# SLME: Strokes Learning Model for Sketch-less Facial Image Retrieval

Anonymous ICME submission 3176

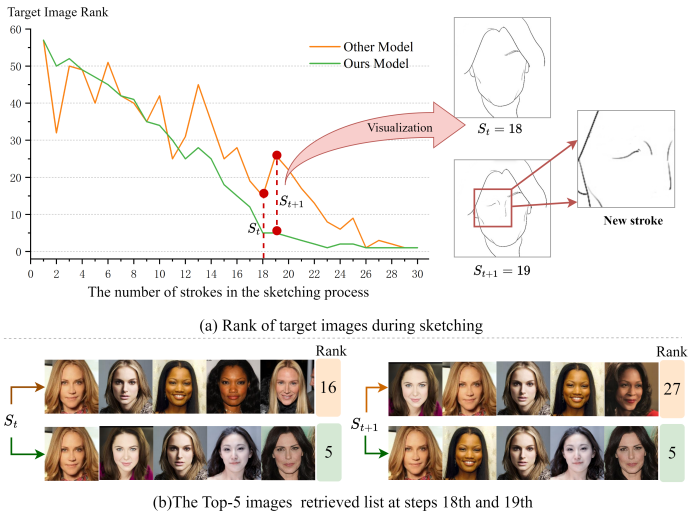
***Abstract*—Sketch Less Facial Image Retrieval (SLFIR) framework retrieves target images with fewer strokes through human-computer interaction, reducing reliance on high-quality sketch. The prevailing focus in SLFIR research lies on enhancing model representation learning; however, the potential for stroke accumulation during the sketching process to introduce noise and degrade retrieval performance is often overlooked. To address this, we introduce a Stroke Learning Model. This model is designed to adaptively identify informative sketch features based on incremental stroke input, thereby mitigating the negative effects of added, extraneous strokes on retrieval efficacy, particularly in the initial stages. Our methods involves: (1) Developing a Facial Alignment and Identity Preservation (FAIP) model to create robust and aligned embeddings for facial images and textual descriptions. (2) Constructing sketch retrieval models utilizing FAIP as a backbone architecture. These components leverage multi-scale feature analysis of both sketch and text to precisely identify salient features within cluttered sketches, enabling effective cross-modal matching. Extensive experiments demonstrates the significant gains in early retrieval accuracy achieved by our proposed model.**

***Index Terms*—Sketch-based image retrieval, Multimodal Fusion, Comparative learning, Mixture-of-Experts**

## I. Introduction

In recent years, sketch-based cross-modal facial recognition has seen widespread application across multiple domains [1]–[8]. For instance, in the field of law enforcement, facial sketches drawn from witness descriptions are frequently used to aid in suspect identification. However, traditional Sketch-Based Facial Image Retrieval frameworks require the use of complete and high-quality facial sketches for retrieval and recognition tasks. This imposes high demands on the skills of the sketch artist, while the process itself is complex and time-consuming, limiting its applicability to a broader user base.To overcome these limitations, Dai et al. [9] introduced the SLFIR framework. This framework dynamically integrates the drawing and retrieval processes, conducting image retrieval in real-time during sketching and providing feedback to the artist to inspire creativity. The objective is to simplify the operational process and enhance retrieval efficiency by using as few strokes as possible, allowing the retrieval of target images from incomplete and low-quality query sketches, thereby increasing its practicality across various application scenarios.

Nevertheless, as the sketching progresses, there is a significant increase in the number of strokes [10]–[14], which concomitantly amplifies the complexity of the image data, as demonstrated in Fig. 1 . Although this proliferation of strokes may enhance the granularity of the images, it potentially compromises the accuracy of the retrieval process prior to the precise delineation of key features. For example, as illustrated Fig. 1, an accumulation of brief and abstract strokes, if failing to comprehensively represent essential features like the eyes, may introduce noise. This noise can generate ambiguity during the model's matching phase, leading to a degradation in the ranking of the target image from 16 to 27. Such issues not only undermine the stability of the retrieval sequence but also impair the user experience. Thus, maintaining the stability of retrieval outcomes in scenarios where sketched features remain incomplete is an imperative challenge that needs addressing.




