2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3315-4364-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/CVPR52734.2025.00373

# GroundingFace: Fine-grained Face Understanding via Pixel Grounding Multimodal Large Language Model

Yue Han[1], Jiangning Zhang[*,1,2], Junwei Zhu[2], Runze Hou[3],
Xiaozhong Ji[2], Chuming Lin[2], Xiaobin Hu[2], Zhucun Xue[1], Yong Liu[†,1,4]
[1]Zhejiang University [2]Youtu Lab, Tencent [3]Tsinghua University
[4]State Key Laboratory of Industrial Control Technology
22132041@zju.edu.cn, yongliu@iipc.zju.edu.cn

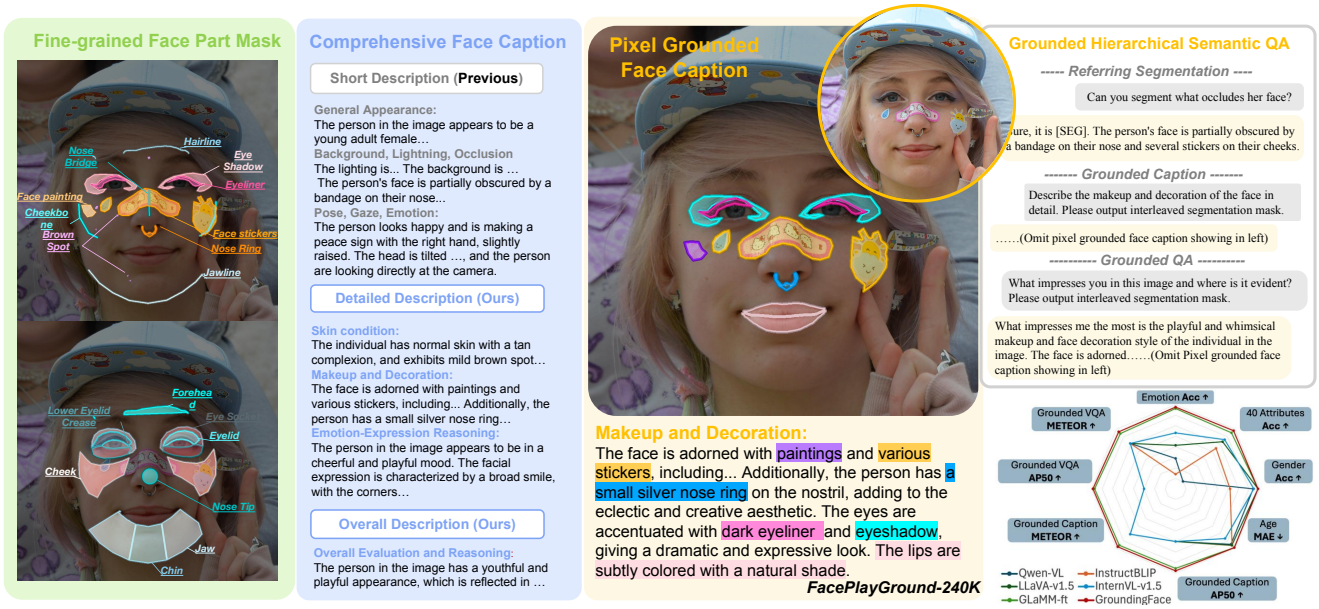


Figure 1. We propose ***FacePlayGround-240K*** for fine-grained face understanding, which includes meticulously annotated fine-grained face part masks and comprehensive face captions. This dataset supports various downstream face-related tasks, and our proposed facial pixel grounding GroundingFace achieves impressive results across multiple tasks, while current open-source models fail.

## Abstract

*Multimodal Language Learning Models (MLLMs) have shown remarkable performance in image understanding, generation, and editing, with recent advancements achieving pixel-level grounding with reasoning. However, these models for common objects struggle with fine-grained face understanding. In this work, we introduce the* ***FacePlayGround-240K*** *dataset, the first pioneering large-scale, pixel-grounded face caption and question-answer (QA) dataset that includes 240K images, 47 mask categories, 5.4M mask annotations, and 7.3M grounded regions, meticulously curated for alignment pretraining and instruction-tuning. We present the* ***GroundingFace*** *framework, specifically designed to enhance fine-grained face understanding. This framework significantly augments the capabilities of existing grounding models in face part segmentation, face attribute comprehension, while preserving general scene understanding. Comprehensive experiments validate that our approach surpasses current state-of-the-art models in pixel-grounded face captioning/QA and various downstream tasks, including face captioning, referring segmentation, and zero-shot face attribute recognition.*

## 1. Introduction

Multimodal Language Learning Models (MLLMs) have demonstrated impressive performance in image understanding, generation, and editing. Recently, grounding

[*] Project lead.
[†] Corresponding author.

2575-7075/25/$31.00 ©2025 IEEE
DOI 10.1109/CVPR52734.2025.00373

MLLMs [5, 13, 36, 51] have extended the understanding capabilities from the image level to the region level, even achieving pixel-level grounding with reasoning capabilities [24]. However, these methods [24, 36, 44] are designed for common scenarios, and their performance on faces is less than satisfactory. Current MLLMs often 'cannot determine' attributes in response to face-related QA problems. This deficiency in fine-grained face understanding impedes the granularity and generalization of models in subsequent generation and editing tasks.

Challenges for detailed face understanding lie in both the dataset and the model. We identify several issues in current datasets for image-level captioning and visual question answering (VQA): ***1) Insufficient attribute:*** Current face caption dataset typically consists of short descriptions with a limited set of facial attributes, failing to cover enough aspects. Additionally, there is a lack of semantic hierarchy, resulting in both finer and broader descriptions being overlooked. ***2) Absence of inter-attribute relationships:*** Template-generated captions lead to disconnected attributes, hindering relational reasoning. ***3) Lack of face QA data:*** The community suffers from a deficiency in face QA data, while the scarcity of relational captions further hampers the generation of QA pairs that possess reasoning capabilities and semantic richness. ***4) Deficiency in region mask:*** There is a significant gap between the vocabulary of facial parts contained in captions and mask categories. The current 19 mask categories in face parsing do not adequately support region-text alignment for pixel grounding.

