FaceCPT: Toward Cross-Modal Facial Representation Learning with Face-Caption Pre-Training

Md Mahedi Hasan
mh00062@mix.wvu.edu
Shoaib Meraj Sami
sms00052@mix.wvu.edu
Nasser Nasrabadi
nnasraba@gmail.com
Jeremy M. Dawson
jeremy.dawson@mail.wvu.edu

Lane Department of Computer Science and Electrical Engineering
West Virginia University
Morgantown, WV, US

Abstract

Facial representation learning (FRL) through weakly-supervised pre-training has shown significant promise across various downstream tasks, highlighting its improved generalizability. However, most existing FRL models excel only in face generation and analysis, and there is a lack of models that address cross-modal tasks. To fill this gap, we pose the following question: Can we learn a universal facial representation by pre-training on web-sourced face-caption pairs for surveillance-related tasks? These tasks include both cross-modal understanding, such as face captioning and text-based face image retrieval, as well as face analysis tasks like face, attribute, expression recognition, and age estimation. In this paper, we take a step toward this objective by introducing **FaceCPT**, a new framework for learning facial representation using **Face C**aption **P**re-**T**raining. However, domain misalignment and information asymmetry between image-text pairs challenge the model's ability to achieve a meaningful interaction. To overcome this, we utilize contrastive learning along with a semantic attribute-aware loss (SAAL) to improve the semantic associations between face-caption pairs and encourage the model to focus on key semantic attributes, respectively. Experiments show that FaceCPT outperforms existing vision-language pre-training and FRL baselines, achieving state-of-the-art results in task-specific fine-tuning and improved zero-shot transferability across both single-modal and cross-modal tasks, even with low-resolution inputs.

1 Introduction

In real-world scenarios, tasks like face recognition (FR) [6, 20, 52], facial attribute recognition (FAR) [32, 44, 45, 46], face image retrieval [8, 56], and face captioning [13, 27] are pivotal in smart surveillance systems, border security, and digital forensics [17]. Improved face recognition in unconstrained, low-resolution images is vital for efficiently identifying suspects [3]. Recognizing facial attributes such as gender, ethnicity, age, and accessories

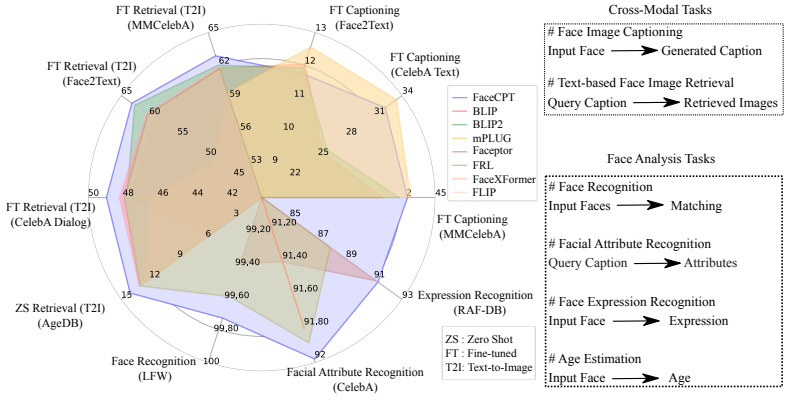

Figure 1: Proposed FaceCPT excels in surveillance-related tasks, such as face analysis and cross-modal understanding, showing strong zero-shot capability and flexible task adaptability.

improves the accuracy of individual tracking and retrieval. Moreover, face captioning can describe specific facial characteristics, aiding in detecting suspicious behavior, such as attempts to cover faces with masks or sunglasses [27]. Additionally, face retrieval plays a crucial role in identifying suspects during criminal investigations. However, real-world scenarios such as occlusion and low light often make it challenging to obtain reliable visual query images, making traditional image-based retrieval systems less effective. Similarly, sketch-based retrieval [36] faces limitations, as it heavily depends on the artist's skill and the accuracy of witness recollections [4]. In contrast, text-based face retrieval [8] provides a practical alternative by enabling searches using natural language without the need for a query image. For example, when a suspect is observed but not captured by the surveillance, witness descriptions of facial characteristics can greatly refine search results in a large database.

The traditional approach to face-related tasks typically relies on supervised, task-specific training with manually annotated datasets [30, 59], which is both labor-intensive and time-consuming [39, 57]. A promising alternative is to develop a unified model that learns general facial representation through weakly supervised pre-training. This approach offers several advantages over task-specific methods: it eliminates the need for separate training per task, significantly reducing computational time and cost. It also improves zero-shot transferability and task adaptation, reducing dependence on costly hand-labeled data [2, 35]. Furthermore, training on web-sourced data enhances the model's robustness to input variations, thereby improving its *adaptability* to real-world scenarios [15].

Recently, vision-language pre-training (VLP) on web-scale image-text pairs has gained momentum, driven by the success of multimodal foundation models. These models learn visual representation through textual supervision and show improved performance across various tasks in both *few-shot* and *zero-shot* settings. While effective at learning general representation, their performance often suboptimal in specialized domains such as face analysis. Meanwhile, several self- and weakly-supervised facial representation learning (FRL) methods [2, 27, 35, 47, 54] have been proposed for tasks such as face analysis and generation. However, the development of FRL methods for cross-modal understanding tasks—such as face captioning and text-based face retrieval, which are crucial for surveillance—remains largely overlooked. Pre-training on noisy, web-sourced face-caption pairs is challenging due to the substantial domain gap between visual and semantic modalities, leading to semantic

misalignment. Moreover, in many state-of-the-art (SOTA) datasets, captions for identical faces can vary widely, reflecting the subjective interpretations of annotators. In addition, these captions are often abstract, describing only a limited set of facial attributes while missing crucial details [13]. This creates *information asymmetry* between the faces and their corresponding captions, as the captions fail to capture the intricate details of the face images.

