The next step is adding positional encodings to the token embeddings to provide information about the positions of tokens in the sequence. The positional encoding for token t_i is denoted a $\mathbf{PE}_i$, and the final embedding for token t_i is $\mathbf{e}_i^{\text{pos}} = \mathbf{e}_i + \mathbf{PE}_i$. This results in a sequence of embeddings with positional information: $\mathbf{E}^{\text{pos}} = \left[\mathbf{e}_1^{\text{pos}}, \mathbf{e}_2^{\text{pos}}, .., \mathbf{e}_n^{\text{pos}}\right]$. The sequence $\mathbf{E}^{\text{pos}}$ passed through the transformer layers to obtain the output embeddings $h_1^{(L)}, h_2^{(L)}, .., h_N^{(L)}$, with L denoting the number of transformer layers and n the number of tokens in the input. The last token in the sequence is used as the final text embedding: $\mathbf{z}_{txt}^{(global)} = h_N^{(L)}$. The global textual embedding $\mathbf{z}_{\text{txt}}^{(glob)}$ is normalized using the normalization $L2$, to make it easier to compare with the global visual embedding.

3.2 Local feature learning

Local feature learning in the context of TBPS aims to associate specific parts of the human body in the image with corresponding noun phrases (NPs) extracted from the textual description. In our approach, we define five semantic body regions: **HEAD**, **UPPERBODY**, **LOWERBODY**, **SHOES**, and **OTHERS**. These regions are obtained using a human parsing network as in [4], which segments the person image into semantically meaningful parts. Examples of parsing results are shown in Fig. 2.

To associate textual descriptions with visual body regions, we extract noun phrases using the spaCy toolkit and map them to body parts using predefined keyword lists in both English and Vietnamese. The mapping rules are as follows:

– **HEAD**:

  – **English keywords**: ”hair”, ”head”, ”face”,

Figure 2: Examples of human parsing images from CUHK-PEDES dataset (five images in the first block) and 3000VnPerson-Search dataset (five images in the right block). The first row shows the original images and the second row depicts the corresponding human parsing images.

”eyes”, ”ear”, ”mouth”, ”nose”, ”teeth”, ”beard”, ”cheeks”, ”bald”, ”chin”, ”eyebrows”, ”lips”, ”spectacles”, ”mustache”, ”goatee”.

- **Vietnamese keywords**: ”tóc”, ”mặt”, ”mắt”, ”mũi”, ”miệng”, ”răng”, ”tai”, ”râu”, ”ria”, ”kính”, ”mép”, ”cằm”, ”má”, ”hói”, ”lông_mày”, ”lông_mi”.

– **UPPERBODY**:

- **English keywords**: ”shirt”, ”t-shirt”, ”jacket”, ”hoodie”, ”sweater”, ”blouse”, ”coat”, ”vest”, ”cardigan”, ”top”, ”jumper”, ”shoulders”, ”chest”, ”arms”, ”biceps”, ”back”, ”neck”, ”robe”, ”cloak”, ”outerwear”.
- **Vietnamese keywords**: ”áo”, ”vai”, ”ngực”, ”tay”, ”lưng”, ”cổ”.

– **LOWERBODY**:

- **English keywords**: ”pants”, ”jeans”, ”trousers”, ”leggings”, ”shorts”, ”skirt”, ”denim”, ”dress”, ”hips”, ”thighs”, ”calves”.
- **Vietnamese keywords**: ”quần”, ”chân”, ”váy”, ”mông”, ”đùi”.

– **SHOES**:

- **English keywords**: ”shoes”, ”sneakers”, ”boots”, ”sandals”, ”slippers”, ”loafers”, ”heels”, ”socks”.
- **Vietnamese keywords**: ”giày”, ”dép”, ”tất”.

– **OTHERS**:

- **English keywords**: ”bag”, ”purse”, ”watch”, ”umbrella”, ”phone”, ”tie”, ”glasses”, ”jewelry”, ”camera”.
- **Vietnamese keywords**: ”túi”, ”cặp”, ”ví”, ”đồng_hồ”, ”ô”, ”điện_thoại”, ”cà_vạt”, ”kính”, ”trang_sức”, ”máy_ảnh”.

Figure 1 illustrates how the extracted NPs (in both English and Vietnamese) are aligned with each semantic region. For instance, ”medium black hair” (”tóc đen vừa”) is mapped to HEAD; ”black jacket”(”áo khoác đen”) to UPPERBODY; ”jeans”(”quần bò”) to LOWERBODY; and so on.

While this rule-based approach is lightweight and interpretable, it has limitations. First, the mapping is inherently *language-dependent*, relying on fixed keyword lists that may not capture paraphrases, figurative language, or culturally specific expressions. This challenge is particularly relevant in low-resource languages like Vietnamese, where lexical diversity can reduce coverage consistency. We provide further discussion of these limitations in Section 5.

To evaluate the robustness of our NP-body-part mapping strategy, we conducted a coverage analysis using the 3000VnPerson-Search dataset. Table 2 shows the total and matched noun phrases per category. The HEAD and UPPERBODY parts achieve high coverage (>85%), while OTHERS shows more variability due to accessory and object descriptions.

This analysis demonstrates the utility of the rule-based strategy in covering core regions. However, the variation observed in OTHERS suggests future extensions could include learned semantic-aware mappings to increase robustness and linguistic generalization.

