Towards Identity-Aware Cross-Modal Retrieval: a Dataset and a Baseline

Nicola Messina*[1], Lucia Vadicamo[1], Leo Maltese[2], and Claudio Gennaro[1]

[1] Institute of Information Science and Technologies, CNR, Pisa, Italy
{nicola.messina, lucia.vadicamo,claudio.gennaro}@isti.cnr.it
[2] University of Pisa, Pisa, Italy - l.maltese1@studenti.unipi.it

Abstract. Recent advancements in deep learning have significantly enhanced content-based retrieval methods, notably through models like CLIP that map images and texts into a shared embedding space. However, these methods often struggle with domain-specific entities and long-tail concepts absent from their training data, particularly in identifying specific individuals. In this paper, we explore the task of *identity-aware cross-modal retrieval*, which aims to retrieve images of persons in specific contexts based on natural language queries. This task is critical in various scenarios, such as for searching and browsing personalized video collections or large audio-visual archives maintained by national broadcasters. We introduce a novel dataset, COCO Person FaceSwap (COCO-PFS), derived from the widely used COCO dataset and enriched with deepfake-generated faces from VGGFace2. This dataset addresses the lack of large-scale datasets needed for training and evaluating models for this task. Our experiments assess the performance of different CLIP variations repurposed for this task, including our architecture, Identity-aware CLIP (Id-CLIP), which achieves competitive retrieval performance through targeted fine-tuning. Our contributions lay the groundwork for more robust cross-modal retrieval systems capable of recognizing long-tail identities and contextual nuances. Data and code are available at https://github.com/mesnico/IdCLIP.

Keywords: Personalized Retrieval · Vision and Language · Cross-modal Person Retrieval · Long-tail Concepts Understanding.

1 Introduction

In modern content-based retrieval literature, deep learning techniques that map images and texts into the same common space – like CLIP [22] – are gaining increasing interest and have been used in many critical downstream multimodal scenarios. These models are particularly effective for cross-modal retrieval, enabling efficient searches for relevant images based on natural language queries using straightforward k-nearest neighbor techniques in the forged common space.

* Corresponding Author. Email: nicola.messina@isti.cnr.it

Consequently, they are widely used in large-scale image or video retrieval, allowing users to easily explore large visual collections using natural language descriptions [31].

However, despite their widespread adoption, these methods exhibit significant limitations. A major challenge is their inability to handle queries containing domain-specific entities and long-tail concepts that were not present within their training set. For instance, CLIP struggles to identify specific individuals, objects, or buildings not included during training, which can hinder performance in scenarios like the query, "*Gianni Morandi sings and dances at the Sanremo Festival in Italy, black and white footage*". This limitation becomes particularly important in personalization scenarios, where users seek to locate specific individuals or places within their video collections, as well as in cultural heritage and audiovisual content preservation applications that require retrieving footage of distinct monuments or historical figures in large archives maintained, for example, by national broadcasters.

Among all the possible long-tail concepts, the ability to recognize people's identities is one of the most critical, with great applicability in many of the above-cited scenarios. Although some studies [4,12] have already taken some steps in this direction, the main obstacle remains the lack of sufficient high-quality data to fine-tune a multimodal model and make it more discriminative with respect to identity features. Additionally, evaluating these models poses its own challenges, as it is not enough to merely retrieve images of a particular person; the context specified in the query must also be taken into account. In the above-mentioned example, we would like not just to retrieve all the possible images depicting *Gianni Morandi*, but only the ones in which he is *singing and dancing* in black and white footage.

To tackle these challenges, we propose a novel dataset, COCO-PFS, derived from COCO [16] and containing persons whose faces have been replaced with entities from VGGFace2 [2] employing deepfake generation techniques. Together with the dataset, we introduce suitable metrics able to monitor both the retrieval of the scene context and the person's identity. We experiment with different CLIP variations to show the challenges this benchmark poses. We demonstrate that the original CLIP struggles to differentiate between identities. We also examine CLIP-PAD [12], a recent CLIP variation that represents a significant step toward creating a domain-independent identity-aware model, capable of taking both visual features of a person's face and text describing the context as input to generate similarity scores with each dataset image. On top of this solution, we devise our Identity-aware CLIP (Id-CLIP), which features an effective, clever fine-tuning strategy for the visual backbone to further boost the results, obtaining a strong baseline for this challenging task.

To summarize, the contributions of this paper are as follows:

- We introduce a novel dataset, COCO-PFS, featuring images with faces replaced by a controlled set of public figures.
- We identify a suitable set of metrics able to assess the system's ability to retrieve both the person's identity and the context.

– We run multiple experiments on various CLIP models, also introducing Id-CLIP which utilizes fine-tuning of the visual backbone in order to create a strong baseline and capture insights on this novel challenging task.