(a) Rank of target images during sketching

(b)The Top-5 images retrieved list at steps 18th and 19th

Fig. 1. Our method displays variabilities in the ranking of target images during retrieval from identical sketch samples, unlike other models. The x-axis illustrates the evolution of strokes, and the y-axis represents the ranking of retrieved images, where a lower value indicates a higher rank. Unlike counterparts that experience notable fluctuations during retrieval, our approach effectively identifies and removes superfluous strokes introduced by the user. This capability ensures minimal variation in retrieval outcomes, enhancing user experience. The accompanying diagram demonstrates the Top-5 target images retrieved at the 19th and 20th stages of the sketching process.

To address this issue, it is imperative to enhance the model's ability to recognize various types of strokes. We initially pre-trained a substantial multi-modal base model, designated as FAIP (Facial Alignment Pre-training with Image to Text), on a dataset encompassing multiple styles of facial sketches [15]. This training model is engineered to enhance the recognition capabilities for diverse styles of sketched images. By leveraging pre-training, the FAIP model markedly improves its accuracy in extracting pivotal facial features. Subsequently, utilizing the FAIP model as a foundational backbone, we develop the Stroke Learning Model(SLME). This model incorporates a hybrid expert system, leveraging contributions

from [16]–[22]. The core concept behind this system is both straightforward and robust: it segments the model into various experts, each specializing in specific tasks or data aspects. The heterogeneity of the MoE model aligns seamlessly with the dynamic characteristics of sketch features that evolve throughout the sketching process. This alignment is further enhanced by an efficient gate scheduling mechanism, which significantly bolsters the model's proficiency in processing disparate information. Consequently, this model maintains a focus on essential sketch features, effectively minimizing distractions caused by the accumulation of extraneous strokes as the sketch develops. This targeted approach ensures more precise and reliable sketch-based facial recognition across varying styles.

During the research process, we encountered a significant issue: in the initial phases of sketching, the strokes are sparse and often exhibit a degree of randomness, which hampers their capacity to effectively represent critical facial features. These preliminary, noise strokes can significantly impede the target recognition process. To address these challenges, we integrat textual descriptions of the sketches to counteract the disruptive effects of the randomness inherent in the initial strokes on the matching results.

As substantiated by both qualitative and quantitative evaluations conducted on multiple public datasets, our methodology has proven to be superior in early retrieval performance during the sketch creation process. In summary, our contributions are as follows:

(1) We introduce the FAIP to provide efficient embedding representations for SLFIR tasks by aligning facial images with corresponding textual descriptions. This approach enhances the accuracy and utility of facial recognition research across diverse contexts.

(2) We propose an innovative SLME model designed to identify key features and semantic information that align with the target image across different sketching processes. By strategically emphasizing these essential elements, the model effectively mitigates the decline in search performance caused by the accumulation of non-essential strokes.

(3) Extensive experimental validation shows that our proposed method significantly alleviates the performance degradation typically caused by the accumulation of unecessary strokes in sketches. These results underscore the efficiency and robustness of our approach in maintaining high retrieval accuracy throughout the sketching process.

## II. Methods

Accurate identification of key facial features is paramount for effective facial sketch retrieval tasks. We have developed a two-stage model to address this requirement. Initially, we employed the FAIP model (refer to Fig. 2) to bolster the recognition capabilities for facial features. Subsequently, to mitigate the influence of randomness in the current sketch strokes on retrieval efficacy, we integrated relevant knowledge to construct the SLME model (refer to Fig. 3).

### A. Backbone: Facial Image-Text Alignment Pre-training

**Architecture:** FAIP consists of two components: a text encoder and an image encoder enhanced by Sparse MoE [23]. ***Text Encoder***: $E_T(\cdot) \rightarrow \mathbb{R}^D$, which employs a standard 12-layer BERT architecture and extracts a summary of the text via the [CLS] token. To enhance multimodal interaction capabilities, we introduce a cross-modal attention layer between the self-attention layers and the feedforward network within the text encoder. This innovation enables the model to effectively integrate visual and textual information. Additionally, we have specifically introduced an [ENC] token for text input, the output embedding of which provides a foundation for the joint representation between image and text. Each text $T_i$ is processed by the text encoder $E_T(\cdot)$ to obtain an embedded vector $\tilde{v}_i^T$. This is then transformed through a fully connected layer $f(\cdot)$ into a low-dimensional feature vector $v_i^T$, which is compared with the low-dimensional feature vector of image.