To address the above deficiencies, we propose a new dataset, termed FacePlayGround-240K, for fine-grained face understanding. The construction pipeline ( 3) involves four stages: ***1) Comprehensive caption generation:*** Discrete attributes are gathered and conditioned on the large VLM [7] to yield captions balancing both attribute reliability and diversity. Captions includes 3 semantic levels, *i.e.*, short, detailed and overall, and 12 subdivided aspects (Sec. 3.2). ***2) Fine-grained part mask annotation*** involves a combination of automatic and manual processes. We automatically generate masks for structural areas and lines, acquire masks for skin details through commercial APIs [11], and manually annotate masks for facial makeup and decorations (Sec. 3.3). ***3) Text-mask alignment for grounded caption:*** The captions and masks obtained from the previous two step are aligned to obtain the grounded caption (Sec. 3.4). ***4) Grounded hierarchical semantic QA for instruction following:*** We use LLM model [1] to generate QA pairs from grounded captions across three hierarchical levels, *i.e.*, concrete attribute-level, abstract trait-level, and overall impression-level. This procedure yields data for referring segmentation (with single object mask) and grounded QA (with multiple object masks) (Sec. 3.5). Fig. 1 showcases one example.

With the constructed dataset, several key questions must be thoroughly examined for the model design: ***1)*** Current models struggle with fine-grained attribute understanding and small part segmentation due to the low-resolution global feature and deep grounding feature that loses low-level information. ***2)*** There is a disparity in the quality of the annotation masks and the precision required for segmenting larger facial structures (e.g., forehead) versus smaller facial parts (e.g., wrinkles). How can we design training strategies to fully utilize high-quality manually annotated data and improve the segmentation accuracy of small parts? ***3)*** With single face closed-up images suitable for detailed attribute grounding, can we leverage the scene understanding capabilities of general grounding models in segmenting common objects and distinguishing multiple instances?

To address these challenges, we propose the GroundingFace framework. For high-resolution face understanding, we reuse the SAM shallow feature for segmentation and introduce a face-prior sampler to effectively extract and compress high-quality face tokens (Sec. 4.3). For high-resolution face part segmentation, we integrate a high-quality adapter into the baseline model for fine-tuning, coupled with a two-stage training strategy. This approach fully utilizes both automatically and manually annotated data to improve fine-grained segmentation accuracy (Sec. 4.4). Additionally, we introduce a LoRA MoE to route high-quality and low-quality tokens to their corresponding two-stage LoRA, preserving the ability to segment common objects and distinguish multiple objects while enhancing the model's fine-grained capabilities (Sec. 4.5).

In summary, our contributions are threefold:

- We propose a novel fine-grained face understanding FacePlayGround-240K dataset, the first large-scale, pixel-grounded face caption and question-answer (QA) dataset, specifically designed for alignment pretraining and instruction-tuning.
- We introduce the GroundingFace framework for fine-grained face understanding, enhancing GLaMM's capabilities in face part segmentation, face attribute understanding, and maintaining general scene capabilities.
- Extensive experiments show that our method surpasses current state-of-the-art models in pixel-grounded face caption/QA and downstream tasks like face captioning, segmentation, and zero-shot face attribute recognition.

## 2. Related Work

**Grounding Multimodal Large Language Models.** Grounding MLLMs extends fine-grained image understanding from the image level to the region level. Kosmos-2 [40] and Shikra [5] use coordinates for box representation, while GPT4RoI [51] and RegionGPT [13] utilize pooled region embeddings. Leveraging SAM [22], LISA [24] achieves pixel-level grounding and enhances reasoning capabilities, enabling segmentation based on complex and implicit queries. GLaMM [44] surpasses single-object grounding by generating natural language responses integrated with multiple object segmentation masks. However, these methods, designed for general

scenarios, perform poorly in facial recognition, which demands extremely fine-grained perception. Therefore, we propose incorporating high-resolution face prior information and a high-precision mask decoding architecture for fine-grained face grounding.

**MLLM datasets.** MLLM datasets for common scenarios include visual-text alignment and instruction-following datasets [6, 30], featuring subtasks like captioning [6], referring segmentation [37, 41, 50], and grounded captioning [42, 44]. However, these datasets lack fine-grained facial concepts due to the rarity of close-up faces in general scenario images. And their region annotations typically come from detection datasets like COCO [29], Objects365 [45], and SAM [22], which often treat the face as a whole or include only a few categories such as eyes, hair, and ears. Currently, MLLM datasets for faces are limited to caption datasets [22] and lack fine-grained semantic descriptions and pixel-level mask annotations, hindering the development of high-precision face understanding models. To address this limitation, this paper introduces the FacePlayGround-240K dataset to facilitate the training of fine-grained face understanding models.

**Fine-grained face understanding.** Facial understanding encompasses a range of fine-grained linguistic perceptions: 1) Static attribute understanding, which includes fine-grained facial attribute recognition [3] like age and gender; 2) Action understanding, which involves expression [27], pose [48], and gaze estimation [52]; 3) Region understanding, which includes landmark detection [23] and face parsing [21]. However, current facial understanding datasets [18, 25, 34] contain a limited number of attribute categories, such as the 40 attributes in the CelebA [34] and 19 semantic areas for CelebAMask-HQ [25]. Consequently, models trained on these datasets are restricted to closed-set categories and cannot generalize to more fine-grained facial understanding. Our FacePlayGround-240K, a large-scale face dataset with fine-grained facial masks and descriptions, which supplies the generalization and fine-grained understanding capabilities of the models trained on it.

# 3. Pipeline for FacePlayGround-240K

## 3.1. Acquisition Sources.

To meet the requirements for fine-grained facial attribute understanding and grounding, the data selection criteria are based on high-resolution, close-up images of individual faces. This includes datasets such as CelebAMask-HQ [25] (30K), FFHQ [19] (70K), and EasyPortrait [18] (40K). To enhance the generalizability of the data, we also supplement with high-resolution individual face images from LAION-Face [53] (98K), which covers a broader range of everyday scenarios. We use high-precision face detectors and aesthetic scoring strategies to exclude low-quality facial data.