2 Related Work

In recent years, vision-language models, and especially image-text matching methods, have gained significant attention for their application in cross-modal retrieval scenarios. Many of this methods adopt hinge-based triplet ranking loss with hard-negative mining using late-fusion approaches [6,14,21,27,34,18,25,22,10] or effective yet less efficient early-fusion methods [32,13,15,17,28,40]. Among late-fusion methods, CLIP [22] has become particularly influential for image-text matching and a strong backbone to solve many downstream vision-language tasks. However, these methods struggle with long-tail or out-of-distribution concepts unless extensively fine-tuned.

Most of the personalized vision-language research has focused on personalized image generation. For instance, [7,20] employed textual inversion to improve personalized feature, while [33,8] proposed encoder-based methods for efficient and accurate customization. Perfusion [30] integrated multiple concepts while preventing overfitting, improving inference speed. Other approaches employ multi-concept extraction from a single image [1], user history-based prompt rewriting [3], and fine-tuning-free method [38].

Some works also started exploring personalized text-image retrieval. Cohen et al. [4] introduced a PerVL setup that leverages few-shot learning to reason about personalized concepts. Differently from our work, they update an already existing generic model, fine-tuning it on a small set of specific concepts. Yeh et al. [35] developed instance-specific embeddings without requiring annotated data, enabling instance retrieval in personal videos through vision-language similarity. The work most similar to ours is by [12], which proposed a method (*CLIP-PAD*) and a dataset (*Celebrities in Action*) for in-context person retrieval. Nevertheless, the proposed dataset collects only very famous actors that may be already present in the CLIP training dataset. Furthermore, our approach builds upon their method by adding a targeted fine-tuning of the CLIP visual backbone to enhance the fine-grained facial understanding abilities.

Entity-aware datasets Most existing computer vision datasets lack named entities in the images or captions, limiting their relevance for identity-aware cross-modal retrieval tasks. While some efforts have leveraged knowledge bases to connect textual names with corresponding entities in multimedia content, they usually involve generic entities, such as *woman*, rather than specific, named individuals [5]. The VELD dataset [41] provides scene graphs linking entities from text to the depicted visual content using DBpedia, but it is not publicly available and covers broader categories such as locations and organizations, which are beyond our scope. WikiPerson [29] offers a dataset of over 48k annotated images with bounding boxes and corresponding Wikidata IDs, but it lacks descriptive

captions, limiting its applicability for identity-aware retrieval. COFAR [9] explores *visual* named entities like business brands and world landmarks, linking them through external knowledge sources, yet it does not include named entities in its captions, hindering its applicability in our context. The recently introduced *This-Is-My* dataset [35] includes annotated video segments with captions for contextualized retrieval, but the available captions are limited to a small query-time dataset split, which only includes 30 instances related to 15 entities. Rosasco et al. [24] proposed *ConCon-Chi*, a benchmark designed to assess the compositional capabilities of models on novel chimeric objects within diverse contexts, revealing limitations in handling new concept-context combinations. However, this is a dataset created in controlled scenarios and has limited variability.

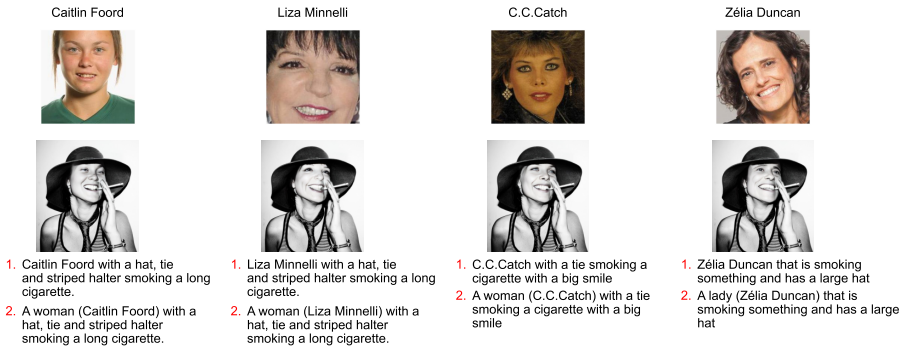
Several other datasets have attempted to link textual names to entities in visual content. For example, [23] presented a dataset with manually annotated names and captions in TV episode videos, which would have been relevant to our task but not publicly available. Similarly, [12] used *Celebrities in Places* and *Celebrities in Action* datasets, containing images and videos of celebrities in various contexts. While highly relevant to our task, these datasets are currently unavailable for download due to privacy concerns.

3 COCO Person FaceSwap (COCO-PFS) Dataset

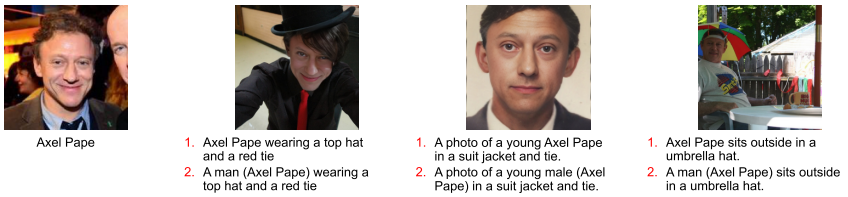
Current widely-used datasets for text-to-image retrieval [16,37,26] generally contain images with generic descriptions, which are not suitable for identity-aware tasks. These datasets often feature random web-scraped images, making it difficult to find multiple images of the same person, and rarely include names in captions due to privacy concerns. These elements are of critical importance for our purpose, which is to understand to which extent state-of-the-art text-to-image retrieval models are able to find the same identity in different contexts.