In this paper, we introduce FaceCPT, a contrastive pre-training framework that overcomes these challenges. Our contributions are as follows:

1. We present FaceCPT, an end-to-end pre-training framework for learning general facial representation that effectively tackles diverse surveillance-related tasks, including face analysis and cross-modal understanding—an area largely overlooked in prior studies.

2. We establish a new benchmark for face image retrieval using textual queries. FaceCPT's robustness for this task is demonstrated through its performance on distorted images, as well as its effective handling of scenarios with missing or incomplete queries.

3. We demonstrate the effectiveness of FaceCPT in both zero-shot transferability and task-specific adaptability, outperforming state-of-the-art VLP and FRL baselines in face analysis and cross-modal tasks, even at lower input resolutions.

2 Related Works

Facial Representation Learning (FRL): Supervised FRL methods [11] typically depend on vast amounts of hand-labeled data for pre-training and often struggle to generalize to downstream tasks due to their overspecialization for the source task. Bulat et al. [2] proposed a large-scale unsupervised pre-training model using in-the-wild face images, which is then fine-tuned for tasks like FR and landmark localization. The authors of [54] introduced MCF, a self-supervised pre-training framework that combines patch-based contrastive learning with masked image modeling to enhance representation learning. It also incorporates knowledge distillation from another pre-trained model. Similarly, Gao et al. [12] introduced FRA, which learns consistent local features using region-aware contrastive learning. Utilizing ResNet [14] as backbone, FRA excels in various face-related classification and regression tasks. However, self-supervised pre-trained models like FRA often show poor zero-shot performance and typically require extensive fine-tuning for task-specific adaptation.

Models such as FaRL [57] and Talk2Face [27] learn general facial representation by pre-training the backbone on large-scale image-text pairs using techniques like contrastive learning and masked image/text modeling, followed by fine-tuning for specific downstream tasks. In particular, FaRL [57] leverages face-caption pre-training by using contrastive learning [18, 40] to capture high-level semantics and masked image modeling [1] to extract low-level details. Talk2Face [27] introduces a sequence-based framework capable of handling a wide array of face analysis and generative tasks, including facial expression recognition and face image captioning. The framework incorporates a sequence modeling objective, reformulating these tasks into a sequence-to-sequence format. However, these models often learn generic facial representation that fail to capture crucial details, leading to less discriminative features and suboptimal performance in face analysis tasks. Additionally, due to pre-training on low-quality pairings in web-sourced data, these weakly-supervised methods struggle to achieve enhanced semantic alignments, which degrades performance on tasks requiring fine-grained understanding, such as face perception.

Recently, unified face perception frameworks [35, 39, 47] have emerged that use shared representations and multi-task learning to handle multiple tasks without task-specific fine-tuning. For example, FaceXFormer [35] employs a transformer-based encoder-decoder to jointly process face and task tokens for robust representation learning. Faceptor [39] utilizes a multi-task learning approach to improve task extensibility, enabling the sharing of representation and knowledge across various tasks. The authors of Q-Face [47] proposed a unified model for multiple face analysis tasks, employing a query-driven decoder with cross-attention to adaptively extract features by capturing correlations among related tasks. However, these models are primarily designed for single-modal tasks where they often demonstrate representational biases toward certain attributes due to imbalanced training data. Extending these models to handle cross-modal tasks remains an open area of research.

Multimodal Face-Caption Datasets: Existing datasets, such as Multi-Modal CelebA-HQ (MMCelebA)[65], Face2Text[49], CelebA-Dialog [19], CelebA-Text [48], and FFHQ-Text [58], have only a limited number of high-quality face-caption pairs, making them inadequate for training FRL models. Although Talk2Face [27] includes 2.3M face-caption pairs, it remains private. Recently, LAION-Face [57], a 50M face-caption pair dataset, was curated from LAION-400M [42] dataset. However, its captions lack detailed facial descriptions, leading to weak correlations between captions and corresponding faces. To address this, FaceCaption-15M [6] was developed as an enhanced version of this dataset with stronger image-text correlations. The authors filtered out non-human faces by applying RetinaFace [7] to crop facial regions and employed an attribute recognition algorithm to generate high-quality captions. Furthermore, FLIP-80M [28], a large-scale dataset consisting of 80M web-crawled face-caption pairs, was introduced. However, in its raw form, it is unsuitable for FRL tasks, as most pairs include non-face images with complex backgrounds.