## 3.2. Comprehensive Face Caption Generation.

In the current dataset, facial captions are primarily generated using grammar templates to organize discrete attributes [34]. This method not only has limitations in the variety of attributes but also, even when rewritten using large language models (LLMs) to enrich expression, the generated descriptions still lack attribute relationships and semantic hierarchy. Additionally, relying solely on large visual language models (VLMs) for annotation faces issues of unreliable attributes due to hallucination effects. To address these problems, we employ a method that combines facial attribute recognition with large visual language models to generate fine-grained descriptions as follows.

**LLM-assisted caption generation.** We utilize a commercial high-precision face recognition API by Face++ to obtain quantitative data, including gender, age, emotion intensity, occlusion area, pose/gaze angle, and skin analysis, which supports the generation of detailed fine-grained facial captions. Using a list of discrete attribute values as premises, we prompt InternVL [7] from seven predefined aspects (top part of Fig. 2) to generate initial text responses. While this approach yields detailed and comprehensive answers, it may result in semantic repetition and redundancy. Additionally, using discrete attributes as strict conditions can lead to responses that directly replicate attribute values, causing incoherent expressions. Therefore, we integrate the original texts from all aspects and employ a large language model (LLM) [1] for deduplication and rewriting.

## 3.3. Fine-grained Part Mask Annotation

To achieve face grounding, it is essential to address the discrepancy in the number of face part vocabularies between captions and masks. Therefore, we utilize SpaCy to perform phrase segmentation, noun phrase extraction, and word frequency statistics on the descriptions obtained in Sec. 3.2. From this, we select high-frequency terms related to facial features, merging similar categories to derive 47 new concepts. These concepts are introduced as mask categories and are grouped into six supercategories in Fig. 4.

**Structural area/line.** We categorize facial regions from facial structure priors into planar regions (e.g., forehead) and linear regions (e.g., cheekbone). These regions are decomposed using MediaPipe's 478 dense facial landmarks [35], and edge optimization is performed on the parsing masks according to the design logic.

**Skin analysis.** Skin details, representing local texture features of the face, are obtained using the commercial skin analysis API [11] for masks of skin issues like wrinkles and spots. We also include one manually annotated class from FFHQ-Wrinkle [38] (1K) and 17 classes from an open-source project [15] (1.45K).

**Hair analysis.** Facial hair includes bangs, moustache, and beard. We determine these categories in images based on descriptions and then use GroundingDINO [32] to locate them. To prevent errors, we constrain the locations using

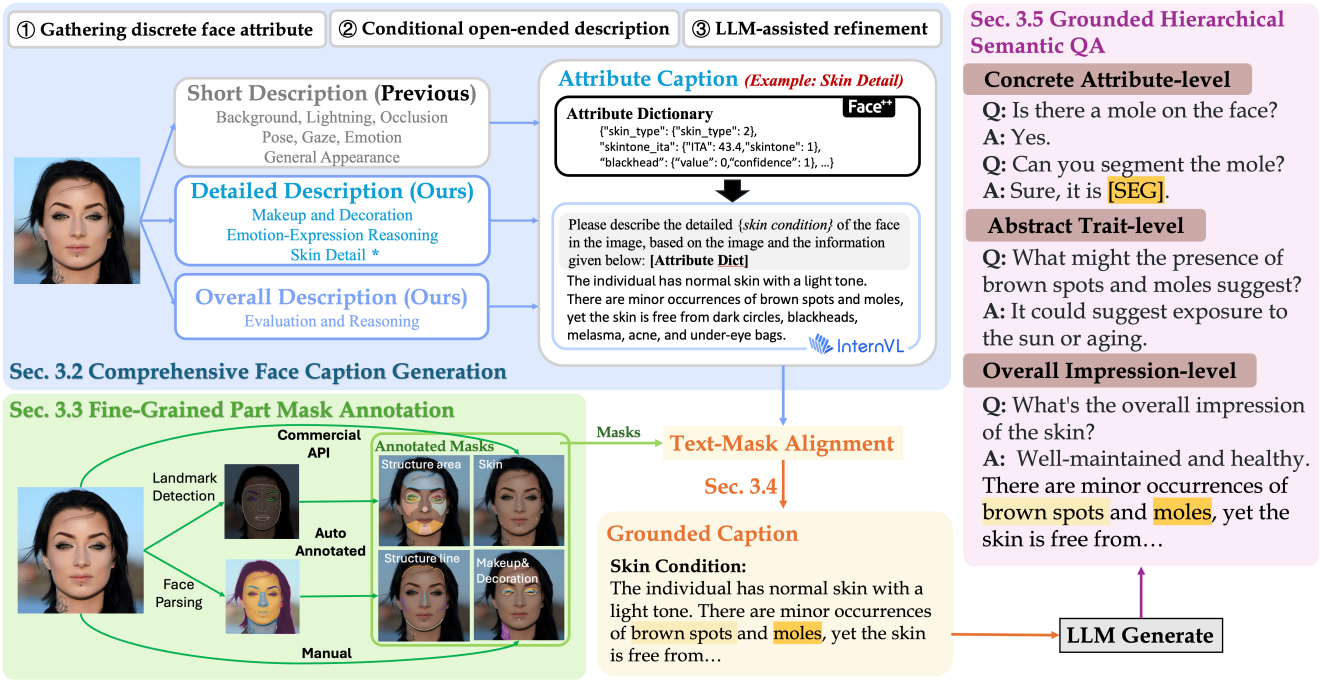


Figure 2. **Construction pipeline of FacePlayGround-240K.** The process involves 4 stages: ***1)*** Comprehensive Face Caption Generation (Sec. 3.2); ***2)*** Fine-grained Part Mask Annotation (Sec. 3.3); ***3)*** Text-Mask Alignment for Grounded Caption (Sec. 3.4); and ***4)*** Grounded Hierarchical Semantic QA for Instruction Following (Sec. 3.5);.