To address this gap, we developed a novel dataset, COCO Person FaceSwap, which transform an existing text-image dataset into an identity-aware benchmark. Our approach includes an automatic pipeline that: (1) selects from the provided dataset only the images containing at least one person, (2) swaps faces with identities from a public face database using deepfake techniques, and (3) enriches captions to include the substituted identity's name. This section describes in detail the stages of the pipeline we utilized to create our benchmark that better suits the evaluation of identity-aware cross-modal retrieval.

Image Pre-Selection The first step involves selecting images containing individuals, as our focus is solely on retrieving scenes with identifiable persons. We selected images from COCO where only a single person appears, given that it is difficult to automatically modify captions to refer to different identities and to ensure a balanced representation of identities across various contexts. To this scope, we first detected faces in the images using MTCNN [39] pretrained on VGGFace2 [2] and CASIA-Webface [36]. We then filtered out images with multiple detected faces and cross-verified the results using COCO's annotations to



(a) One context, multiple identities.



(b) One identity, multiple contexts.

Fig. 1: Examples from our COCO-PFS dataset showing images obtained by substituting (a) different identities in the same context or (b) the same identity in different contexts. The captions depicted are obtained through two different templates, starting from one of the five original COCO captions.

minimize false positives. For each selected image, we extracted the facial features of the depicted individual using an Inception Resnet (V1), also trained on VGGFace2 and CASIA-Webface.

Gender-Ethnicity Detection and Face-swapping In this stage, we substitute the faces in COCO images with those from a VIP face database, using the VGGFace2 test set, which includes 500 identities. To ensure a balanced dataset, we ensure that each identity is present in at least two different contexts, and each image is swapped with multiple faces to depict various people. We used *Roop* tool[3], an advanced, easy-to-use face-swapping software, to reliably swap the faces. To make these swaps realistic, we don't select faces randomly. Instead, we use *DeepFace tool*[4] – which implements advanced face-analysis models – to detect the gender and ethnicity of the original faces in COCO, selecting a suitable matching face from VGGFace2 for the swap. Before the actual face-swapping phase, we keep only the images in which the individuals' faces are large in proportion to the image size. This filtering step is crucial to ensure the opti-

[3] `https://github.com/s0md3v/roop`
[4] `https://github.com/serengil/deepface`

mal performance of tools such as Deepface and Roop, particularly when dealing with face swapping. In fact, these tools tend to produce superior results when the facial features are prominently visible, and the face occupies a significant portion of the image.

Caption Enrichment To create more personalized and realistic captions, we transform COCO's generic descriptions using a template-based approach. For each image I_i, where N is the swapped person's name, we modify the five original COCO captions $(C_{i1}, \ldots, C_{i5})$ by replacing generic noun referring to the person in the image (e.g., *person*, *man*, *woman*, *kid*, etc.) with the specific name N using k templates, $\{\mathbf{T}^t\}_{t=1}^{k}$. As a results, we generated k alternative versions of each caption, $C_{ij}^t = \mathbf{T}^t(C_{ij}, N)$, for $t = 1, \ldots, k$.

For example, assuming a set of three different templates, the original caption *"A person is running at the park"* and the name N = *"Gianni Morandi"* can be transformed into: (i) "*Gianni Morandi is running at the park*", (ii) "*An image with Gianni Morandi. The person is running at the park*"; (iii) "*The famous person Gianni Morandi is running at the park*". Our dataset has been constructed using $k = 11$ different templates for each caption. For ease of notation, we flatten the indexes j and t to obtain again C_{ij}, where j this time spans from 1 to $5 \cdot k$.

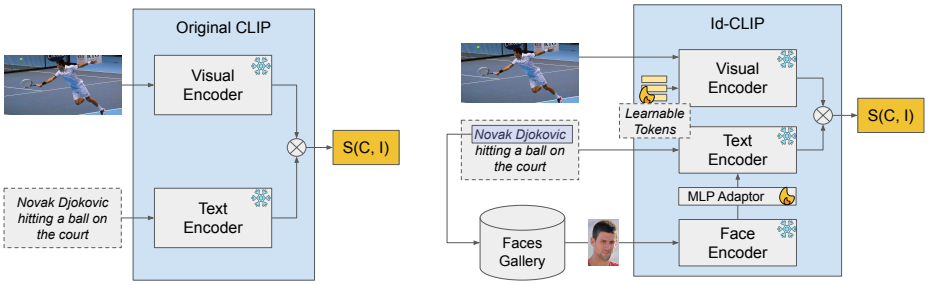
We also release the captions with a [ENTITY] placeholder on behalf of the actual name so that methods can directly employ this information without having to rely on external NER modules for re-extracting the name.