3 Proposed Method

3.1 FaceCPT Architecture

As illustrated in Figure 2, FaceCPT contains two unimodal encoders for encoding input pairs, a cross-modal encoder for image-text interaction, and a decoder for caption generation. In this study, we employ ResNets [14] as our image encoder for its efficiency in extracting high-quality facial representation from low-resolution images. The encoder extracts a global feature vector $\bar{\boldsymbol{f}}_g \in \mathbb{R}^{512}$ along with local features $\boldsymbol{L}_3 \in \mathbb{R}^{256\times14\times14}$ and $\boldsymbol{L}_2 \in \mathbb{R}^{128\times14\times14}$ from the third and second residual blocks, respectively. These features pass through a visual adapter (VA) to produce the final image representation $\boldsymbol{f}_g \in \mathbb{R}^{768}$.

For the text encoder, we utilize a 6-layer Transformer with 12 attention heads [51], which generates output embeddings $\boldsymbol{W} \in \mathbb{R}^{768\times T}$ for up to T input tokens, applying padding or truncation as needed. A learnable $[CLS]$ token is prepended to the input tokens to produce the caption embedding, $\boldsymbol{f}_c \in \mathbb{R}^{768}$. Initially, $\boldsymbol{f}_c$ and $\boldsymbol{f}_g$ are misaligned due to the inherent heterogeneity and information asymmetry between the image and text modalities. To resolve this, we apply the image-text contrastive (ITC) loss to learn global semantic associations between these two modalities. Next, the image and textual representations interact through another 6-layer, 12-head cross-modal encoder, which employs cross-attention to effectively fuse information between modalities, generating a rich cross-modal representation $\boldsymbol{f}_{CM}$.

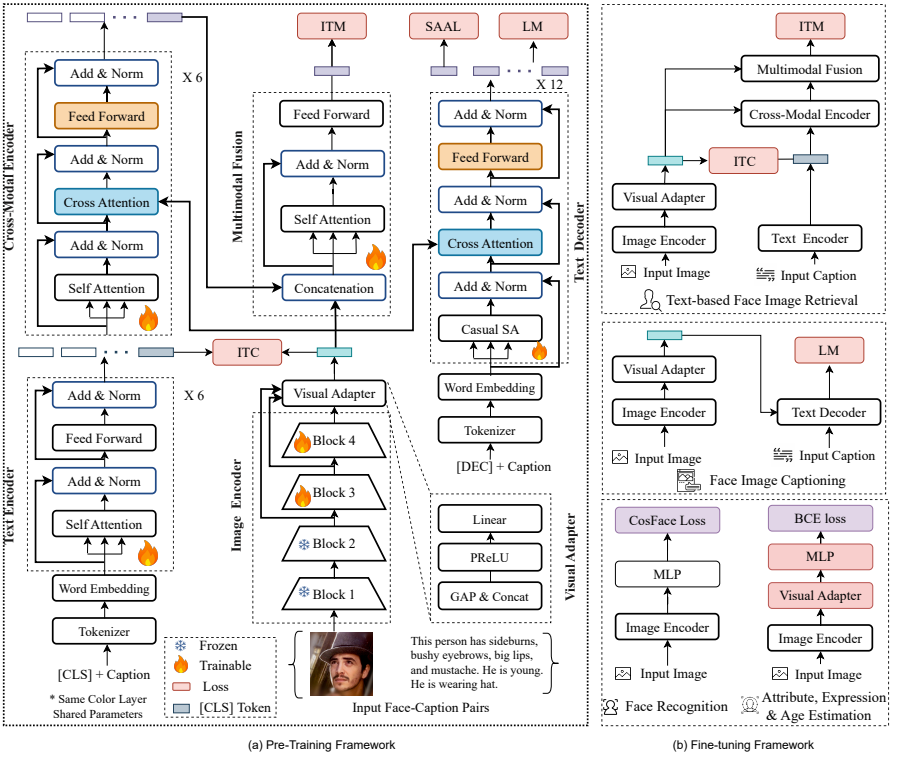

Figure 2: (a) Illustration of the pre-training framework of FaceCPT. It employs two unimodal encoders to encode input pairs and enhance their deep semantic associations using an image-text contrastive (ITC) loss. (b) Fine-tuning frameworks for downstream tasks.

Finally, we concatenate $\boldsymbol{f}_{CM}$ with $\boldsymbol{f}_g$ within a multimodal fusion block to compute the image-text matching (ITM) loss. This loss enhances the model's ability to learn image-text semantic relationships, improving its performance in cross-modal tasks. Given the inherent information asymmetry between face-caption pairs, fusing $\boldsymbol{f}_{CM}$ with $\boldsymbol{f}_g$ has proven effective. Our fusion block, consisting of a self-attention and feed-forward layer, generates the multimodal embedding $\boldsymbol{f}_{MF}$, which is then fed into the ITM head for image-text matching.

3.2 Pre-training Tasks

Image-Text Contrastive (ITC) loss is a widely adopted alignment loss [18, 21, 24, 40] that improves semantic associations between the image representation $\boldsymbol{f}_g$ and the text representation $\boldsymbol{f}_c$. Its primary objective is to minimize the distance between positive pairs while maximizing the distance for negative pairs, which are derived from non-matching pairs within the batch. To implement this loss, we optimize the similarity score, $s(I,T) = g_i(\boldsymbol{f}_g) \cdot g_t(\boldsymbol{f}_c)$, for the normalized feature representations of the face-caption pair (I,T). Here, $g_i(\boldsymbol{f}_g) = \boldsymbol{f}_g^\top \boldsymbol{w}_I$ and $g_T(\boldsymbol{f}_c) = \boldsymbol{f}_c^\top \boldsymbol{w}_T$ denote linear projections that map image and text representations into a shared space, where $\boldsymbol{w}_I, \boldsymbol{w}_T \in \mathbb{R}^{768\times256}$, enabling alignment with reduced complexity and faster convergence.