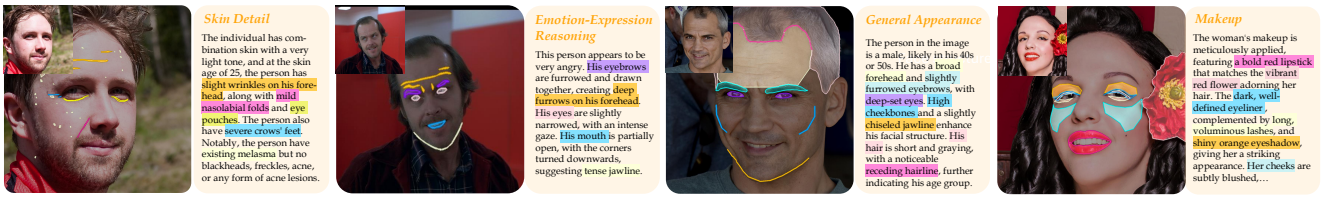


Figure 3. Examples of grounded caption in FacePlayGround-240K.

landmark priors and then provide the bounding boxes to SAM [22] to obtain the corresponding masks.

**Manually-labeled makeup and face decoration.** The fuzzy boundaries of makeup make annotation challenging. Therefore, we filter images with heavy makeup through face captions and manually annotate eyeliner, eyeshadow, and eyelashes with clear boundaries, while omitting annotations for makeup areas with fuzzy boundaries such as highlighter and blush. Lipstick and eyebrow shadow are directly annotated using parsing masks. We employed 10 individuals, resulting in 3,745 annotated images.

### 3.4. Text-Mask Alignment for Grounded Caption

**Phrase decomposition.** To align fine-grained masks with local phrases in long texts, we first need to obtain semantically clear phrases. We use SpaCy to decompose the original text and extract phrases and corresponding noun chunks.

**Text-mask Alignment.** The masks for hair, face makeup, and decoration are annotated based on the descriptions in the captions, thus they have a one-to-one correspondence and can be directly matched. We only need to match the face structure area, skin mask, and captions. By merging the high-frequency word list, we obtained a synonym list corresponding to the class names. We traverse all noun phrases to determine if they are in the synonym list, allowing for direct matching of noun vocabulary. However, this approach can lead to issues. For example, "lines on the forehead" (corresponding to the forehead wrinkle mask) contains two noun phrases: "line" and "forehead," which may result in one phrase matching multiple class masks. To address this problem, we further use SpaCy to calculate the similarity between class names and phrases, matching the mask class name with the phrase that has a higher similarity above a given threshold. Results are shown in Fig. 3.

### 3.5. Grounded Hierarchical Semantic QA

We observe that general-purpose MLLMs tend to generate face captions when answering questions about facial attributes. This tendency is due to the imbalance between the amount of face caption data and detailed attribute question-

Table 1. **Comparison for popular facial datasets from multiple dimensions.** FacePlayGround-240K contains a larger volume of data along with detailed pixel-level and textual annotations in multiple granularity. ✔: Satisfied; ✘: Unsatisfied. -: Inapplicable.

| Dataset | Data Source | Samples | Mask | | Text | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Part | Labeling Manner | Attibute | Short Description | | | | | Detailed Description | | | Overall Evaluation | Labeling Manner |
| | | | | | | General appearance | Pose Gaze | Emotion | Occlusion | Lighting Background | Skin detail | Makeup Face decoration | Expression | | |
| FFHQ-Text [19] | FFHQ | 760 | - | - | 162 | ✔ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | manual |
| MM-CelebA [47] | CelebA-HQ | 30K | 19 | manual | 40 | ✔ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | - |
| CelebA-Dialog [17] | CelebA-HQ | 30K | - | - | 40 | ✔ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | - |
| CelebV-Text [49] | self-collected | 70K | - | - | 77 | ✔ | ✔ | ✔ | ✘ | ✔ | ✔ | ✘ | ✔ | ✘ | automatic |
| LAION-Face [53] | self-collected | 50K | - | - | - | - | - | - | - | - | - | - | - | - | automatic |
| FaceCaption-15M [8] | LAION-Face | 15M | - | - | 40 | ✔ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | automatic |
| EasyPortrait [18] | self-collected | 40k | 9 | manual | - | - | - | - | - | - | - | - | - | - | - |
| **FacePlayGround -240K (Ours)** | CelebAMask-HQ EasyPortrait, FFHQ LAION-Face, SEWA | **240K** | **66** | **automatic manual** | **open-ended** | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | **automatic** |