Finally, the matched NPs and their corresponding image regions are used in the local feature alignment module, leveraging the same ViT-B/32 and GPT-2 encoders used in the global branch. A contrastive loss is applied to optimize region-to-phrase alignment in a joint embedding space.

Table 2: NP-to-body-part mapping coverage by category on the 3000VnPerson-Search dataset

Body Part	Total NPs	Matched	Coverage (%)
HEAD	243	217	89.3
UPPERBODY	398	348	87.4
LOWERBODY	366	310	84.7
SHOES	155	129	83.2
OTHERS	271	190	70.1
Overall	1433	1194	83.3

3.3 Joint learning of image and text

Join learning of image and text is done in an end-to-end manner with the cross-entropy loss function being used to compare the cosine similarities between the image and text embeddings (logits) at both global and local branches. The output is typically used in symmetric loss (image-to-text and text-to-image), and the cross-entropy loss is calculated along both dimensions (image-to-text and text-to-image).

Considering the global feature embedding, we have the Cross Entropy Loss for Image-to-Text as shown in Eq. (7) as follows:

$$\mathcal{L}_{img}^{\text{global}} = -\sum_{i=1}^{N} \log \left(\frac{\exp\left(\text{logits}_{i,i}\right)}{\sum_{j=1}^{N} \exp\left(\text{logits}_{i,j}\right)} \right) \quad (7)$$

where $\text{logits} = \left(\mathbf{z}_{img} \cdot \mathbf{z}_{txt}^{T}\right) \cdot \exp(t)$; $\left(\mathbf{z}_{img} \cdot \mathbf{z}_{txt}^{T}\right)$ is the cosine similarity between the normalized image and text embedding vectors; $\exp(t)$ is applied element-wise to scale the logits, making the similarity values sharper; $\exp\left(\text{logits}_{i,i}\right)$ is the similarity of the image with the correct matching text (positive class); $\sum_{j=1}^{N} \exp\left(\text{logits}_{i,j}\right)$ is the sum of exponentiation logits over all possible text embeddings (this accounts for all possible text candidates). The temperature parameter t is a scalar hyperparameter used to scale the similarity logits. A smaller t produces a sharper probability distribution, emphasizing high-confidence matches. In our implementation, we set $t = 0.07$, following standard practice in contrastive learning frameworks.

Similarly, the cross-entropy loss for text-to-image in the global branch is calculated as shown in Eq. (8) as follows:

$$\mathcal{L}_{txt}^{\text{global}} = -\sum_{i=1}^{N} \log \left(\frac{\exp\left(\text{logits}_{i,i}\right)}{\sum_{j=1}^{N} \exp\left(\text{logits}_{i,j}\right)} \right) \quad (8)$$

The symmetric cross-entropy loss for the global branch is the average of the two loss elements: a global loss function of image-to-text ($\mathcal{L}_{img}^{\text{global}}$) and a global loss function of text-to-image ($\mathcal{L}_{txt}^{\text{global}}$) as illustrated in Eq. (9) as follows:

$$\mathcal{L}_{symCE}^{\text{global}} = \frac{1}{2}\left(\mathcal{L}_{img}^{\text{global}} + \mathcal{L}_{txt}^{\text{global}}\right) \quad (9)$$

Similarly, we have the symmetric cross-entropy loss for local branch, it is also composed of two local loss functions as shown in Eq. (10) as follows::

$$\mathcal{L}_{symCE}^{\text{local}} = \frac{1}{2}\left(\mathcal{L}_{img}^{\text{local}} + \mathcal{L}_{txt}^{\text{local}}\right) \quad (10)$$

Finally, the total loss function is combined of the global loss part and the local loss part as shown in Eq. (11) as follows:

$$\mathcal{L}_{total} = \alpha \cdot \mathcal{L}_{symCE}^{\text{global}} + (1-\alpha) \cdot \mathcal{L}_{symCE}^{\text{local}} \quad (11)$$

Where α is a hyper-parameter that controls the weight of the global versus local features.

4 Experiments and results

4.1 Dataset and configuration parameters

4.1.1 Experimental dataset

In this work, we used four datasets for text-based person search experiments in English and Vietnamese description languages. These datasets are CUHK-PEDES [8], CUHK-PEDES-VN [13], 3000VnPersonSearch [14], and 3000Vn-V2E [24]. Table 3 summarizes these datasets and their main characteristics.

CUHK-PEDES-VN is the Vietnamese-translated version of the CUHK-PEDES dataset [8], which is the first large-scale dataset for text-based person search. CUHK-PEDES contains 40,220 images of 13,003 individuals collected from various person re-identification datasets. Each image is annotated with two natural language descriptions written in English by independent workers on Amazon Mechanical Turk (AMT), resulting in a total of 80,440 descriptions. These descriptions are typically long, diverse in vocabulary, and contain minimal repetition.

To support Vietnamese-language experiments in text-based person search, all English descriptions were translated into Vietnamese using the Google Translate API, without altering the original image-description pairings. This translation forms the CUHK-PEDES-VN dataset, which retains 80,440 Vietnamese descriptions. The average description length is 26.10 words, with a standard deviation of 9.83; the longest description contains 115 words and the shortest 9 words. Furthermore, the noun phrases (NP) extracted from the original English descriptions were also translated into Vietnamese for use in NP-to-body-part mapping experiments in this study.