Method	Visual Backbone	Res	MMCelebA [53]				Face2Text [49]				CelebA-Text [48]			
			B@4	M	RL	C	B@4	M	RL	C	B@4	M	RL	C
Vision-Language Pre-training (VLP)														
BLIP-base [24]†	ViT-B/16	384^2	40.43	61.0	56.05	27.48	12.04	41.0	40.85	22.65	25.53	48.65	45.78	30.77
BLIP2-OPT$_{2.7B}$ [25]†	ViT-g/14	364^2	41.90	54.98	47.72	26.69	11.92	39.79	34.13	22.92	26.02	49.54	46.71	23.24
OFA-base [53]†	ResNet101	480^2	39.40	60.56	54.85	26.42	12.01	41.10	41.85	22.45	25.65	49.79	45.80	29.40
mPLUG-base [21]†	ViT-B/16	384^2	**42.87**	27.93	54.75	**92.09**	**12.56**	23.03	40.35	**57.54**	28.38	23.14	41.73	**55.48**
Facial Representation Learning (FRL)														
Talk2Face [2]†	-	256^2	39.58	55.61	52.38	34.71	11.50	28.50	40.62	26.90	**33.40**	28.40	53.50	40.0
Ours (FaceCPT-R50)	ResNet50	112^2	40.54	61.60	57.50	30.05	11.50	41.50	41.86	22.90	31.38	51.98	51.74	30.80
Ours (FaceCPT-R101)	ResNet101	112^2	42.60	**62.15**	**58.90**	32.08	11.76	**42.30**	**42.20**	23.25	32.20	**53.18**	**53.92**	35.69

Table 1: **Overview of the FaceCPT performance on face captioning** across three state-of-the-art face-caption datasets. † indicates reproduction using the official codebase.

Image-Text Matching (ITM) is a binary classification loss that trains the model to learn joint image-text representation by distinguishing between matching and mismatched face-caption pairs. Optimizing this objective improves the model's ability to discriminate between pairs based on fine-grained details. It also provides a similarity score, which is crucial for ranking retrieved images according to their relevance to the query. Following the approach in [23], hard negative mining is used to identify more challenging in-batch negatives, pairs with higher contrastive similarity, thereby improving the robustness of the matching.

Language Modeling (LM) guides the text decoder to generate coherent captions. It employs cross-entropy loss to maximize the conditional likelihood of output tokens given the preceding tokens. The full pre-training objective of FaceCPT is as follows:

$$L = L_{ITC} + L_{ITM} + L_{LM}. \tag{1}$$

Semantic Attribute-Aware Loss (SAAL): Prior works often employ CIDEr-based reward optimization [21] during fine-tuning to enhance caption quality. However, captions generated through this approach tend to be generic, focusing only on common facial attributes and thus lacking distinctiveness. In this study, we replace CIDEr-based optimization with SAAL, a differentiable objective that maximizes alignment between facial attributes and those mined from the generated captions. SAAL penalizes the text decoder for misrepresenting key facial features, leading to more attribute-driven captions that enhance zero-shot face retrieval.

In this study, we apply SAAL during fine-tuning, as a multi-label cross-entropy (CE) loss to penalize FaceCPT for missing or incorrect attributes, leading to more accurate and distinctive captions. To address attribute imbalance—since only a few attributes are typically present per face—we employ a weighted CE loss that assigns higher weights to less frequent attributes. Let a_{ij} denotes the label of the j^{th} attribute. For predicted attributes, $a_{ij} = 1$ indicates the positive (*e.g.* man), and $a_{ij} = -1$ denotes the negative (*e.g.* woman) of the attribute in the input caption. Missing attributes are always marked as 0.

$$L_{SAAL}(I) = \frac{1}{N} \sum_i \sum_j -w_{a_j} log P(a_j \mid I_i). \tag{2}$$

4 Experiments

4.1 Downstream Tasks

Facial analysis tasks such as face recognition (FR), attribute recognition (FAR), and attribute estimation (AE) require learning discriminative identity representation, which contrasts with

Pre-trained Models	Resolution	CelebA-Dialog [19]		MMCelebA [65]		Face2Text [49]	
		R@5	R@10	R@5	R@10	R@5	R@10
Vision-Language Pre-training							
CLIP [40]	224^2	23.42	42.30	44.06	59.48	37.89	52.47
ALBEF [23]	384^2	31.38	47.35	47.51	61.34	47.22	59.54
BLIP-base [22]	384^2	32.09	47.56	48.08	61.75	45.20	59.34
mPLUG-large [21]	336^2	29.29	46.50	46.35	59.56	38.39	48.65
BLIP2-OPT$_{2.7B}$	364^2	32.04	47.78	47.90	62.02	48.58	62.60
Facial Representation Learning							
FaRL [57]	224^2	32.84	47.90	47.83	61.70	46.34	60.27
FLIP [28]	224^2	33.08	48.21	48.51	61.92	46.65	60.60
Ours(FaceCPT-R50)	112^2	32.36	48.44	49.50	62.02	45.86	59.48
Ours(FaceCPT-R101)	112^2	**33.48**	**48.94**	**50.01**	**62.88**	**48.76**	**63.12**

Table 2: Overview of FaceCPT performance in text-based face image retrieval.