Dataset Overview We applied the proposed pipeline to the training, validation, and test sets of the COCO dataset. The resulting training set contains 49,957 images with 500 distinct entities, resulting in 2,747,635 captions. The validation and test sets each contain 1,760 images, with 466 and 467 entities, respectively, and 96,800 captions per set. We report an example in Figure 1.

4 Identity-aware Cross-modal Retrieval

Identity-aware cross-modal retrieval extends the classic text-to-image retrieval task, where the goal is to find the most relevant images in a potentially large database given a natural language query. In our scenario, textual queries are enriched with person names, and the task is to retrieve all images featuring the specified person in the described context.

To train or evaluate an identity-aware model, we assume the availability of a *dataset* $\mathcal{D} = \{(I_i, \{C_{ij}\}_j), i = 1 \ldots, n\}$, which consists of n images I_i, each associated with one or more descriptive natural language captions $\{C_{ij}\}$. Additionally, we consider a *face gallery* $\mathcal{G} = \{(F_l, N_l)\}_{l=1}^{m}$ containing m identities. Each person in the gallery is represented by an image of their face F_l and their name N_l (e.g., *Mario Rossi*). Each image I_i should contain individuals from $\mathcal{G}$ in various contexts, with the corresponding captions C_{ij} explicitly mentioning the person's name (e.g., *Mario Rossi is riding a bike down the city streets*). This is



(a) Original CLIP. (b) Identity-Aware CLIP (Id-CLIP).

Fig. 2: (a). The original CLIP model. In order to digest entity names, should be properly finetuned and specialized to a given gallery $\mathcal{G}$. (b) Id-CLIP solves the problem by factoring out the entity from the caption. Through $\mathcal{G}$, the name is used to lookup the face, which becomes an additional input to the system. The learnable tokens in input to the Visual Encoder implement *visual prompt tuning*.

exactly how our COCO-PFS dataset is designed. The face gallery – VGGFace2 in our case – serves as an external knowledge base, allowing us to retrieve a face F_l given a person's name N_l. In this setting, given any of the descriptions C_{ij} as a query, the identity-aware retrieval objective is to rank all n images of the dataset by decreasing relevance, ensuring that I_i – the *correct* image containing the person mentioned in C_{ij} – is ranked as high as possible.

A straightforward solution to this task involves training a text-to-image retrieval model on a dataset containing all identities from $\mathcal{G}$. Once trained, the model would contain all the entity-specific knowledge in its weights (Figure 2a). However, this would require re-training the model for each gallery to transfer knowledge of specific person identities into the model weights. While feasible, this is impractical for real-world applications where users may want to fine-tune the system with their own galleries, such as smartphone photo collections.

A much more application-friendly formulation of this problem would use a single general-purpose trained model that adapts, at inference time, to any user-provided gallery of identities $\mathcal{G}$. The idea is to leverage the general knowledge that cross-modal models like CLIP [22] have about the world – which helps in understanding generic image-text correspondences – together with the domain-specific knowledge present in the user-provided gallery $\mathcal{G}$ – necessary to distinguish between different identities. To achieve this, during training, we remove the person's name from the query C_{ij}, substituting it with its face instead. This can be done by extracting the person's name N_i from C_{ij} using off-the-shelf named entities recognition methods and using it to look up the corresponding face F_i in $\mathcal{G}$. This is exactly the idea behind Id-CLIP, presented in the next section and summarized in Figure 2b.

4.1 Identity-aware CLIP (Id-CLIP)

In this work, we adapt existing CLIP-based models to search for relevant images given a *compound multimodal query* $(\tilde{C}, F)$, where $\tilde{C}$ is an anonymized caption and F represents an image of a person's face. We call this adaptation *Identity-aware CLIP(Id-CLIP)*, which extends the CLIP-PAD architecture from [12] by incorporating a *visual prompt tuning* stage. This enhancement serves as a targeted, non-invasive fine-tuning strategy for the CLIP vision encoder, allowing the model to better focus on person identities.

Overview Id-CLIP is inspired by the work in [12], in which the authors produce an anonymized caption $\tilde{C}$ by substituting the person name in C with a special token called [TOK]. This special token is not associated with a learnable embedding, as is the case with normal word tokens. Instead, its embedding $\mathbf{w}^{[\text{TOK}]}$ is obtained as a multi-layer perceptron (MLP) projection of the visual embedding extracted from the person's face F through a state-of-the-art face recognition network (FRN): $\mathbf{w}^{[\text{TOK}]} = \text{MLP}(\text{FRN}(F))$. The whole network is then subject to the same CLIP contrastive learning objective used to pull relevant image-text pairs and repel irrelevant ones. The network is trained using a suitable dataset $\mathcal{D}^{\text{train}}$ containing images of persons in specific contexts. During training, the CLIP-PAD network only updates the MLP projection network, while the rest of the CLIP weights are kept frozen. The MLP network learns to map the facial features into an actual word token that the frozen pre-trained text encoder can process alongside the other caption tokens. The training phase's sole purpose is to learn meaningful semantic mapping between the face feature space and the CLIP token embedding space. Therefore, we obtain the caption embedding $\mathbf{c}$ by applying the CLIP text encoder to the sequence of tokens containing the special token $\mathbf{w}^{[\text{TOK}]}$ together with the original sentence tokens $\mathbf{w}_i$: $\mathbf{c} = \text{CLIP}^{\text{txt}}([\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}^{[\text{TOK}]}, \ldots, \mathbf{w}_n])$. Notice that $\mathbf{w}^{[\text{TOK}]}$ is inserted in specific positions inside the caption $\tilde{C}$, as better explained in Section 5.