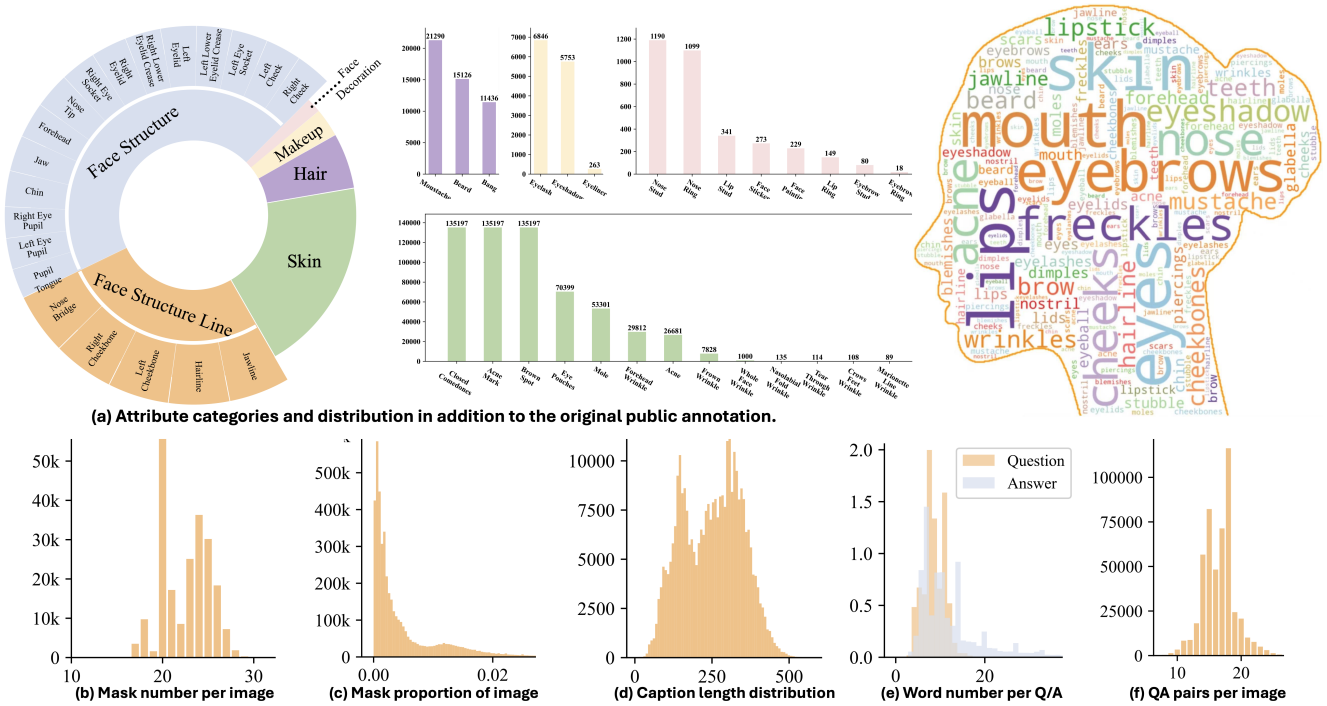


Figure 4. Comprehensive statistics of FacePlayGround-240K in several aspects. Zoom in for better viewing.

answering data available in the community. On the other hand, excessively long captions increase the difficulty of visual-text alignment for the model. Therefore, in pursuit of comprehensiveness, the description of each aspect's details is limited. In contrast, question-answering tasks allow for detailed inquiries into specific aspects, which better facilitates the model's fine-grained understanding of faces.

**Hierarchical semantic QA.** There exist different semantic levels and inferential relationships among facial attributes. For instance, age-related attributes include skin condition and hair color, while emotions are closely related to facial expressions. In the descriptions obtained in Sec. 3.2, these associations are already present. We highlight and emphasize these connections through a question-and-answer format. Specifically, we categorize the semantic levels into three: "*concrete attribute-level*" (*e.g.*, hair color, wrinkles), "*abstract trait-level*" (*e.g.*, age), and "*overall impression-level*" (*e.g.*, elderly person). We use contextual prompts to guide the LLM in generating question-and-answer pairs from grounded caption for these three semantic levels.

**Data for grounded QA and referring segmentation.** The obtained grounded Hierarchical Semantic QA data is used for the grounded QA task. QA pairs that only point to a single object mask are used for the referring segmentation.

### 3.6. Dataset Analysis

**Comparing with counterpart datasets.** The current facial datasets primarily exhibit the following issues: 1) limited data scale; 2) the face part vocabulary is restricted to the 19 categories proposed by CelebAMask-HQ [25]; 3) limited textual attribute annotations; 4) lack of attribute relationships and hierarchical semantic descriptions; 5) insufficient alignment between mask and text annotations; 6) absence of fine-grained understanding and reasoning VQA data. These factors contribute to low accuracy or poor generalization in multimodal fine-grained facial understanding. Our FacePlayGround-240K addresses these issues by expanding the previous parsing dataset [25] with addi-

tional high-resolution FFHQ [19]. Additionally, we have filtered LaionFace [53] based on resolution. Ultimately, through a meticulous multi-level annotation process, the proposed dataset demonstrates significant improvements in data scale, fine-grained and precise attribute diversity, and annotation/alignment, as illustrated in Fig. 2, which highlights the differences compared to related datasets.

**Dataset statistics.** Fig. 4 presents a comprehensive statistical analysis of FacePlayGround-240K. In addition to the originally published attribute annotations, Fig. 4(a) shows the distribution of the fine-grained facial attribute categories we provided, as discussed in Sec. 3.3, along with their sub-attributes in FacePlayGround-240K. Note that each sub-attribute in the Face Structure and Line category is present in every image, hence they are not listed. Word clouds illustrate the diversity and frequency distribution of text annotations. Fig. 4(b) shows the mask number per image. Fig. 4(c) illustrates the distribution of the relative area occupied by mask annotations, which varies widely and provides diverse pixel-level grounding annotations. Fig. 4(d), (e), and (f) present the distribution of the caption length, word number per question / answer, and QA pair per image.

# 4. Methodology: GroundingFace

## 4.1. Research definition.

This work extends face understanding task to a finer granularity of facial pixel-wise grounding in MLLM. Given a prompt $P$ and a facial image $I$, the MLLM generates text responses $T$ interleaved with one or multiple face part segmentation masks $M$. After the training phase, this model can be applied to various downstream tasks, including grounded face captioning, zero-shot face attribute recognition, and face-oriented grounded question answering. We use the general GLaMM framework based on MLLM [44] as the foundation to develop the GroundingFace model for fine-grained face understanding, as illustrated in Fig. 5.

## 4.2. Motivation of Different Components

**Instance-level *vs*. Part-level.** Typically, MLLM for general object pixel grounding tends to understand inter-instance relationships and perform instance-level segmentation [44]. However, the current global visual encoder, which uses low-resolution input, fails to meet the requirements for understanding fine-grained attributes within instances. Face understanding is particularly sensitive to detailed parts. Additionally, the mask decoder based on SAM is inadequate for fine-grained part segmentation due to the coarse-grained nature of the generated masks. To address these issues, we propose Fine-grained Face Part Segmentation (Sec. 4.3) and Fine-grained Face Attribute Understanding (Sec. 4.4).

**Coarse auto-generated masks *vs*. Accurate manual-annotated masks.** Compared to larger face structure areas (*e.g*., forehead), smaller facial parts (*e.g*., wrinkles) demand higher precision in mask edges, requiring different levels of annotation quality. Training all granularity masks together can negatively impact the segmentation accuracy of small parts. Therefore, we introduce an additional Stage-2, where we fine-tune the model using high-quality data specifically for small facial parts and manually annotated data.

**Distinguish multiple instances and prevent common object forgetting.** To achieve fine-grained face grounding, we utilize a close-up single-face image dataset without multiple-person scenes and common object labeling. However, the general scene base model inherently excels in common object grounding and distinguishing multiple instances. Thus, we incorporate stage-aware LoRA [14] and MoE [16] for efficient fine-tuning to retain the model's general scene understanding capabilities Sec. 4.5.