Pre-trained Models	Res.	LFW [16]		Age-DB [34]	
		R@10	R@20	R@10	R@20
Vision-Language Pre-training					
BLIP-base [22]	384^2	4.89	8.27	5.56	8.51
mPLUG-large [21]	336^2	9.52	11.79	12.30	15.13
BLIP2-OPT$_{2.7B}$	364^2	7.85	12.63	12.88	15.92
Facial Representation Learning					
FaRL [57]	224^2	9.85	14.36	12.98	16.90
FLIP [28]	224^2	9.90	14.47	13.11	17.06
FaceCPT-R50	112^2	10.91	15.94	13.12	18.51
FaceCPT-R101	112^2	11.25	16.13	14.06	20.10
FaceCPT-R50	224^2	10.91	17.33	13.25	19.26
FaceCPT-R101	224^2	**11.56**	**16.90**	**15.21**	**21.46**

Table 3: Comparison of zero-shot transferability among FaceCPT and SOTA VLP and FRL baselines for text-to-face self-retrieval, with all methods fine-tuned on the MM-CelebA dataset.

Methods	Training	CelebA [30]	LFWA
DMM† [53]	FSL	91.70	86.56
SlimCNN† [44]		91.24	76.02
AFFAIR† [72]		91.45	86.13
SSPL† [45]	WSL	91.77	86.53
MCF* [52]		91.33	-
FRL* [1]		89.65	-
FRA* [10]		**92.20**	-
Faceptor [39]		91.39	-
FaceXFormer [35]		91.79	-
FaRL† [57]		91.88	86.69
Ours (FaceCPT-R50)		91.92	86.40
Ours (FaceCPT-R101)		92.12	**86.72**

Table 4: Comparison with various fully supervised (FSL), self-supervised (SSL), and weakly supervised (WSL) learning methods for FAR. Note that † and * indicate results taken from [57] and [10], respectively.

learning facial representation via face-caption matching. Nevertheless, FaceCPT needs only moderate fine-tuning to adapt its image encoder for discriminative representation. Face captioning is typically evaluated using BLEU-4 [37] (B@4), METEOR [9] (M), ROUGE-L [29] (RL), and CIDEr [51] (C). However, these metrics often overlook the true semantic quality of a distinctive caption, as they mainly rely on n-gram overlap. For example, a model may receive low scores even if it generates descriptive facial descriptions, simply because certain attributes are not present in the reference captions. This limitation poses a significant challenge for evaluating face captioning. In this study, we focus on generating descriptive captions to improve text-based face retrieval. We further enhance retrieval performance by optimizing FaceCPT with ITC and ITM contrastive losses, which ensure strong semantic alignment between face and caption representations. Furthermore, SAAL and the LM loss help to bridge the image-text modality gap by improving FaceCPT's semantic understanding.

4.2 Implementation Details

We pre-train our model end-to-end using FaceCaption-15M [5] and a refined version of FLIP-80M [28], processed using a method similar to that employed for FaceCaption-15M. We compare our approach with VLP baselines, including CLIP [40], ALIGN [18], AL-

Methods	Res	LFW	CFP-FP	AgeDB	CALFW	CPLFW
Cosface [52]	112^2	99.81	**98.12**	98.11	95.76	92.28
ViT* [10]	224^2	**99.83**	96.19	97.82	**95.92**	92.55
FaRL* [57]	224^2	99.60	96.70	95.55	95.38	92.38
Faceptor* [39]	112^2	99.40	96.34	93.65	94.75	92.27
FaceCPT-R50	112^2	99.68	96.88	98.01	95.58	92.42
FaceCPT-R101	112^2	99.73	97.09	**98.14**	95.66	**92.59**

Table 5: Comparison with FRL methods on benchmark FR evaluation datasets using 1:1 verification. All methods have been fine-tuned using the CosFace [52] loss to ensure a fair comparison. *: results are taken from [39]

Methods	RAF-DB [26]	AgeDB [34]
FaRL [57]	88.31	5.64
Faceptor [39]	91.26	4.96
SwinFace [38]	90.97	-
QFace [42]	**92.86**	4.67
FaceCPT-R50	91.42	4.91
FaceCPT-R101	91.82	**4.63**

Table 6: Comparison with SOTA FRL methods for AE on the AgeDB dataset using mean absolute error (MAE ↓) and for FER on the benchmark RAF-DB dataset using mean accuracy (↑).

Objectives	Retrieval-ZS		Caption-FT		
	R@5	R@10	B@4	M	RL
ITM+ITC	27.88	42.65	-	-	-
LM	-	-	37.72	52.48	48.14
ITM+ITC+LM	28.62	43.81	38.96	59.05	54.86
ITM+ITC+LM+VA	30.95	46.35	39.18	59.35	55.25
ITM+ITC+LM+VA+SAAL	-	-	40.54	61.60	57.50

Table 7: Analysis of different pre-training components of FaceCPT.