Visual Prompt Tuning The rationale behind Id-CLIP is that the person's identities are discriminated solely through their visual features, both at the query and image encoding stages. At query encoding time, the model can just leverage learning a suitable MLP projection network to discriminate between two different queries that carry the same context but different identities – relying on the discriminative ability of the off-the-shelf face recognition network. From the image dataset encoding perspective, instead, CLIP-PAD assumes that the visual encoder is already able to discriminate between different identities, assuming equal context. Considering that CLIP has not been directly trained to recognize fine-grained facial differences and that this information is not present in the generic image-text training datasets, it is likely that the CLIP visual encoder fails in such a task. In order to compensate for this potential drawback, we enrich the visual encoder of Id-CLIP with some additional learnable visual tokens $\tilde{\mathbf{v}}_i$, employing the approach presented in [11]. These additional visual tokens are purposed to

learn the discriminative characteristics of the person's faces. Specifically, the final image feature $\mathbf{v}$ is obtained through the application of the CLIP visual encoder to the original image tokens $\{\mathbf{v}_i\}_{i=1}^{wh}$ concatenated to the series of the special learnable prompt tuning tokens $\{\tilde{\mathbf{v}}_i\}_{i=1}^{p}$: $\mathbf{v} = \text{CLIP}^{\text{vis}}([\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_p, \mathbf{v}_1, \ldots, \mathbf{v}_{wh}])$, where p is the number of prompt tuning tokens (we empirically found that $p = 5$ works best in our scenario), and wh is the total number of visual patches of the image. Notice that, in this setup, the MLP and the added visual tokens are learned together during the training procedure, while the core CLIP weights remain frozen.

Objective As our image-text contrastive objective, we employ an Info-NCE loss [19], formulated as follows:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\text{sim}(\mathbf{c}_i, \mathbf{v}_i))}{\sum_{j=1}^{B}\exp(\text{sim}(\mathbf{c}_i, \mathbf{v}_j))}, \tag{1}$$

where $\mathbf{c}_i$ and $\mathbf{v}_j$ are the textual and visual features in output from CLIP, as explained in the previous paragraphs, $\text{sim}(\cdot, \cdot)$ is the cosine similarity, and B is the batch size employed during training.

5 Experiments

We evaluate the performance of Id-CLIP mainly on the targeted identity-aware cross-modal retrieval task, which we refer to as the *entity-in-context retrieval* – that focuses on retrieving images that contain both the requested person and the specified context. However, we can probe the same model also on a second related task, which we call *entity-only retrieval*. This secondary task aims at finding all images depicting a given entity, regardless of the specific context in which they appear. This second task is mainly crafted to understand the abilities of the trained model to actually understand the person's features without the potential noise introduced by the context.

Inference Strategies We apply two different inference strategies depending on the task we are tackling, i.e., *entity-in-context* retrieval or *entity-only* retrieval.

Concerning the entity-in-context retrieval subtask, our proposed COCO-PFS dataset already comes with the anonymized captions $\tilde{C}_{ij}$, which contain the [ENTITY] placeholder instead of the actual name[5]. In the most simple case, [ENTITY] could be substituted with the [TOK] token expected by Id-CLIP. However, other meaningful strategies may be applied, such as replacing [ENTITY] solely with the original text name (e.g., *Gianni Morandi*), or a combination of textual and visual features – like *[TOK] Gianni Morandi*, or *Gianni Morandi*

[5] In a real scenario, we should use NER techniques to extract the entity from the query. However, in this setup, we already have this information available considering how we built the dataset.

[TOK]. Other strategies are possible, like placing the visual feature at the beginning of the string and then substituting [ENTITY] to the sole textual name – e.g., *The famous kid [ENTITY] is running at the park* may become *[TOK]. The famous kid Gianni Morandi is running at the park.*. For entity-in-context retrieval, we consider each caption $\tilde{C}_{ij}$ as a separate query with which the whole image database is searched, and we tested different templates to substitute [ENTITY] with [TOK] and the original person's name.