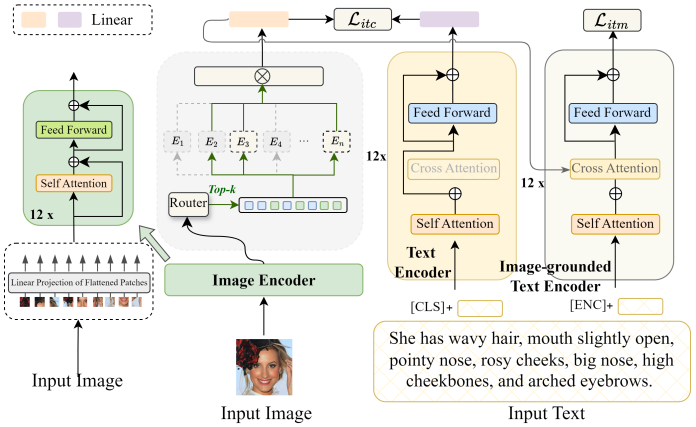



Fig. 2. Backbone: FAIP consists mainly of a text branch and an image branch. The image branch consists of an image encoder and MoE architecture.

***Vision Encoder***: The image branch primarily comprises a visual encoder $E_I(\cdot) \rightarrow \mathbb{R}^D$ and a Sparse MoE module $\mathfrak{M}(\cdot)$. It maps image content to a D-dimensional vector space. $E_I(\cdot)$ is constructed using a Visual Transformer (ViT) composed of 12 layers of Transformer encoder layers, specifically designed to process image information. Each image $I_i$ is processed through $E_I(\cdot)$ to obtain an embedded vector $\tilde{v}i^I$. Although features from the image are extracted as embedded vectors at this stage, these primarily represent global features and do not highlight key details or specific parts of the image. Therefore, the Sparse MoE module $\mathfrak{M}(\cdot)$ is connected after $E_I(\cdot)$. This module processes and optimizes the $\tilde{v}i^I$ vector using $M$ experts, dynamically adjusting the feature processing post-image encoding to enhance the expression capability for complex data. This process includes top-level gating selection and noise augmentation strategies aimed at enhancing overall model performance. The vector processed by $\mathfrak{M}(\cdot)$ is then transformed by $f(\cdot)$ to yield $v_i^I$, mapped to the same dimensionality as $v_i^T$.

**Loss Functions**: To enhance the alignment precision between images and text, we employed Image Text Correspondence

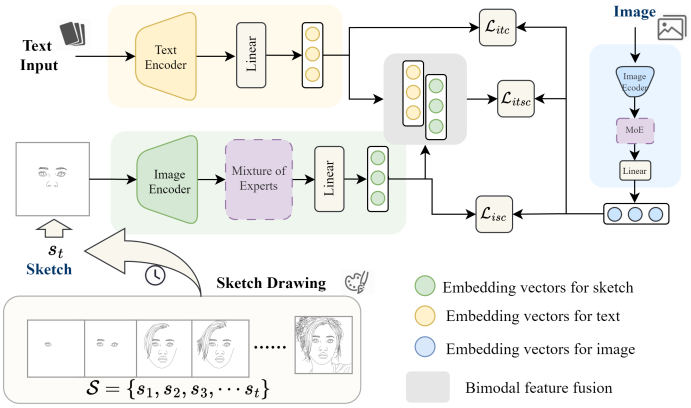



Fig. 3. Model Architecture. Using FAIP as the backbone, ours model essentially consists of two separate branches the Image Branch, encapsulating an image encoder and a MOE model, and the Text Branch comprising of a text encoder.

(ITC) Loss [24] and Image Text Matching (ITM) Loss [25], [26] to train the model. This approach optimizes the model's performance and ensures its ability to precisely match and retrieve relevant images.