Table 3: Summary of datasets used in the experiments

Dataset	Language	#Person IDs	#Images	#Descriptions	Avg. Length	Notes
CUHK-PEDES	English	13,003	40,220	80,440	∼26.10 words/tokens	Two English descriptions per image written by AMT workers
CUHK-PEDES-VN	Vietnamese	13,003	40,220	80,440	∼26.10 words/tokens	Vietnamese version of CUHK-PEDES, translated via Google Translate
3000VnPerson-Search	Vietnamese	3,000	6,302	12,602	∼30.02 words/tokens	Collected from surveillance videos; annotated by Vietnamese speakers
3000Vn-V2E	English	3,000	6,302	12,602	∼30.02 words/tokens	English translation of 3000VnPerson-Search using Google Translate

To ensure correct evaluation and avoid data leakage, we strictly follow a non-overlapping person ID split between the training and test sets across all datasets. Specifically, for the CUHK-PEDES dataset, we use the standard split of 11,003 person IDs for training and 1,000 distinct IDs for testing, with no identity overlap. For the 3000VnPerson-Search and 3000Vn-V2E datasets, we divide the 3,000 person IDs into 2,000 for training and 1,000 for testing, ensuring exclusive person IDs in each subset. CUHK-PEDES-VN follows the same split as the original English dataset. This split strategy ensures that the model is always evaluated on unseen identities, conforming to the standard TBPS protocol.

The 3000Vn-V2E dataset consists of English translations of Vietnamese descriptions from the 3000VnPersonSearch dataset [1]. The translations were automatically generated using the Google Translate API and no changes were made to the image description pairings.

The original 3000VnPersonSearch dataset includes 3,000 person identities, 6,302 person images, and 12,602 Vietnamese textual descriptions. Each image is described by two independent annotators to ensure linguistic diversity, with a focus on visual appearance. Each description contains one or more sentences. The average description length is 30.02 tokens, with the longest containing 95 tokens and the shortest 7 tokens. The dataset contains 1,827 unique Vietnamese tokens with a total of 378,274 occurrences.

The person images were collected from surveillance footage in crowded public areas, primarily during the daytime under normal weather and lighting conditions. Human bounding boxes were annotated using a combination of manual tools (e.g., LabelImg) and automated object detection via YOLOv8.

4.1.2 Configuration parameters

Frameworks and environment: All experiments were implemented using Python 3.8, PyTorch v1.13.1, and the HuggingFace Transformers library v4.28.1. The CLIP model was based on OpenAI's official repository (commit `cc5d2e0`), and the PhoBERT model was obtained from the HuggingFace model hub (`vinai/phobert-base`).

[1] Both the 3000VnPersonSearch and 3000Vn-V2E datasets are released for non-commercial academic research purposes only and are available upon request.

Fine-tuning hyperparameters: We fine-tune the model for 20 epochs with a learning rate of 5×10^{-6}. The learning rate is unitless and is applied per optimization step. The batch size is set to 32 image-text pairs and 512 tokens for textual input. We use the Adam optimizer with a weight decay of 0.01. A dropout rate of 0.1 is applied in both the vision and text transformer backbones to mitigate overfitting. For image preprocessing, we apply standard normalization and random horizontal flipping. Text descriptions are tokenized and lowercased using the respective tokenizer of each pre-trained language model.

Vision transformer parameters: We use the ViT-B/32 architecture [21] as the image encoder, consisting of 12 transformer encoder layers, 12 attention heads and an embedding dimension of 768. The input images are resized to 224×224, divided into 32×32 patches, and passed through a linear projection followed by positional encoding and layer normalization.

Text transformer parameters: For English descriptions, we use token embedding, positional embedding and a transformer with 12 residual attention layers, each with 8 attention heads, and embedding dimension of 512. For Vietnamese descriptions, we use **PhoBERT-base** [23], which shares the same architectural configuration.

Loss weighting parameter: In the joint loss formulation (Equation 11), we empirically evaluated the weighting parameter $\alpha \in \{0.2, 0.5, 0.8\}$ to balance global and local alignment. The best trade-off was observed at $\alpha = 0.5$, which we adopt in all final experiments.

Hardware setup: All experiments were conducted on a machine equipped with a single NVIDIA Corporation Device 2504 GPU and 48GB of RAM. The average training time per experimental scenario (20 epochs) was approximately 5 hours. This setup was sufficient for training both global-only and joint global-local models, though we acknowledge limitations in scaling to larger architectures (e.g. ViT-L or ViT-H), as discussed in Section 5.1 and Section 6.

4.2 Evaluation metric

In this work, the top k ranking metric is used for experimental evaluations of text-based person search. The proposed system produces a list of person images that were found for each query sentence, sorted by confidence score. If the corresponding person's image appears in the top k images,

a relevant person retrieval has been performed. The Top-1 (Rank-1), Top-5 (Rank-5), and Top-10 (Rank-10) accuracies are reported in our experiments.

4.3 Result

The experimental results are presented in multiple scenarios to demonstrate the challenges of text-based person search (TBPS) in different languages, particularly in low-resource settings like Vietnamese. In addition, the results also highlight the performance differences between intra-dataset and cross-dataset evaluations.