Differently, for entity-only retrieval, the sole query is the name of the person to be searched. In this context, we treat each person as a different categorical class. We can solve this task using Id-CLIP by employing [ENTITY] as the sole input to the model. Since we are interacting with a language model that understands natural language sentences, we surround the [ENTITY] tag with some pre-defined templates, like *An image with [ENTITY]* or *The famous [ENTITY]*, performing *prompt ensembling* like in CLIP [22]. [ENTITY] is then expanded in the same way as for the entity-in-context retrieval, using a combination of [TOK] and the original person's name.

5.1 Metrics

In the experiments, we report the standard metric commonly employed to evaluate the cross-modal retrieval models, the average recall@k over all the tested queries, varying $k \in \{1, 5, 10, 50\}$.

$$\text{Recall@k} = \frac{|\text{Relevant instances} \cap \text{Retrieved top-k instances}|}{\min(k, |\text{Relevant instances}|)} \tag{2}$$

In the case of *entity-in-context retrieval*, we have only one relevant image in the dataset for each query text. Therefore, the average recall@k corresponds to the percentage of text queries that successfully retrieve the correct image containing the appropriate entity in the right context within the top k results.

We also assess model performance on the *entity-only retrieval* task, where the query text includes only the entity's name without any context. In this case, the recall@k measures the percentage of items in the top-k results that match the specified identity, regardless of the image's context.

We also report Rsum $= \sum_{k \in \{1,5,10,50\}}$ Recall@k as an aggregated measure.

5.2 Implementation Details

We optimized Id-CLIP using the Adam optimizer. We employed a learning rate of 5e-5. The MLP used to convert face features in a textual token is composed of one hidden layer without bias weights and ReLU non-linearities. The model is trained for a maximum of 10 epochs and validated on the main entities-in-context retrieval task. During training, as a robust augmentation strategy, we employ the face crops of the face-swapped persons instead of the original ones as F. During testing, instead, F is a real face taken from $\mathcal{G}$. Concerning captions, at training time, a single template is chosen randomly among all the available ones. This

Table 1: Entity-in-context Retrieval Results

		Recall@k				
Model	VPT	k=1	k=5	k=10	k=50	Rsum
CLIP (Original)	–	12.10	56.20	69.10	92.20	229.60
CLIP N	–	26.20	58.20	70.10	91.80	246.30
CLIP-PAD T (All $\mathbf{T}_i$)	–	36.30	66.60	77.80	95.20	275.90
CLIP-PAD T+N (All $\mathbf{T}_i$)	–	36.40	65.00	75.80	94.40	271.50
CLIP-PAD T ($\mathbf{T}_1$)	–	39.30	69.70	80.80	96.50	286.20
CLIP-PAD T+N ($\mathbf{T}_1$)	–	38.00	67.10	78.30	95.50	278.90
CLIP (Original)	✓	12.40	57.90	70.60	93.20	234.10
CLIP N	✓	27.90	61.70	72.90	93.00	255.50
Id-CLIP T (All $\mathbf{T}_i$)	✓	39.20	68.20	78.50	95.70	281.60
Id-CLIP T+N (All $\mathbf{T}_i$)	✓	37.50	66.10	76.30	94.30	274.10
Id-CLIP T ($\mathbf{T}_1$)	✓	**43.20**	**71.20**	**80.80**	**96.70**	**291.90**
Id-CLIP T+N ($\mathbf{T}_1$)	✓	39.90	68.60	78.10	95.00	281.60

can be considered as an additional text augmentation procedure, which enables the network to generalize to different phrasings of the same (context, identity) pair, in turn stabilizing the learning procedure.

5.3 Results

For the entity-in-context retrieval, we report the results in Table 1, which is split into two equal sets of experiments: one without Visual Prompt Tuning (VPT) and one with VPT engaged. The first and the second lines report the two trivial baselines for this task, which are represented by the original CLIP prompted with the original caption and by the original CLIP prompted with COCO-PFS captions containing the person names (CLIP N).

By focusing on the recall@1 metric, we can notice how CLIP is already able to exploit some biases when prompted with the person's names – most likely, certain names imply specific ethnical groups, which imply certain facial features. We then first re-evaluate CLIP-PAD (which does not have visual prompt tuning active) on our benchmark, considering the most successful inference strategies. The performance improves significantly with CLIP-PAD (All T_i), which averages recall across 11 different templates evaluated independently, resulting in an improvement of more than 38% compared to CLIP N. The two different reported versions, T and T+N, relate to how [ENTITY] is substituted – with only [TOK] in the first case and with [NAME] [TOK] in the second one, where the name is the person's name. Interestingly, CLIP-PAD seems not to improve as much as the CLIP baselines when presented with original names. In fact, the difference between CLIP-PAD T and CLIP-PAD T+N is quite marginal with respect to the huge improvement observed in the baselines. This may be due to the ethical biases present in the pre-trained CLIP transforming into unuseful noise when the

Table 2: Entity-only Retrieval Results. In configuration (1), [ENTITY] is expanded as "[NAME] [TOK]", while in (2), the sentence always starts with "An image with [TOK]." and is then followed by the original caption where [ENTITY] is substituted with [NAME].