### B. Stroke Learning a Mixture-of-Experts Model

**Architecture:** The inputs of ours model: sketch sequence and textual description. A sequence of $t$ incomplete sketches constitutes the sketching process $\mathcal{S} = \{s_1, s_2, \ldots, s_t\}, \mathcal{S} \in \mathcal{D}^S$. The textual description $T \in \mathcal{D}^T$. $\mathcal{S}$ and $T$ form a text-sketch sequence pair $\mathcal{P} = \{(s_1, T), (s_2, T), \ldots, (s_t, T)\}$. The encoders $E_I(\cdot)$ and $E_T(\cdot)$ from the FAIP model are retained as the basic encoders. At time $t$, the sketch $s_t$ and text $T$ are respectively processed by $E_I(\cdot)$ and $E_T(\cdot)$, resulting in $\tilde{v}_t^S = E_I(s_t)$ and $\tilde{v}^T = E_T(T)$. Subsequently, the embedding vector of the sketch $\tilde{v}_t^S$ is further processed by MoE module $E_M(\cdot)$ to extract key features of the sketch. Thereafter, it is passed through the fully connected layer $f(\cdot)$ to map the image to a low-dimensional vector $v_t^S$, which is used for distance calculation with the image. $\tilde{v}^T$ is directly passed through the Linear layer $f(\cdot)$ to generate a low-dimensional vector $v^T$ for similarity calculation [24] with the image. To tackle the issue of sparse initial strokes in sketches, we combined two vectors as shown in Fig. 3. This integration occurs in the Bimodal Fusion Feature module, producing a composite feature $\tilde{v}_t^{\wp}$. This feature is then processed by $f(\cdot)$ to yield a low-dimensional vector $v_t^{\wp}$ for further image computation.

**Training Procedure**: We employed a contrastive learning approach to train ours model, aiming to enhance the similarity within positive sample pairs while reducing it among negative sample pairs. Specifically, we implemented three types of loss functions for optimization: the ITC loss [24] for text-image pairs, the ISC loss for sketch-image pairs(Eq. 1), and the ITSC loss(Eq. 2) for integrated sketch-image pairs. This strategy significantly improved the model's discriminative ability across various types of data pairs. In detail, the ISC loss calculates the distance between a sketch and its corresponding facial image $(s_i, I)$, mapping them to a shared feature space, thereby effectively enhancing feature consistency and robustness.

$$\mathcal{L}_{isc} = -\sum_{i=1}^{N} \log \frac{\exp(sim(v_i^s, v_i^I)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(v_i^s, v_j^I)/\tau)} \tag{1}$$

ITSC loss is calculated by assessing the similarity between the features $v_t^{\wp}$, obtained from the fusion of text and sketches, and the images, as demonstrated in Eq. 2.

$$\mathcal{L}_{itsc} = -\sum_{i=1}^{N} \log \frac{\exp(sim(v_t^{\wp}, v_i^I)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(v_t^{\wp}, v_j^I)/\tau)} \tag{2}$$

where, $N$ denotes the number of samples in a batch. $\tau$ is a temperature parameter used to calibrate the scale of similarity scores.

## III. Experiment

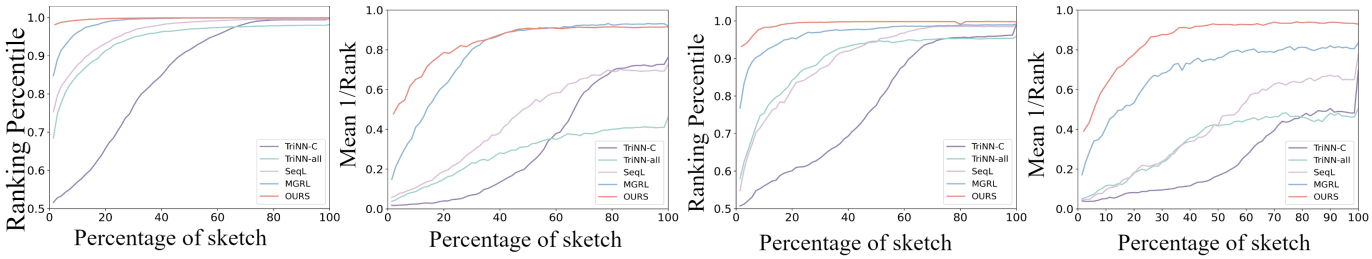
**Datasets**: CelebAMask-HQ [15] is a large-scale facial image dataset consisting of 30,000 high-resolution facial images selected from the CelebA dataset, following the CelebA-HQ protocol. FS2K-SDE1 [9] comprises 75,530 sketches and 1,079 images for training, with the remaining portion for testing; FS2K-SDE2 consists of 23,380 sketches and 334 images for training, and the leftover sketches and images are used for testing.