4.3.1 Evaluation on English language

To evaluate our proposed method **GLAlign** and compare it with existing state-of-the-art (SOTA) methods for English TBPS, we used the CUHK-PEDES dataset and the 3000Vn-V2E dataset. Four experimental scenarios are implemented. The first scenario (**SE1**) uses 11k IDs from CUHK-PEDES for fine-tuning and 1k IDs for testing. This is a standard setting commonly used in prior works. The second scenario (**SE2**) involves the 3000Vn-V2E dataset, with 2k IDs for training and 1k IDs for testing. The third scenario (**SE3**) evaluates cross-dataset performance: the model is trained on 11k IDs from CUHK-PEDES and tested on 1k IDs from 3000Vn-V2E. The fourth scenario (**SE4**) enriches the training data by combining 11k IDs from CUHK-PEDES with 2k IDs from 3000Vn-V2E, while testing remains on the remaining 1k IDs from 3000Vn-V2E.

Table 4 presents the results for the first scenario (**SE1**). Our method achieves a Rank-1 accuracy of 80.75%, outperforming recent transformer-based methods such as IRRA [10] (73.38%), PLOT [11](75.28%) and RaSa [12] (76.51%). Compared to earlier CNN- or RNN-based methods (e.g., GNA-RNN [8], Dual-path CNN [15], ViTAA [4], SCRF [9]), the improvement ranges from approximately 25% to over 60% in Rank-1 accuracy.

Table 5 reports the results for the second scenario (**SE2**). Our model, fine-tuned on 2k IDs from 3000Vn-V2E, achieves Rank-1, Rank-5, and Rank-10 accuracies of 67.50%, 92.96% and 96.82%, respectively. This performance significantly exceeds M-IRRA [24] (Rank-1 = 37.07%). However, performance is slightly lower than that of the CUHK-PEDES-trained model (Table 4), due to the smaller training dataset.

Table 6 illustrates the results of the third scenario (**SE3** for cross-dataset). Our method achieves an accuracy of 65.40% Rank-1, surpassing ViTAA (25.31%) and IRRA (51.35%). However, it underperforms compared to the intra-dataset scenarios, which yielded 80.75% and 67.50% in the first and second settings, respectively.

As shown in Table 7, the fourth scenario (**SE4**) benefits from the enrichment of training data, improving performance to 70.86% Rank-1 accuracy, compared to 65.40% in the third scenario. The additional training data boosts the model's generalization across domains.

4.3.2 Evaluation on Vietnamese language

We evaluated our method on Vietnamese TBPS using three scenarios in the CUHK-PEDES-VN and 3000VnPerson-Search datasets. In the first scenario (**SV1**), we fine-tuned 11k IDs and tested 1k IDs of CUHK-PEDES-VN. The second scenario (**SV2**) trains on CUHK-PEDES-VN and tests on 100 manually annotated samples from 3000VnPerson-Search. The third scenario (**SV3**) scales this evaluation to the full 3k IDs in the 3000VnPerson-Search.

Table 8 shows the results of the first experimental scenario (**SV1**). Our proposed method outperforms the methods in [14] and [13]. The Rank-1 accuracy of our method is 77.72% compared to 52.40% in [14] and 33.07% in [13]. Compared with the results on the same scenario but for the English language (Table 4), the precision for Vietnamese descriptions is about 3% lower than the accuracy of English descriptions in all ranks. This is due to the quality of the translation from English to Vietnamese using Google API for descriptive sentences. The semantics of the description may not be perfectly transferred from English to Vietnamese. This presents a linguistic challenge when used for TBPS. In order to effectively image search for a particular description language, the model should be executed with data from that language itself. Using descriptive data from other languages through text translation helps reduce the cost of data building, but can reduce the efficiency of TBPS.

To statistically assess whether the performance difference between the English and Vietnamese descriptions is significant, we conducted McNemar's test [25] on 1,000 paired queries using the prediction results of CUHK-PEDES (English) and CUHK-PEDES-VN (translated Vietnamese). McNemar's test is a non-parametric statistical test used to evaluate whether there is a significant difference between the proportions of two related binary outcomes. It is particularly suitable for paired nominal data, such as comparing whether a retrieval result for the same query is correct under two different systems or conditions.

The test yielded a p value less than 0.01, indicating that the difference in Rank-1 retrieval accuracy (80.75% vs. 77.72%) is statistically significant. This means that the likelihood of observing such a performance gap under the null hypothesis of no difference is less than 1%, supporting the conclusion that automatic translation introduces meaningful semantic noise affecting the accuracy of the retrieval.

Table 9 presents the experimental results with the second scenario of Vietnamese text-based person search (**SV2**): fine-tuning the model on 11k IDs of the CUHK-PEDES-VN dataset and testing on 100 IDs of the 3000VnPerson-Search dataset. This experimental scenario is closer to the real problem of text-based person search, as the model is trained on a dataset and tested on a new data set with descriptive sentences written by people from another culture. In fact, descriptions for the same person image when done by different people will not be exactly the same, depending on each descriptor's perspective, writing style, and culture.