Encoders	Patch	Retrieval-ZS		Caption-ZS		
		R@5	R@10	B@4	M	RL
ViT-base	16^2	30.04	44.68	10.40	31.92	24.81
ViT-base	8^2	30.48	45.25	10.91	31.96	25.12
ViT-large	16^2	31.18	46.54	10.98	32.11	25.21
Ours	-	30.95	46.35	10.27	32.66	24.92

Table 8: Ablating different image encoder variants in FaceCPT

BEF [23], mPLUG [21], BLIP [24], and BLIP2 [25], all of which are fully fine-tuned on the respective task-specific datasets. Additionally, we evaluate FaceCPT against prevalent FRL methods—Talk2Face [27], FaRL [57], and FLIP [28].

The image encoder, initialized with ArcFace weights [6] pre-trained on MS1MV2 [6]. The text encoder and cross-modal encoder are initialized using the first and last six layers of the pre-trained BERT-base model [10], respectively. The text decoder, also a BERT-base model, shares some layers with the cross-modal encoder. To evaluate text-based face retrieval, we first select the top K candidates based on face-caption similarity, then re-rank them using their pairwise ITM scores. For all datasets, we set $K = 128$. All of the CelebA-based fine-tuning datasets provide 40 facial attributes [55] as ground truth to train our attribute prediction model, which is then used to predict attributes for unlabeled datasets like Face2Text [49]. During pre-training, FaceCPT is optimized using the AdamW [31] optimizer with a weight decay of 0.05. We use a batch size of 128 and begin with a learning rate of 1e-6. Additional details on datasets, baselines, and hyperparameters are provided in the supplementary materials.

4.3 Main Results

Evaluation on Cross-Modal Tasks: Table 1 compares FaceCPT with existing VLP baselines for face image captioning. FaceCPT, using a ResNet-101 encoder, outperforms other VLPs that rely on higher-resolution inputs and greater computational resources. In contrast, FaceCPT operates on lower-resolution inputs (112^2) during both training and evaluation, resulting in significantly faster inference. Although mPLUG [21] was fine-tuned to optimize CIDEr-based rewards using the self-critical sequence training (SCST) framework [41], it achieves higher CIDEr scores at the cost of generating less distinctive captions. This tendency to mimic reference captions also leads to lower self-retrieval performance. In this study, we focus on generating distinctive captions to improve self-retrieval; thus, we do not implement any reference-based optimization to boost metrics like CIDEr. Table 2 compares

the performance of FaceCPT with other methods, showing improvements over baselines. Specifically, FaceCPT exceeds FLIP [28] by 1.50%, 2.11%, and 0.40% in R@5 on the MM-CelebA, Face2Text, and CelebA-Dialog datasets, respectively. In addition, Table 3 compares zero-shot retrieval performance of FaceCPT with SOTA baselines fine-tuned on MM-CelebA [55]. FaceCPT significantly outperforms them on both LFW [16] and AgeDB [34], demonstrating strong robustness to missing modalities and incomplete textual queries.

Evaluation on Facial Analysis: Following the standard protocol [57], we evaluate FaceCPT for attribute recognition, as shown in Table 4, where most SOTA methods use input resolutions of 224×224. Although FaceCPT is not specifically pre-trained for discriminative tasks, it achieves state-of-the-art performance in FAR with only moderate fine-tuning, outperforming both FaRL [57] and FaceXFormer [35]. Furthermore, Table 6 presents FER results on RAF-DB [26] and AE results on AgeDB [34], where FaceCPT surpasses existing FRL methods and reach the performance of task-specific models. Similarly, Table 5 reports the 1:1 verification accuracy of various methods on benchmark datasets. Our method once again outperforms FRL baselines, and achieves performance comparable to off-the-shelf FR methods [11, 52], demonstrating its effectiveness in learning general facial representation.

4.4 Ablation Study

We analyze FaceCPT through ablations on the MMCelebA [55] dataset using a ResNet-50 image encoder.

Ablating Pre-training Components: We examine the impact of different components of FaceCPT in Table 7. For the zero-shot retrieval task, we start with the ITM+ITC components, which form the core of FaceCPT's retrieval architecture. Adding the LM components slightly improves performance by enabling the model to capture fine-grained details. Incorporating the visual adapter (VA) into the image encoder enhances performance by combining global and local image features into a more comprehensive representation. For face captioning, we start with LM loss and incrementally add components, with ITM+ITC losses further improving captioning. Moreover, the application of SAAL loss has a significant impact.

Ablating Image Encoders: We further evaluate the impact of different image encoders on FaceCPT, as demonstrated in Table 8. ResNets-based encoder performs comparably to ViT-Base [11] with patch sizes of 16×16 and 8×8. Although ViT-B/8 and ViT-L may provide better performance with larger pre-training datasets, their computational cost is higher compared to the ResNet encoders.

Ablating Image Quality: We analyze the impact of image quality on FaceCPT's performance in cross-modal tasks by simulating various levels of image degradation using atmospheric turbulence [33]. As reported in Table 9, our model maintains strong performance in face captioning but experiences a decline in text-based face retrieval. The distortions in low-quality images obscure key visual cues, leading to poor facial representation.

Visualization: In Figure 3, we present the Grad-CAM-generated class activation maps [43] to illustrate how FaceCPT attends to key facial regions based on the semantic attributes of the

Sample Distorted Faces	Distortion Levels(↓)	Turbulence Strengths	Retrieval-FT R@5	Retrieval-FT R@10	Caption-FT B@4	Caption-FT M	Caption-FT RL	Caption-FT C
0 1 2 3 	0	0	49.50	62.02	40.54	61.60	57.50	30.05
	1	0.5	38.49	54.08	40.32	60.44	56.48	25.77
	2	1	33.26	48.97	39.25	59.97	55.25	23.61
	3	1.5	27.85	42.65	37.45	59.35	54.68	22.83

Table 9: Ablation study on the impact of face quality on cross-modal performance.