**Evaluation Metrics**: Our aim is to retrieve the target image using the fewest strokes in the minimal amount of time, with a focus on promptly surfacing those images at the top of the results list. To evaluate this, we employ m@A (the ranking percentile) [3] and m@B (the reciprocal of the ratio between the ranking and the sketch completion percentage) [3] as metrics. These indicators allow for a comprehensive assessment of the average retrieval performance at each phase of the sketching process. For early sketch retrieval performance, we utilize W@mA [9] and W@mB [9], which provide greater weight to earlier sketches. Additionally, the Top-$k$ metric is used to quantitatively evaluate the model's retrieval efficacy for fully completed sketches.

**Implementation detail**: (1) Our computational setup includes an RTX 4090 GPU and a batch size of 16. We initiated training with pre-trained weights from FVLP [27], leveraging the CelebAMask-HQ dataset [15] and employing AdamW as the optimizer. The configuration included eight expert modules, with a k value set at 4, and the learning rate was managed via a cosine scheduling strategy. (2) Training the model proceeded over 50 epochs, utilizing the AdamW optimizer with a weight decay factor of 0.05. The model architecture featured four expert networks in the feature observer, setting the parameter k at 2. A cosine annealing learning rate schedule was applied, allowing the rate to vary from 3e-5 to zero. The output dimension was configured to 256.

### A. Comparison With the State-of-the-Art Methods

We evaluate the performance of SLME in the initial phases of sketch-based retrieval both subjectively (see Fig. 4) and objectively (see Table I). As depicted in Fig. 4, the retrieval




(a) FS2K-SDE1 (b) FS2K-SDE2

Fig. 4. SLME method demonstrates a significant performance advantage over baseline methods in early-stage sketch retrieval. The horizontal axis denotes the percentage of strokes, and the vertical axis measures the retrieval metrics, with higher values indicating enhanced performance in early-stage retrieval.

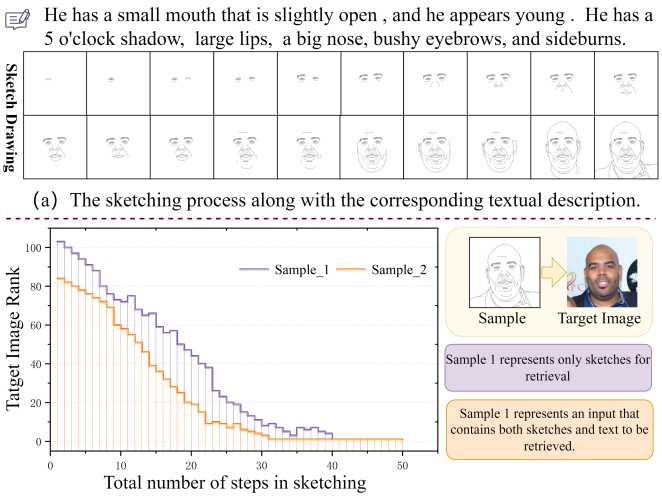
TABLE I
COMPARATIVE RESULTS WITH OTHER BASELINE METHODS ON FS2K-SDE DATASETS.

| Model | FS2K-SDE1 | | | | | FS2K-SDE2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | m@A | m@B | w@mA | w@mB | Top-5 | m@A | m@B | w@mA | w@mB | Top-5 |
| **TriNN-C**(2016 [28]) | 84.77 | 32.69 | 50.40 | 15.83 | 91.33 | 77.83 | 24.59 | 46.00 | 12.47 | 94.41 |
| **TriNN-all**(2020 [3], [28]) | 94.16 | 28.58 | 58.18 | 15.83 | 64.22 | 89.77 | 34.14 | 54.99 | 18.96 | 69.23 |
| **TriNN-all-RL**(2022 [3]) | 84.42 | 22.76 | 51.52 | 12.21 | 51.78 | 85.65 | 26.70 | 51.91 | 14.59 | 69.23 |
| **SeqL**(2023 [9]) | 96.22 | 45.48 | 59.57 | 24.56 | 90.00 | 90.22 | 41.55 | 54.85 | 22.22 | 95.82 |
| **MGRL**(2024 [29]) | 98.80 | 78.92 | 61.69 | 46.20 | 97.11 | 96.65 | 69.19 | 60.12 | 40.80 | 95.10 |
| **MITRL**(2024 [30]) | 99.70 | 80.48 | 62.33 | 48.47 | 98.66 | 97.82 | 70.02 | 60.82 | 41.06 | 96.50 |
| **Clip-based***(2021 [31]) | 98.01 | 57.27 | 60.97 | 32.58 | - | 96.02 | 57.32 | 59.38 | 33.36 | - |
| **Blip-based***(2022 [26]) | 97.64 | 46.72 | 60.80 | 27.01 | - | 97.10 | 64.06 | 60.24 | 37.50 | - |
| **FVIP-based***(2024 [27]) | 98.14 | 56.70 | 61.16 | 32.58 | - | 95.41 | 59.40 | 58.99 | 34.37 | - |
| **TAIP-based*** | 98.56 | 58.97 | 61.47 | 34.06 | - | 96.75 | 63.47 | 60.01 | 37.02 | - |
| **Ours** | **99.58** | **82.21** | **62.07** | **49.72** | **98.68** | **98.36** | **82.41** | **61.07** | **49.38** | **97.23** |