		Recall@k				
Model	VPT	k=1	k=5	k=10	k=50	Rsum
CLIP N	–	12.80	6.00	8.40	17.50	44.80
CLIP-PAD T	–	15.40	11.10	13.50	28.00	68.00
CLIP-PAD T+N (1)	–	20.40	11.80	13.90	28.00	74.10
CLIP-PAD T+N (2)	–	21.80	11.80	14.20	28.20	75.90
CLIP N	✓	15.40	7.30	9.20	19.10	51.00
Id-CLIP T	✓	21.20	11.50	14.20	**31.90**	78.90
Id-CLIP T+N (1)	✓	22.60	11.60	14.60	29.70	78.60
Id-CLIP T+N (2)	✓	**24.30**	**13.70**	**16.20**	31.40	**85.60**

facial feature is injected through the [TOK] token. We also report CLIP-PAD (T_1) – coming as well in the T and in T+N configurations, which, differently from CLIP-PAD (All T_i), only tests the best-performing template, which is T_1 = *"[ENTITY] in the image. [CAPTION]"*. We can notice that the best template behaves particularly better than the "All T_i" configuration, meaning that, on average, some reasonable query behaves particularly worse than the average. Shifting the attention to the results where visual prompt tuning is active, we can appreciate how much our proposed Id-CLIP improves over CLIP-PAD for all the probed inference schemas. In particular, Id-CLIP obtains an improvement on recall@1 of 55% with respect to CLIP N, with an average increase over CLIP-PAD of 6%, proving the effectiveness of this targeted fine-tuning strategy for the visual encoder. Notice also that the improvement on the recall@k for higher values for k is limited given that, at a certain point, the model will have retrieved all the pictures of the different identities in a given context[6], thus in the limit behaving similarly to the original CLIP model. We obtain the best overall results over all the metrics with the Id-CLIP T (T_1) model.

Concerning the entity-only retrieval, we report the results in Table 2. Again, we can immediately notice how the original CLIP model prompted with names (CLIP N) behaves poorly on this subtask. Nevertheless, differently from the results in entity-in-context retrieval, we can notice how the original name plays a more important role, as far as CLIP-PAD is concerned. In fact, the configurations denoted as CLIP-PAD T+N obtain a significant improvement over CLIP-PAD T, meaning that the name biases present in CLIP help in bringing reasonable identities to the top of the ranked list. This effect is actually damped in Id-CLIP, where visual prompt tuning effectively increases the importance of the

[6] We remind that, in our dataset, each image has been swapped with multiple identities, and this applies to the test set as well.

visual features over the textual biases. For completeness, we report two different [ENTITY] expansion strategies: T+N (1) represents the case in which [ENTITY] is expanded as "[NAME] [TOK]"; (2) represents the case where the sentence is always started with "An image with [TOK]." and then followed by the original caption where [ENTITY] is substituted with only [NAME]. The difference between the two versions is not so pronounced, meaning that the model seems robust to different entity expansion formulations. Even in this case, we reach the best result through Id-CLIP, again proving the importance of VPT. Specifically, the best-performing model results to be Id-CLIP T+N (2), which increases the recall@1 by 57% with respect to CLIP N and by 11% with respect to the corresponding configuration of CLIP-PAD.

6 Conclusions

In this work, we tackled the *identity-aware cross-modal retrieval* task, introducing a dataset and some strong baselines that are designed to avoid a deep fine-tuning of the model on specific identity galleries. More in detail, we first introduced the COCO-PFS dataset, obtained by replacing the faces in the COCO dataset with controlled, identity-specific entities from VGGFace2. We have shown that this task is particularly challenging for existing multimodal models like CLIP, which struggle to handle domain-specific entities and long-tail concepts such as unique identities not present in the training data. As a second step, we also proposed Id-CLIP, a strong baseline that demonstrated significant improvements over standard CLIP models. In particular, we showed that while original CLIP can leverage biases related to names and ethnic groups, it is still limited in its ability to accurately retrieve images that both match the identity and the context described by the text query. Our experiments revealed that modifying CLIP augmenting the query with the face features of the person taken from an external gallery and incorporating a visual prompt tuning leads to notable enhancements in both identity and context retrieval. As a future work, we plan to explore different pre-trained CLIP models and to augment Id-CLIP to actively weigh the importance of context and identity directly at inference time.