Table 4: The comparative results of our method (**GLAlign**) with other SOTA methods on the first experimental scenario-**SE1**: fine tuning on 11k IDs and testing on 1k IDs of CUHK-PEDES dataset

Method	Backbone		Rank-1	Rank-5	Rank-10
	Image	Text			
GNA-RNN [8]	VGG16	LSTM	19.71%	43.20%	55.24%
Dual-path CNN [15]	ResNet50	ResNet50	39.06%	61.67%	70.69%
Method in [14]	ResNet50	ResNet50	41.48%	62.89%	72.59%
ViTAA [4]	ResNet50	Bi-LSTM	55.97%	75.84%	83.52%
SCRF [9]	ResNet50	BERT	64.04%	82.99%	88.81%
IRRA [10]	CLIP-ViT-B/16	CLIP-Xformer	73.38%	89.93%	93.71%
PLOT [11]	CLIP-ViT-B/16	CLIP-Xformer	75.28%	90.42%	94.12%
RaSa [12]	ALBEF	ALBEF	76.51%	90.29%	94.25%
GLAlign (Ours)	ViT-B/32	GPT-2	**80.75%**	**97.85%**	**99.38%**

Table 5: The comparative results of our method (**GLAlign**) with other methods on the second experimental scenario-**SE2**: fine tuning on 2k IDs and testing on 1k IDs of 3000Vn-V2E dataset

Method	Backbone		Rank-1	Rank-5	Rank-10
	Image	Text			
GNA-RNN [8]	VGG16	LSTM	26.85%	52.47%	66.40%
ViTAA [4]	ResNet50	Bi-LSTM	27.08%	51.38%	63.00%
M-IRRA [24]	CLIP-ViT-B/16	Multilingual CLIP	37.07%	66.23%	76.51%
GLAlign (Ours)	ViT-B/32	GPT-2	**67.50%**	**92.96%**	**96.82%**

Table 6: The comparative results of our method (**GLAlign**) with other methods on the third experimental scenario-**SE3**: fine tuning on 11k IDs of CUHK-PEDES dataset and testing on 1k IDs of 3000Vn-V2E dataset

Method	Backbone		Rank-1	Rank-5	Rank-10
	Image	Text			
ViTAA [4]	ResNet50	Bi-LSTM	25.31%	66.20%	59.72%
IRRA [10]	CLIP-ViT-B/16	CLIP-Xformer	51.35%	78.03%	85.78%
GLAlign (Ours)	ViT-B/32	GPT-2	**65.40%**	**94.80%**	**98.40%**

Table 7: The comparative results of our method (**GLAlign**) with other methods on the fourth experimental scenario-**SE4**: fine tuning on 11k IDs of CUHK-PEDES dataset plus 2k IDs of 3000Vn-V2E dataset, and testing on the remaining 1k IDs of 3000Vn-V2E dataset

Method	Backbone		Rank-1	Rank-5	Rank-10
	Image	Text			
ViTAA [4]	ResNet50	Bi-LSTM	31.17%	58.12%	70.35%
IRRA [10]	CLIP-ViT-B/16	CLIP-Xformer	59.57%	82.99%	89.71%
GLAlign (Ours)	ViT-B/32	GPT-2	**70.86%**	**95.10%**	**98.70%**

Table 8: The comparative results of our method (**GLAlign**) with other methods on Vietnamese TBPS, considering the first experimental scenario-**SV1**: training on 11k IDs and testing on 1k IDs of CUHK-PEDES-VN dataset

Method	Backbone		Rank-1	Rank-5	Rank-10
	Image	Text			
Method in [14]	ResNet50	ResNet50	33.07%	53.39%	62.97%
Method in [13]	ResNet50	Bi-LSTM	52.40%	72.28%	80.78%
GLAlign (Ours)	ViT-B/32	GPT-2	**77.72%**	**94.31%**	**97.02%**

Table 9: The comparative results of our method (**GLAlign**) with other methods on Vietnamese TBPS, considering the second experimental scenario-**SV2**: training on 11k IDs of CUHK-PEDES-VN dataset and testing on 100 IDs of 3000VnPerson-Search dataset

Method	Backbone		Rank-1	Rank-5	Rank-10
	Image	Text			
Method in [14]	ResNet50	ResNet50	42.32%	70.04 %	81.52 %
Method in [13]	ResNet50	Bi-LSTM	60.02 %	81.23 %	89.57 %
GLAlign (Ours)	ViT-B/32	GPT-2	**85.72%**	**97.31%**	**99.70%**

Table 10: The comparative results of our method with other methods on Vietnamese TBPS, considering the third experimental scenario-**SV3**: training on 11k IDs of CUHK-PEDES-VN dataset and testing on 3k IDs of 3000VnPerson-Search dataset

Method	Backbone		Rank-1	Rank-5	Rank-10
	Image	Text			
Method in [14]	ResNet50	ResNet50	8.27%	20.92%	29.11%
Method in [13]	ResNet50	Bi-LSTM	28.49	37.26 %	48.94 %
GLAlign (Ours)	ViT-B/32	GPT-2	**50.72%**	**67.31%**	**79.70%**

This shows the challenge of natural language processing in the context of a TBPS problem. The experimental results obtained by our method significantly outperform the SOTA methods in [14] and [13]. The Rank-1, Rank-5, and Rank-10 of our method are 85.72%, 97.31%, and 99.70%, respectively. These results in [14] and [13] are only 42.32% and 60.02% for Rank-1; 70.04% and 81.23% for Rank-5; 81.52% and 89.57% for Rank-10, respectively.