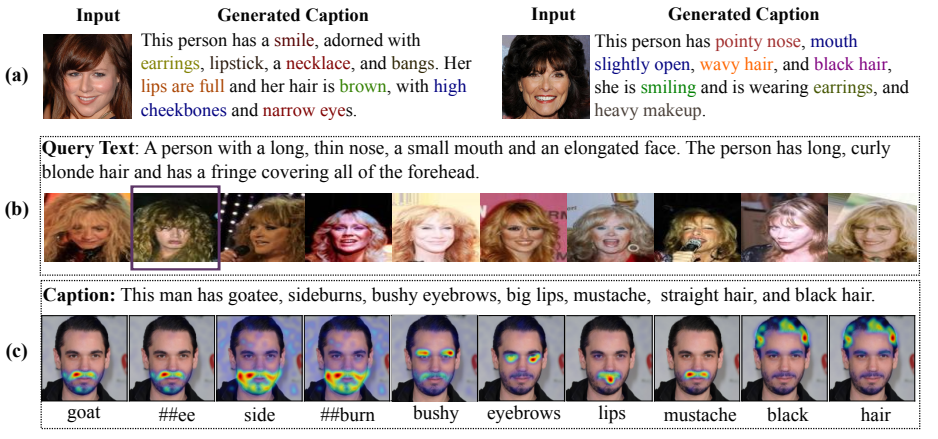

Figure 3: (a) Visualization of sample captions generated by FaceCPT. (b) Retrieved images from text-based face retrieval task. (c) Class activation maps generated using Grad-CAM.

query caption. The Figure also includes sample captions and retrieved images for text-based face retrieval. (See supplementary materials for additional visualizations.).

5 Conclusion

We present FaceCPT, a novel weakly supervised framework for learning general facial representation. While VLP models are typically optimized to align with references, they often generate less distinctive outputs. In contrast, FaceCPT incorporates the SAAL during fine-tuning, which encourages the model to capture fine-grained attributes, leading to more distinctive captions. As a result, FaceCPT outperforms VLP methods and shows strong zero-shot transferability. However, it still requires fine-tuning with labeled data for facial analysis. In future, we plan to pre-train on a larger dataset to further improve its task adaptability.

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *Int. Conf. Learn. Represent.*, 2022.

[2] Adrian Bulat, Shiyang Cheng, Jing Yang, Andrew Garbett, Enrique Sánchez-Lozano, and Georgios Tzimiropoulos. Pre-training strategies and datasets for facial representation learning. In *Eur. Conf. Comput. Vis.*, volume 13673, pages 107–125, 2022.

[3] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. Low-resolution face recognition. In *Asi. Conf. Comput. Vis.*, pages 605–621, 2019.

[4] Dawei Dai, Shiyu Fu, Yingge Liu, and Guoyin Wang. Vision-language joint representation learning for sketch less facial image retrieval. *Information Fusion*, 112:102535, 2024.

[5] Dawei Dai, YuTang Li, YingGe Liu, Mingming Jia, Zhang YuanHui, and Guoyin Wang. 15M multimodal facial image-text dataset. *arXiv preprint arXiv:2407.08515*, 2024.

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: additive angular margin loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4685–4694, 2019.

[7] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-shot multi-level face localisation in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5202–5211, 2020.

[8] Yuchuan Deng, Qijun Zhao, Zhanpeng Hu, and Zixiang Xu. Text-based face retrieval: Methods and challenges. In *Chinese Conf. on Biometric Recog.*, pages 150–159, 2023.

[9] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Ninth workshop on statistical machine translation*, pages 376–380, 2014.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, 2019.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Int. Conf. Learn. Represent.*, 2021.

[12] Zheng Gao and Ioannis Patras. Self-supervised facial representation learning with facial region awareness. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2081–2092, 2024.

[13] Md Mahedi Hasan, Shoaib Meraj Sami, Nasser Nasrabadi, and Jeremy Dawson. Learning multi-scale knowledge-guided features for text-guided face recognition. *IEEE Trans. on Biom., Behav., and Ident. Sc.*, 2024.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016.

[15] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *Int. Conf. Mach. Learn.*, pages 2712–2721, 2019.

[16] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in Real-Life Images: detection, alignment, and recognition*, 2008.

[17] Anil K Jain, Brendan Klare, and Unsang Park. Face matching and retrieval in forensics applications. *IEEE multimedia*, 19(1):20, 2012.

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, and et al. Scaling up visual and vision-language representation learning with noisy text supervision. In *Int. Conf. Mach. Learn.*, pages 4904–4916, 2021.

[19] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-To-Edit: Fine-grained facial editing via dialog. In *Int. Conf. Comput. Vis.*, pages 13799–13808, 2021.

[20] Minchul Kim, Anil K Jain, and Xiaoming Liu. AdaFace: quality adaptive margin for face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18750–18759, 2022.

[21] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022.

[22] Jianshu Li, Fang Zhao, Jiashi Feng, Sujoy Roy, Shuicheng Yan, and Terence Sim. Landmark free face attribute prediction. *IEEE Trans. Img. Proc.*, 27(9):4651–4662, 2018.