Acknowledgments. This work was partially funded by MUCES - a MUltimedia platform for Content Enrichment and Search in audiovisual archive project (CUP B53D23026090001), and by PNRR-M4C2 (PE00000013) "FAIR-Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded under the NextGeneration EU program.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Avrahami, O., Aberman, K., Fried, O., Cohen-Or, D., Lischinski, D.: Break-a-scene: Extracting multiple concepts from a single image. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–12 (2023)

2. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). pp. 67–74. IEEE (2018)
3. Chen, Z., Zhang, L., Weng, F., Pan, L., Lan, Z.: Tailored visions: Enhancing text-to-image generation with personalized prompt rewriting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7727–7736 (2024)
4. Cohen, N., Gal, R., Meirom, E.A., Chechik, G., Atzmon, Y.: "this is my unicorn, fluffy": Personalizing frozen vision-language representations. In: European conference on computer vision. pp. 558–577. Springer (2022)
5. Dost, S., Serafini, L., Rospocher, M., Ballan, L., Sperduti, A.: Jointly linking visual and textual entity mentions with background knowledge. In: Natural Language Processing and Information Systems: 25th International Conference on Applications of Natural Language to Information Systems, NLDB 2020, Saarbrücken, Germany, June 24–26, 2020, Proceedings 25. pp. 264–276. Springer (2020)
6. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improving visual-semantic embeddings with hard negatives (2018), https://github.com/fartashf/vsepp
7. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023 (2023), https://openreview.net/pdf?id=NAQvF08TcyG
8. Gal, R., Arar, M., Atzmon, Y., Bermano, A.H., Chechik, G., Cohen-Or, D.: Encoder-based domain tuning for fast personalization of text-to-image models. ACM Transactions on Graphics (TOG) **42**(4), 1–13 (2023)
9. Gatti, P., Penamakuri, A.S., Teotia, R., Mishra, A., Sengupta, S., Ramnani, R.: Cofar: Commonsense and factual reasoning in image search. arXiv preprint arXiv:2210.08554 (2022)
10. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
11. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision. pp. 709–727. Springer (2022)
12. Korbar, B., Zisserman, A.: Personalised clip or: how to find your vacation videos (2022)
13. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)
14. Li, K., Zhang, Y., Li, K., Li, Y., Fu, Y.: Visual semantic reasoning for image-text matching. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4654–4662 (2019)
15. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16. pp. 121–137. Springer (2020)
16. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)

17. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems **32** (2019)
18. Messina, N., Stefanini, M., Cornia, M., Baraldi, L., Falchi, F., Amato, G., Cucchiara, R.: Aladin: Distilling fine-grained alignment scores for efficient image-text matching and retrieval. In: International Conference on Content-based Multimedia Indexing. pp. 64–70 (2022)
19. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
20. Pang, L., Yin, J., Xie, H., Wang, Q., Li, Q., Mao, X.: Cross initialization for face personalization of text-to-image models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8393–8403 (2024)
21. Qu, L., Liu, M., Cao, D., Nie, L., Tian, Q.: Context-aware multi-view summarization network for image-text matching. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 1047–1055 (2020)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/radford21a.html
23. Ramanathan, V., Joulin, A., Liang, P., Fei-Fei, L.: Linking people in videos with "their" names using coreference resolution. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. pp. 95–110. Springer (2014)
24. Rosasco, A., Berti, S., Pasquale, G., Malafronte, D., Sato, S., Segawa, H., Inada, T., Natale, L.: Concon-chi: Concept-context chimera benchmark for personalized vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22239–22248 (2024)
25. Sarafianos, N., Xu, X., Kakadiaris, I.A.: Adversarial representation learning for text-to-image matching. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5814–5824 (2019)
26. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565 (2018)
27. Stefanini, M., Cornia, M., Baraldi, L., Cucchiara, R.: A novel attention-based aggregation function to combine vision and language. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 1212–1219. IEEE (2021)
28. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vl-bert: Pre-training of generic visual-linguistic representations. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=SygXPaEYvH
29. Sun, W., Fan, Y., Guo, J., Zhang, R., Cheng, X.: Visual named entity linking: A new dataset and a baseline. arXiv preprint arXiv:2211.04872 (2022)
30. Tewel, Y., Gal, R., Chechik, G., Atzmon, Y.: Key-locked rank one editing for text-to-image personalization. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)
31. Vadicamo, L., Arnold, R., Bailer, W., Carrara, F., Gurrin, C., Hezel, N., Li, X., Lokoc, J., Lubos, S., Ma, Z., et al.: Evaluating performance and trends in interactive video retrieval: Insights from the 12th vbs competition. IEEE Access (2024)

32. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., Wei, F.: Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
33. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15943–15953 (2023)
34. Wen, K., Gu, X., Cheng, Q.: Learning dual semantic relations with graph attention for image-text matching. IEEE transactions on circuits and systems for video technology **31**(7), 2866–2879 (2020)
35. Yeh, C.H., Russell, B., Sivic, J., Heilbron, F.C., Jenni, S.: Meta-personalizing vision-language models to find named instances in video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19123–19132 (2023)
36. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
37. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014)
38. Zeng, Y., Patel, V.M., Wang, H., Huang, X., Wang, T.C., Liu, M.Y., Balaji, Y.: Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6786–6795 (2024)
39. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters **23**(10), 1499–1503 (2016)
40. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5579–5588 (2021)
41. Zheng, Q., Wen, H., Wang, M., Qi, G.: Visual entity linking via multi-modal learning. Data Intelligence **4**(1), 1–19 (2022)