## 4.3. Fine-grained Face Part Segmentation

The vanilla SAM decoder is trained for general instance segmentation without considering the need for high-precision segmentation. Therefore, we introduce HQ-SAM [20] to improve the face-oriented mask decoder in MLLM, referred to as the HQ-SAM Adapter, to achieve fine-grained mask generation.

**High-quality mask decoder.** We visualize the features of different ViT layers in Fig. 6 and observe that shallow layer features exhibit finer pixel details, while deep layer features are more oriented towards high-level semantics. Considering that facial understanding requires high detail but small parts are almost invisible in deep-layer features, we introduce shallow-layer features to enhance fine-grained representation. To better utilize fine-grained features, we employ a Shallow-Deep Fusion method to integrate ViT features of different depths. This module is implemented using a combination of naive deconvolution and linear layers.

## 4.4. Fine-grained Face Attribute Understanding


Figure 6. Visualization of different scale (by rows) / depth (by columns) features of CLIP ViT [10, 43].

To achieve high-resolution image understanding, the current mainstream approach in MLLM methods is to adopt a dynamic resolution crop strategy. This method enlarges local regions to capture more image details while keeping the visual encoder unchanged, but it results in a significant increase in the number of tokens, thereby increasing the computational cost of the LLM component [13, 51]. Some works focus on adaptive token selection and sampling, but this is unnecessary for faces since their locations can be easily obtained through priors, such as a pretrained face detector. Another approach involves introducing higher resolution visual encoders and multi-stage features, which also significantly increase computational costs. Unlike these methods, we propose reusing the shallow and deep features extracted by SAM, as described in Sec. 4.3, which

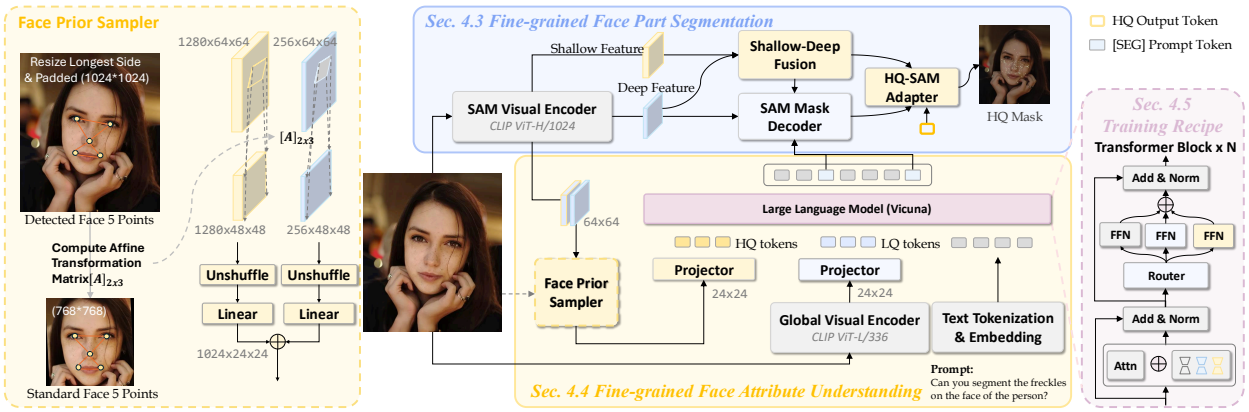


Figure 5. **Overview: The GroundingFace framework is specifically designed for *fine-grained face understanding*,** which enhances GLaMM's fine-grained facial understanding capabilities on three aspects: *1)* Fine-grained Face Part Segmentation (Sec. 4.3) that integrates shallow and deep features for small part mask generation. *2)* Fine-grained Face Attribute Understanding (Sec. 4.4) that reuses SAM features and employs a face prior sampler to efficiently inject facial information. *3)* Training Recipe (Sec. 4.5) introduces MoE and LoRA to maintain the model's general scene capabilities while incorporating fine-grained facial concept understanding.

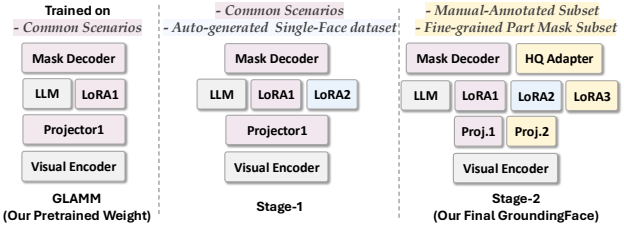


Figure 7. Two-stage training recipe for GroundingFace.

contain rich detailed features.

**High resolution facial tokens sampler.** Directly using SAM features reduces the computational cost of feature extraction, but it still retains a large number of tokens to maintain detailed features. We propose to sample the high-resolution feature tokens corresponding to the facial region in the first stage feature using a face prior sampler: 1) First, determine that the target size of the 64×64 SAM features after sampling is 48×48. Compute the affine transformation matrix using a 1024-resolution SAM image input and a 768-resolution standard five-point facial landmark. By applying the matrix and transforming the features, we obtain the cropped and aligned facial region tokens. This approach increases the proportion of the facial region while reducing the feature space resolution, effectively compressing the number of tokens with minimal loss. Subsequently, we further downsample and compress the tokens using pixel unshuffle [46] and learnable linear layers. Since these high-resolution facial tokens are not spatially aligned with the original image, and the SAM features and Global Visual Encoder features are not in the same feature space, we introduce an additional projector for feature alignment.

## 4.5. Training Recipe

**Efficient Adaption with LoRA.** We introduce a two-stage training strategy, as shown in Fig. 7, to preserve the pretrained knowledge of the baseline GLaMM and enhance the understanding and segmentation of small parts. In the first stage, we fine-tune using both general data and facial data: we freeze the base model's LoRA1 and fine-tune the trainable LoRA2, Projector1, and mask decoder. This step primarily aims to integrate the newly introduced facial concepts with the general scene knowledge. In the second stage, we freeze Projector1, LoRA1/2, and the Mask Decoder, and fine-tune LoRA3, Projector2, and the HQ Adapter on a high-quality hand-labeled dataset and a fine-grained part mask subset. This approach allows the model to learn fine-grained understanding and segmentation while keeping the previous weights unchanged to retain the general knowledge as much as possible.