performance of all examined methods improves and stabilizes as the sketch nears completion. Notably, in the early stages of sketching, our model demonstrates significantly superior retrieval results compared to other methods. To accurately assess the performance of our model, we conducted comparative analyses with the current baseline methods.

(1) Table I illustrates the comparative performance between our model and other baselines (FAIP-Based*: where the encoder weights of FAIP are fixed, and only the last fully connected layer is trained). The results indicate that our model exhibits marked superiority during the initial phases of retrieval, especially when contrasted with the baseline utilizing a contrastive learning strategy. This advantage arises primarily because facial features are not distinctly expressed in early sketch stages. Sole reliance on the triplet net strategy to minimize the distance between the sketch and positive samples might result in mismatches due to sparse sketches, thereby disturbing the model's optimization direction and diminishing retrieval accuracy. Additionally, the incorporation of text descriptions in our model enhances the precise definition of image features. Under identical fine-tuning conditions, FAIP outperforms the LLMS model and demonstrates significant improvements over various other contrastive learning methods. This further underscores FAIP's efficacy in image and text embedding representation.

(2) In order to delve deeper into the stability characteristics of our model's feedback in the sketch handling process, we have conducted research and demonstrated the evolution of the target image results throughout the retrieval process for




(a) The sketching process along with the corresponding textual description.

(b) Sample (sketch) in the sketching process to get the target image to retrieve the results (ranking) change curve.

Fig. 5. The changing state of the target image search ordering of sketch samples during retrieval. The left side of the line graph represents the ranking of the target image, the lower the curve indicates, the higher the ranking of the target image.

a set of samples. The specific details are depicted in Fig. 5. Fig. 5.a accurately portrays the creation process of a sketch and provides the corresponding textual description. Fig. 5.b then reveals the transformation of the target image positioning during the retrieval procedure. The horizontal axis reflects the total number of steps necessary to complete the sketch, whereas the vertical axis represents the ranking or position of the target image; a smaller number indicates a higher

ranking. We have compared the search results of solely using the sketch and of incorporating both text and sketch inputs simultaneously. The results exhibit significant enhancement in retrieval performance when a double input approach, relative to relying solely on sketch search, is applied. The gradual decline of the curves throughout the entire sketch sequence retrieval process, as manifested by the changing dynamics of the two curves, suggest an extremely stable overall retrieval process without significant fluctuations. Regardless of whether single-mode or dual-mode input was adopted, the display of the curves in various situations convincingly demonstrates the robust stability inherent in our proposed method, proving to be a powerful safeguard for user retrieval experience.

TABLE II
ARCHITECTURAL ABLATION.

| Dataset | Model | m@A | m@B | w@mA | w@mB |
|---|---|---|---|---|---|
| FS2K-SDE1 | FAIP_based* | 98.56 | 58.97 | 61.47 | 34.06 |
| | FAIP_based | 99.51 | 76.88 | 61.36 | 45.95 |
| | FAIP_based+MoE | 98.33 | 79.42 | 61.72 | 46.97 |
| | Ours | **99.58** | **82.21** | **62.07** | **49.72** |
| FS2K-SDE2 | FAIP_based* | 96.75 | 63.47 | 60.01 | 37.02 |
| | FAIP_based | 97.86 | 79.04 | 60.24 | 45.31 |
| | FAIP_based+MoE | 98.00 | 80.39 | 60.98 | 48.50 |
| | Ours | 98.36 | **82.41** | **61.07** | **49.38** |