[23] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Adv. Neural Inform. Process. Syst.*, 34:9694–9705, 2021.

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Int. Conf. Mach. Learn.*, pages 12888–12900, 2022.

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Int. Conf. Mach. Learn.*, pages 19730–19742, 2023.

[26] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2852–2861, 2017.

[27] Yudong Li, Xianxu Hou, Zhe Zhao, Linlin Shen, Xuefeng Yang, and Kimmo Yan. Talk2Face: A unified sequence-based framework for diverse face generation and analysis tasks. In *ACM Int. Conf. Multimedia*, pages 4594–4604, 2022.

[28] Yudong Li, Xianxu Hou, Zheng Dezhi, Linlin Shen, and Zhe Zhao. Flip-80m: 80 million visual-linguistic pairs for facial language-image pre-training. In *32nd ACM International Conference on Multimedia*, pages 58–67, 2024.

[29] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Ass. for comput. ling.*, pages 605–612, 2004.

[30] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3730–3738, 2015.

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Int. Conf. Learn. Represent.*, 2019.

[32] Longbiao Mao, Yan Yan, Jing-Hao Xue, and Hanzi Wang. Deep multi-task multi-label CNN for effective facial attribute classification. *IEEE Transactions on Affective Computing*, 2020.

[33] Zhiyuan Mao, Nicholas Chimitt, and Stanley H Chan. Accelerating atmospheric turbulence simulation via learned phase-to-space transform. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14759–14768, 2021.

[34] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. AgeDB: the first manually collected, in-the-wild age database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 51–59, 2017.

[35] Kartik Narayan, Vibashan VS, Rama Chellappa, and Vishal M Patel. FaceXFormer: A unified transformer for facial analysis. *arXiv preprint arXiv:2403.12960*, 2024.

[36] Lu Pang, Yaowei Wang, Yi-Zhe Song, Tiejun Huang, and Yonghong Tian. Cross-domain adversarial feature learning for sketch re-identification. In *ACM Int. Conf. Multimedia*, pages 609–617, 2018.

[37] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Association for Computational Linguistics*, pages 311–318, 2002.

[38] Lixiong Qin, Mei Wang, Chao Deng, Ke Wang, Xi Chen, Jiani Hu, and Weihong Deng. SwinFace: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[39] Lixiong Qin, Mei Wang, Xuannan Liu, Yuhang Zhang, Wei Deng, Xiaoshuai Song, Weiran Xu, and Weihong Deng. Faceptor: A generalist model for face perception. In *Eur. Conf. Comput. Vis.*, 2024.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, and et al. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, pages 8748–8763, 2021.

[41] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[42] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *Adv. Neural Inform. Process. Syst.*, 35:25278–25294, 2022.

[43] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 618–626, 2017.

[44] Ankit Kumar Sharma and Hassan Foroosh. Slim-CNN: A light-weight cnn for face attribute prediction. In *International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 329–335, 2020.

[45] Ying Shu, Yan Yan, Si Chen, Jing-Hao Xue, Chunhua Shen, and Hanzi Wang. Learning spatial-semantic relationship for facial attribute recognition with limited labeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11916–11925, 2021.

[46] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: Simplifying semi-supervised learning with consistency and confidence. *Adv. Neural Inform. Process. Syst.*, 33:596–608, 2020.

[47] Haomiao Sun, Mingjie He, Shiguang Shan, Hu Han, and Xilin Chen. Task-adaptive Q-Face. *arXiv preprint arXiv:2405.09059*, 2024.

[48] Jianxin Sun, Qi Li, Weining Wang, Jian Zhao, and Zhenan Sun. Multi-caption text-to-face synthesis: Dataset and algorithm. In *ACM Int. Conf. Multimedia*, pages 2290–2298, 2021.

[49] Marc Tanti, Shaun Abdilla, Adrian Muscat, Claudia Borg, Reuben A Farrugia, and Albert Gatt. Face2Text revisited: Improved data set and baseline results. *Proceedings of the workshop on People in Vision, Language, and the Mind*, pages 41–47, 2022.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, and et al. Attention is all you need. *Adv. Neural Inform. Process. Syst.*, 30, 2017.

[51] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4566–4575, 2015.

[52] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. CosFace: large margin cosine loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5265–5274, 2018.

[53] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *Int. Conf. Mach. Learn.*, pages 23318–23340, 2022.

[54] Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Liang Liu, Yabiao Wang, and Chengjie Wang. Toward high quality facial representation learning. In *ACM Int. Conf. Multimedia*, pages 5048–5058, 2023.

[55] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. TediGAN: Text-guided diverse face image generation and manipulation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2256–2265, 2021.

[56] Alireza Zaeemzadeh, Shabnam Ghadar, Baldo Faieta, Zhe Lin, Nazanin Rahnavard, Mubarak Shah, and et al. Face image retrieval with attribute manipulation. In *Int. Conf. Comput. Vis.*, pages 12116–12125, 2021.

[57] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18676–18688, 2022.

[58] Yutong Zhou and Nobutaka Shimada. Generative adversarial network for text-to-face synthesis and manipulation with pretrained BERT model. In *International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08, 2021.

[59] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, et al. WebFace260M: A benchmark unveiling the power of million-scale deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10492–10502, 2021.