**Boosting Fine-grained Understanding with MoE.** Although the high-quality data subset has been filtered, it only increases the proportion of high-quality hand-labeled masks in the data. There are still coarse masks present in the grounded caption and VQA data. The model needs to decide whether to use coarse LoRA2 or fine-grained LoRA3 based on the input. Therefore, we introduce a MoE [16] to address this issue. Inspired by MOELoRA [12, 26, 31], we employ a Top-K (K=1) router to assign HQ and LQ token to the corresponding LoRA expert of two training stages. Specifically, it computes the probability $S^l$ of the input token $\mathbf{x}$ being routed to an expert with trainable parameters $\boldsymbol{W_r}$ in layer $l$. The whole process is denoted as follows:

$$S^l(x) = \text{Top-1}\left(\text{Softmax}\left(W_r^l x\right)\right),$$

$$h^l = W_0^l x + S^l(x) A^l B^l x,$$

$$A^l B^l = \begin{cases} A_{HQ}^l B_{HQ}^l & \text{if } S^l(x) = S_{HQ}^l(x) \\ A_{LQ}^l B_{LQ}^l & \text{if } S^l(x) = S_{LQ}^l(x) \end{cases},$$

$h^l$ is the output embedding, $W_0^l \in \mathbb{R}^{d_q \times d_p}$ is pretrained weight matrix, and $A^l, B^l$ are the low-rank matrices, where $A^l \in \mathbb{R}^{d_q \times r}, B^l \in \mathbb{R}^{r \times d_p}$, and $r \ll \min(d_q, d_p)$. Following [16], we apply load balancing loss to mitigate the

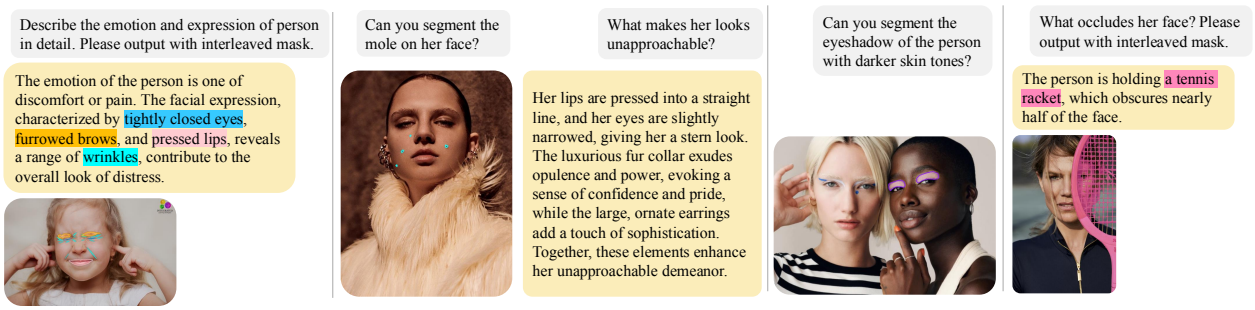


Figure 8. Quantitative capability demonstration across different downstream grounded caption / VQA and referring segmentation tasks.

unbalanced load for experts when training.

## 5. Experiments

**Benchmarking setting.** We conduct a quantitative evaluation of GroundingFace on four benchmarks: 1) Pixel Grounded Face Caption, 2) Face Referring Segmentation, 3) Face-oriented Grounded Caption and Question Answering, and 4) Zero-shot Face Attribute Recognition.

**Dataset and evaluation criteria.** We split the FacePlayGround-240K into training, validation, and test sets comprising 232.5K, 2.5K, and 5K images, respectively. For each model, we evaluate three key aspects: 1) dense caption quality, 2) mask quality, and 3) phrase-mask correspondence accuracy. We include METEOR [4] for captions, class-agnostic mask AP for grounding, and mask IoU for segmentation. We select the currently most powerful and related GLaMM [44] as our baseline since there are no prior pixel grounding works for the face field.

### 5.1. Experimental Results

**Pixel Grounded Face Caption** Tab. 2 shows that GLaMM, trained on general scenario data, performs poorly on the fine-grained face grounding task (Idx.1). However, fine-tuning GLaMM on FacePlayGround-240K significantly improves its performance (Idx.2). Additionally, GroundingFace demonstrates enhanced performance compared to the baseline, as well as obtaining consist conclusion on pixel grounded face caption and VQA Tab. 3. Qualitative results are presented in Fig. 8.

**Face Referring Segmentation.** Given explicit and implicit user instructions, the model outputs a segmentation mask for the specified single target. Qualitative results in Tab. 2 and quantitative cases in Fig. 8 demonstrate our superiority.

**Zero-shot Face Attribute Recognition** We compare our method with general state-of-the-art MLLMs by reorganizing widely-used face attribute recognition benchmarks into question-answer pairs. Our method demonstrates a significant advantage, as shown in Tab. 4.

**Ablation Analysis** To more clearly observe the capabilities of GroundingFace in face understanding and segmentation, we conduct ablation studies on the face captioning and referring segmentation tasks, as shown in Tab. 2. ① **Key components in GroundingFace .** As shown in Tab. 2-Idx.4/5/6/7, each key component contributes to the model's final performance. ② **Different depth SAM features.** By default, we use the 6th and 24th SAM features. We explore the impact of using additional 6th, 12th, 18th, and 24th SAM features on the model and find that it leads to a performance drop despite the increased computational cost, as shown in Tab. 2-Idx.7/8. ③ **Fine-tuning stages.** The two-stage training strategy significantly outperforms the one-stage approach (Idx.3/6 in Tab. 2).