The comparative results of our method and other methods for the third experimental scenario (**SV3**) are shown in Table 10. In this scenario, the model is also trained on 11k IDs of CUHK-PEDES-VN as in the second scenario, but the testing is implemented on all samples (3k IDs) of 3000VnPerson-Search. This is much more challenging than the second test scenario. The experimental results in the third scenario are much lower than the results in the second scenario. The Rank-1 accuracy of our method decreases from 85.72% in the second scenario to 50.72% in the third scenario. This decline also occurred in Rank-5 and Rank-10. Compared to other methods in [14] and [13], the experimental results of our method are still much higher at all Rank values. These experimental results indicate the challenge of TBPS with cross-dataset evaluation and searching on a massive image database.

4.3.3 Impact of global vs. local feature alignment

To quantify the contribution of each component in our proposed framework, we conduct ablation studies on two representative datasets: CUHK-PEDES (English) and 3000VnPerson-Search (Vietnamese). We evaluated three variants of the model:

- **Global-only model:** Uses ViT-B/32 as the vision backbone and GPT-2 or PhoBERT as the text encoder. Only the global [CLS] representation is used for matching; local features and part-level alignment are disabled.
- **Local-only model:** Human body parts and noun phrases are extracted and aligned using the same transformer encoders. The global [CLS] embedding is excluded; only the part-to-part alignment is used.
- **Full model (**GLAlign**):** Combines global and local representations. The final loss is the weighted sum of global and local symmetric cross-entropy losses as defined in Equation (11).

Table 11 presents the ablation results. From the results, we observe that the global-only model performs reasonably well due to the strong representation capacity of pretrained transformers. However, the incorporation of local alignment brings significant additional gains: +6.54% in CUHK-PEDES and +7.84% in 3000VnPerson-Search. This highlights the importance of fine-grained part-level matching in TBPS tasks, especially in linguistically complex or multilingual environments. We also note that the local-only model, while weaker than the global-only model, still retains competitive performance. This suggests that part-to-part alignment alone carries valuable semantic information but works best when complemented with global contextual embeddings.

5 Discussion

In this section, we provide a deeper analysis of the experimental results and explain the observed performance differences between datasets and language settings compared to existing state-of-the-art (SOTA) methods summarized in Table 1. The discussion focuses on three main aspects: model architecture, cross-lingual performance, and generalization across datasets.

5.1 Effectiveness of joint global-local alignment

Our proposed method **GLAlign** outperforms all baselines in English and Vietnamese datasets, achieving a Rank-1 score of 80.75% in CUHK-PEDES and 85.72% in 3000VnPerson-Search. These gains can be primarily attributed to the integration of global and local feature alignment in a transformer-based framework.

Unlike previous transformer-based methods such as CLIP [6] or RaSa [12], which mainly focus on global visual-textual alignment, our framework incorporates local-level matching through noun phrase (NP) to body part mapping. By leveraging human parsing and NP extraction, we create fine-grained semantic correspondences between specific textual attributes (e.g., ”red jacket”, ”black shoes”) and corresponding visual regions (e.g., upper body, shoes). This granularity enhances discriminative power during retrieval, especially when the global context is ambiguous or insufficient.

Table 11: Ablation study of Rank-1 accuracy on CUHK-PEDES (English) and 3000VnPerson-Search (Vietnamese) under experimental scenarios **SE1** and **SV2**, respectively. We compare global-only, local-only, and full (joint global-local) model variants.

Model Variant	CUHK-PEDES (English)	3000VnPerson-Search (Vietnamese)
Global-only model	74.21%	77.88%
Local-only model	68.10%	72.35%
Full model (GLAlign)	**80.75%**	**85.72%**

5.2 Cross-lingual performance and linguistic variance

Our framework also shows strong performance across languages. In particular, the Rank-1 accuracy on the Vietnamese-native 3000VnPerson-Search dataset (85.72%) is even higher than on CUHK-PEDES (80.75%). This improvement stems from the use of PhoBERT, a pre-trained Vietnamese language model that captures native linguistic features more accurately than translated sentences.

In contrast, when evaluating on CUHK-PEDES-VN (a machine-translated version of the English dataset), we observe a slight performance drop (Rank-1 = 77.72%). This is due to semantic drift and structural inconsistency introduced during translation, which can misalign noun phrases with visual attributes. To illustrate this, we empirically estimated the translation-induced mismatch rate by sampling 200 randomly selected description-image pairs from the translated dataset and manually checking whether the extracted Vietnamese noun phrases semantically match the associated body parts. We observed an approximate error rate of 18.50%, mainly due to lexical ambiguity or loss of specificity.

Despite overall strong performance, we also analyzed failure cases to better understand the model's limitations in cross-lingual settings. These errors often fall into four categories:

- **Ambiguous or generic descriptions:** Queries such as "a man in dark clothes" or "a person walking" lack distinctive visual details, resulting in a high search-and-retrieval confusion.
- **Low-quality or poorly illuminated images:** In the 3000VnPerson-Search dataset, images captured under poor lighting or motion blur reduce visual discriminability, especially in local regions.
- **Partial occlusion:** When key attributes such as "yellow handbag" or "white shoes" are occluded or not visible due to cropping, part-based matching fails to align these cues correctly.
- **Translation-induced mismatch:** Subtle errors in automatic translation (as discussed above) can cause mismatches in noun phrases or misinterpretation of clothing and objects.