Table 2. Results on face caption (METEOR) and referring segmentation (gIoU). Our full model achieves obvious improvement over GLaMM baseline. The effect of each component is ablated.

| Idx. | Method | HQ-SAM Adapter Sec. 4.3 | HQ Facial Tokens Sec. 4.4 | MoE Sec. 4.5 | METEOR | gIoU |
|---|---|---|---|---|---|---|
| 1 | GLaMM | ✘ | ✘ | ✘ | 1.1 | 6.6 |
| 2 | Stage-1 | ✘ | ✘ | ✘ | 18.2 | 83.6 |
| 3 | Stage-1 | ✔ | ✔ | ✘ | 20.6 | 86.2 |
| 4 | Stage-2 | ✔ | ✘ | ✔ | 17.5 | 87.1 |
| 5 | Stage-2 | ✘ | ✔ | ✔ | 22.3 | 85.3 |
| 6 | Stage-2 | ✔ | ✔ | ✘ | 21.3 | 88.2 |
| 7 | Full Model | ✔ | ✔ | ✔ | **23.1** | **88.9** |
| 8 | 6/12/18/24th Feats. | ✔ | ✔ | ✔ | 22.3 | 88.1 |

Table 3. Results on pixel grounded face caption and VQA.

| Method | Grounded Caption | | | Grounded VQA | | |
|---|---|---|---|---|---|---|
| | METEOR | AP50 | mIoU | METEOR | AP50 | mIoU |
| GLaMM | 0.9 | 0.0 | 27.0 | 1.6 | 5.0 | 5.4 |
| Stage-1 | 19.8 | 30.6 | 48.6 | 18.3 | 52.1 | 73.2 |
| **Ours** | **21.9** | **32.9** | **52.0** | **20.6** | **53.7** | **76.9** |

Table 4. Performance of GroundingFace in zero shot face attribute recognition. Our method perform best in all datasets.

| Method | Params. | RAF-DB [28] | LFWA [33] | AgeDB | [39] |
|---|---|---|---|---|---|
| | | Emotion↑ | Attribute↑ | Gender↑ | Age↓ |
| Qwen-VL [2] | 7B | 42.9 | 0.0 | 93.9 | 11.44 |
| InstructBLIP [9] | 7B | 27.9 | 49.5 | 71.9 | 9.97 |
| LLaVA-v1.5 [30] | 7B | 55.4 | 58.3 | 98.5 | 10.22 |
| InternVL-v1.5 [7] | 26B | 67.2 | 61.1 | 94.4 | 19.26 |
| **Ours** | **7B** | **91.7** | **73.1** | **98.5** | **6.92** |

## 6. Conclusion

We introduce a novel FacePlayGround-240K dataset, the first large-scale, pixel-grounded face caption and question-answer (QA) dataset designed for alignment pretraining and instruction-tuning, which provides comprehensive captions and QA data rich in attribute relationships. Additionally, we propose the GroundingFace framework, enhancing fine-grained facial understanding through improved face part segmentation and attribute understanding.

**Acknowledgement.** This work is supported by the State Key Laboratory of Industrial Control Technology, China (Grant No. ICT2024A09).

## References


[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 2, 3

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 8

[3] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *WACV*, 2016. 3

[4] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL*, 2005. 8

[5] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 2

[6] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *ECCV*, 2024. 3

[7] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 2, 3, 8

[8] 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. 5

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2304.13509*, 2023. 8

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 6

[11] Face++. Face++ cognitive services. https://www.faceplusplus.com.cn/. Accessed: 2024-10-10. 2, 3

[12] Chongyang Gao, Kezhen Chen, Jinmeng Rao, Baochen Sun, Ruibo Liu, Daiyi Peng, Yawen Zhang, Xiaoyuan Guo, Jie Yang, and VS Subrahmanian. Higher layers need more lora experts. In *ICLR*, 2024. 7

[13] Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. Regiongpt: Towards region understanding vision language model. In *CVPR*, 2024. 2, 6

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 6

[15] Elisabeth Hui. Skin face detection model, 2023. Accessed: 2024-10-05. 3

[16] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 1991. 6, 7

[17] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-edit: Fine-grained facial editing via dialog. In *ICCV*, 2021. 5

[18] Alexander Kapitanov, Karina Kvanchiani, and Kirillova Sofia. Easyportrait - face parsing and portrait segmentation dataset. *arXiv preprint arXiv:2304.13509*, 2023. 3, 5

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 3, 5, 6

[20] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. In *NeurIPS*, 2024. 6

[21] Khalil Khan, Rehan Ullah Khan, Kashif Ahmad, Farman Ali, and Kyung-Sup Kwak. Face segmentation: A journey from classical to deep learning paradigm, approaches, trends, and directions. *IEEE Access*, 2020. 3

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 2, 3, 4

[23] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *CVPRW*, 2017. 3

[24] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, 2024. 2

[25] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 3, 5

[26] Dengchun Li, Yingzi Ma, Naizheng Wang, Zhiyuan Cheng, Lei Duan, Jie Zuo, Cal Yang, and Mingjie Tang. Mixlora: Enhancing large language models fine-tuning with lora based mixture of experts. *arXiv preprint arXiv:2404.15159*, 2024. 7

[27] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 2020. 3

[28] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*, 2017. 8

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024. 3, 8

[31] Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. When moe meets llms: Parameter efficient fine-tuning for multi-task medical applications. In *ACM SIGIR*, 2024. 7

[32] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024. 3

[33] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 8

[34] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 3

[35] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 3

[36] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *ECCV*, 2025. 2

[37] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 3

[38] Junho Moon, Haejun Chung, and Ikbeom Jang. Facial wrinkle segmentation for cosmetic dermatology: Pretraining with texture map-based weak supervision. *arXiv preprint arXiv:2408.10060*, 2024. 3

[39] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *CVPRW*, 2017. 8

[40] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. In *ICLR*, 2024. 2

[41] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *ICLR*, 2024. 3

[42] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 3

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 6

[44] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *CVPR*, 2024. 2, 3, 6, 8

[45] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, 2019. 3

[46] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016. 7

[47] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *CVPR*, 2021. 5

[48] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *ICCV*, 2023. 3

[49] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. CelebV-Text: A large-scale facial text-video dataset. In *CVPR*, 2023. 5

[50] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 3

[51] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 2, 6

[52] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *TPAMI*, 2017. 3

[53] 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 *CVPR*, 2022. 3, 5, 6