We illustrate representative cases in Figure 3, where the global-only model often fails due to these issues, while local alignment partially mitigates some of them.

5.3 Cross-dataset generalization

Although intra-dataset training and testing yield the best results, we also evaluate the model's cross-dataset generalization. When fine-tuned on CUHK-PEDES and tested on 3000Vn-V2E (cross dataset), Rank-1 drops to 65.40%. This performance reduction is expected due to domain change (e.g., surveillance image conditions, clothing diversity, and cultural description styles).

However, compared to other methods under the same conditions (e.g., IRRA at 51.35%), our framework still achieves superior generalization. This is likely because the shared embedding space learned from both global and local features is more transferable and can better accommodate unseen data distributions.

5.4 Summary of observations

The findings discussed in the previous subsections highlight several key takeaways about the effectiveness, robustness, and generalization capacity of our proposed framework. We summarize the most notable insights as follows.

- Performance improvement is mainly driven by the combination of local-level NP-to-body-part mapping and global transformer-based embeddings.
- PhoBERT-based textual encoding significantly boosts results in native Vietnamese descriptions compared to translated datasets.
- The model demonstrates strong cross-dataset generalization, with lower sensitivity to domain and language changes than prior methods.

Overall, the experimental and comparative analysis confirms that joint global-local alignment, when integrated into a fine-tuned transformer framework, leads to superior performance in both monolingual and multilingual TBPS tasks.

5.5 Limitations and future work

Although our method achieves state-of-the-art performance in multiple TBPS scenarios, there are several limitations that suggest promising directions for future work.

Model size and computational cost. In this study, we adopted ViT-B/32 as the visual encoder due to its balance between performance and computational feasibility on our hardware setup (a single NVIDIA GTX 1650 GPU with

(a) Ambiguous description: A person wearing dark clothes (người đàn ông mặc quần áo tối màu)

The global-only model fails due to ambiguous description; full model also struggles under lack of visual cues.

(b) Poor lighting: A man in a red striped sweatshirt, light colored pants, yellow backpack, red shoes (người đàn ông mặc áo nỉ kẻ màu đỏ, quần sáng màu, đeo bao lô vàng, đi giày màu đỏ.)

In low-light conditions, visual cues such as color and texture become less distinguishable, leading to incorrect retrieval by the global-only model. The full model still succeeds by focusing on localized part-level features that remain discriminative despite reduced image quality.

(c) Partial occlusion: A woman carrying a white handbag (Một người phụ nữ mang một chiếc túi xách màu trắng)

When the target object (e.g., a white handbag) is partially occluded, the global-only model fails to identify the correct match. The full model correctly retrieves the target by leveraging localized alignment, which focuses on visible cues such as clothing and posture.

(d) Translation mismatch: A man wears a gray T-shirt with white logo (Áo phông xám có logo màu trắng)

Translation artifacts introduce semantic drift that affects both global and local matching. In this case, neither model retrieves the correct result due to a mismatch between the translated description and the actual visual attributes.

Figure 3: Qualitative examples illustrating four representative failure types in text-based person search: (a) ambiguous descriptions, (b) poor image quality, (c) partial occlusion, and (d) translation-induced semantic drift. Each row shows (from left to right): the input query image, the top-1 retrieved result from the global-only model, and the top-1 result from our full model (global + local features). Red bounding boxes indicate incorrect retrievals, while green bounding boxes denote correct matches. While the full model improves performance in many cases, it still struggles under extreme ambiguity or visual noise.

8GB RAM). However, recent work shows that larger transformer models such as ViT-L/16 and ViT-H/14 [21] achieve better performance in vision tasks that require high granularity. These deeper models have more parameters (ViT-

B: 86M, ViT-L: 307M, ViT-H: 632M), but they also incur higher memory demands and training time. Additionally, training larger models on limited datasets such as 3000VnPerson-Search may cause overfitting. In future work, we plan to investigate the use of deeper backbones with distillation or low-rank adaptation strategies to balance efficiency and accuracy.

Rule-based NP-to-body-part mapping. Our current local alignment is based on matching the rules of noun phrases with body parts using predefined keyword lists. This approach is language-dependent and can be brittle when handling paraphrases, synonyms, or figurative language. Errors are more likely in low-resource languages, such as Vietnamese, where linguistic diversity is greater. Future research may explore neural phrase grounding or attention-guided part discovery methods to enhance the flexibility and robustness of local alignment.

Dataset scale and generalization. The performance of our model varies between datasets, especially in cross-dataset or cross-lingual settings. Although our method generalizes well compared to previous work, further gains could be achieved by training on larger multilingual datasets, including real-world surveillance footage and human-written multilingual descriptions. This would also mitigate semantic drift caused by automatic translation, as discussed in Section 5.3.

Real-world deployment challenges. We observe several failure modes under real-world conditions, including poor lighting, partial occlusion, and vague descriptions. These cases highlight the need to incorporate uncertainty modeling, visual quality enhancement, and query disambiguation techniques into future versions of the TBPS system.

In general, while our framework establishes a strong foundation, we believe that improving scalability, multilingual adaptability, and visual grounding mechanisms will be critical to advancing TBPS systems in practical deployments.

6 Conclusion

In this work, an efficient framework for TBPS called GLAlign is proposed. In this framework, both global and local features of image and text pairs are considered when fine-tuning the large-scale and pretrained transformer models on standard datasets. Several experimental scenarios are implemented to show the outperforming results of our proposed solution compared to other SOTA methods in both descriptive languages, English and Vietnamese. Fine-tuning the large-scale and pretrained transformers on a large enough dataset as CUHK-PEDES brings higher searching performance than a smaller dataset of 3000Vn-V2E dataset. The TBPS performance of the cross-dataset evaluation is decreased compared to the intra-dataset evaluation. Considering the experimental results with English descriptions in the dataset, the results with Vietnamese descriptions are lower. These experimental observations show the challenges of the TBPS problem. Although our proposed method has improved results compared to other SOTA methods, further improvements in feature learning and matching, as well as the enhanced TBPS database, should be implemented in the future to improve the efficiency of the TBPS.

References

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", *arXiv preprint arXiv:1409.1556*, 2014.

[2] C. Gao, G. Cai, X. Jiang, *et al.*, "Contextual non-local alignment over full-scale representation for text-based person search", Jan. 2021. DOI: 10.48550/arXiv.2101.03036.

[3] X. Han, S. He, L. Zhang, and T. Xiang, "Text-based person search with limited data", *arXiv preprint arXiv:2110.10807*, p. 13, 2021. DOI: 10.48550/arXiv.2404.18106.

[4] Z. Wang, Z. Fang, J. Wang, and Y. Yang, "Vitaa: Visual-textual attributes alignment in person search by natural language", in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, Springer, 2020, pp. 402–420. DOI: 10.1007/978-3-030-58610-2_24.

[5] *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, DOI: 10.1109/TNNLS.2023.3310118.

[6] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision", in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763. DOI: 10.48550/arXiv.2103.00020.

[7] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation", in *International conference on machine learning*, PMLR, 2022, pp. 12 888–12 900. DOI: 10.48550/arXiv.2201.12086.

[8] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description", in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5187–5196. DOI: 10.1109/CVPR.2017.551.

[9] W. Suo, M. Sun, K. Niu, *et al.*, "A simple and robust correlation filtering method for text-based person search", in *European conference on computer vision*, Springer, 2022, pp. 726–742. DOI: 10.1007/978-3-031-19833-5_42.

[10] M. Cao, Y. Bai, Z. Zeng, M. Ye, and M. Zhang, "An empirical study of clip for text-based person search", in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 465–473. DOI: 10.48550/arXiv.2308.10045.

[11] J. Park, D. Kim, B. Jeong, and S. Kwak, "Plot: Text-based person search with part slot attention for corresponding part discovery", in *European Conference on Computer Vision*, Springer, 2025, pp. 474–490. DOI: 10.1007/978-3-031-72664-4_27.

[12] Y. Bai, M. Cao, D. Gao, *et al.*, "Rasa: Relation and sensitivity aware representation learning for text-based person search", 2023. DOI: 10.24963/ijcai.2023/62.

[13] T. T. T. Pham, V.-T. Nguyen, H.-Q. Nguyen, *et al.*, "Person search by natural language description in vietnamese using pre-trained visual-textual attributes alignment model", in *2021 13th International Conference on Knowledge and Systems Engineering (KSE)*, IEEE, 2021, pp. 1–6.

[14] T. T. T. Pham, H.-Q. Nguyen, H. Phan, *et al.*, "Towards a large-scale person search by vietnamese natural language: Dataset and methods", *Multimedia Tools and Applications*, vol. 81, no. 19, pp. 27 569–27 600, 2022. DOI: 10.1007/s11042-022-12138-1.

[15] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, "Dual-path convolutional image-text embeddings with instance loss", *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 2, pp. 1–23, 2020. DOI: 10.1145/3383184.

[16] D. Mladenić and M. Grobelnik, "Automatic text analysis by artificial intelligence", *Informatica*, vol. 37, no. 1, pp. 27–33, 2013.

[17] B. Xiao, H. Wu, W. Xu, *et al.*, "Florence-2: Advancing a unified representation for a variety of vision tasks", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4818–4829. DOI: 10.1109/CVPR52733.2024.00461.

[18] J. Li, R. R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation", in *Advances in Neural Information Processing Systems*, 2021.

[19] D. Mladenić and B. Fortuna, "Multilingual document classification and sentiment analysis", *Informatica*, vol. 37, no. 3, pp. 241–252, 2013.

[20] A. Chauhan, A. Sharma, and R. Mohana, "An enhanced aspect-based sentiment analysis model based on roberta for text sentiment analysis", *Informatica*, vol. 49, no. 14, pp. 193–202, 2025. DOI: 10.31449/inf.v49i14.5423.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale", in *International Conference on Learning Representations ICLR (2021)*, 2021.

[22] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, *Language models are unsupervised multitask learners*, Technical report, OpenAI, 2019.

[23] D. Q. Nguyen and A. T. Nguyen, "Phobert: Pre-trained language models for vietnamese", pp. 1037–1042, 2020. DOI: 10.18653/v1/2020.findings-emnlp.92.

[24] H. P. T. Tran, T. H. P. Phan, T. B. N. Nguyen, *et al.*, "M-irra: A multilingual model for text-based person search", in *The 3rd APSIPA Workshop on Signal and Information Processing (SIP) in Vietnam*, 2024, pp. 1–6.

[25] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages", *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.