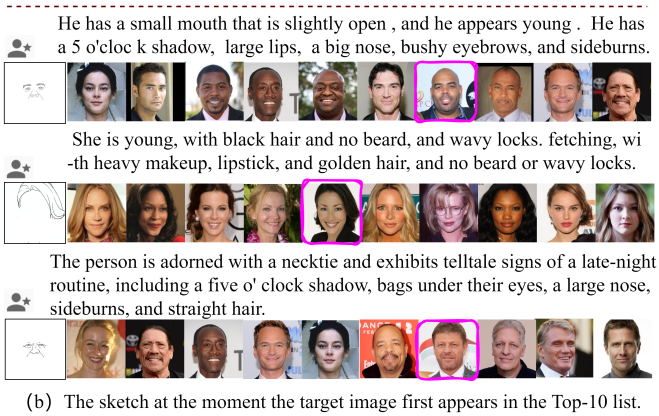
### B. Ablation Study

We conducted an ablation study on the different components of the ours model, including the training model's presence, the existence of MoE, and the application of ITSC loss. As shown in Table II, the experimental results indicate that our model outperforms tasks that only use FAIP for fine-tuning, regardless of the configuration. By incrementally adding components, we observed a significant increase in the performance of all evaluation metrics. It is particularly noteworthy that the introduction of MoE significantly enhances the feature recognition ability during the initial sketching and improves early retrieval efficiency. With the integration of text information and the application of ITSC loss, the integrity of the early stage of the sketch is significantly enhanced, thereby further improving the model's retrieval performance.

### C. Comparison under different modal inputs

To mitigate the inherent sparsity of initial sketches, we supplement our approach with auxiliary textual semantics to enhance image representation learning. We performed a comprehensive multi-modal analysis to rigorously evaluate our proposed method in the single-modality (sketch-only) setting. As shown in Table III, performance comparisons were conducted under both sketch-only and combined sketch+text input scenarios. Our method demonstrates superior retrieval performance even in the absence of textual augmentation. Notably, in the sketch-only condition, our approach surpasses all benchmarked methods. Furthermore, when incorporating both textual and sketch modalities, our method attains state-of-the-art performance. These findings highlight the robustness

(a) The sketch at the moment the target image first appears in the Top-10 list. (Sketch)




(b) The sketch at the moment the target image first appears in the Top-10 list. (Sketch+Text)

Fig. 6. (a) The sketch at the moment when the target image initially appears in the Top-10 list (Input sketch only).(b) The sketch at the moment when the target image initially appears in the Top-10 list (Input sketch and text).

TABLE III
EXPERIMENTS WITH AND WITHOUT TEXT FUSION BASE MODEL ON DATASET FS2K-SDE1.

| Input | Model | FS2K-SDE1 m@A | m@B | w@mA | w@mB |
|---|---|---|---|---|---|
| Sketch | TriNN-C(2016) | 84.77 | 32.69 | 50.40 | 15.83 |
| | TriNN-all(2020) | 94.16 | 28.58 | 58.18 | 15.83 |
| | TriNN-all-RL(2022) | 84.42 | 22.76 | 51.52 | 12.21 |
| | SeqL(2023) | 96.22 | 45.48 | 59.57 | 24.56 |
| | MITRL(2024) | 97.34 | 77.15 | 59.73 | 37.21 |
| | MGRL(2024) | 98.80 | 78.92 | 61.69 | **46.20** |
| | Ours | **98.81** | **78.93** | **61.72** | 44.20 |
| Sketch+Text | Clip-based*(2021) | 98.01 | 57.27 | 60.97 | 32.58 |
| | Blip-based*(2022) | 97.64 | 46.72 | 60.80 | 27.01 |
| | SeqL(2023) | 96.22 | 45.48 | 59.57 | 24.56 |
| | MITRL(2024) | **99.70** | 80.48 | **62.33** | 48.47 |
| | Ours | 99.58 | **82.21** | 62.07 | **49.72** |

of our proposed approach in augmenting sketch-based image retrieval, regardless of the presence of textual descriptions.

Subsequently, we illustrate our findings with specific case examples. Figure 6 displays three experimental samples, presenting the incomplete sketches at the exact moments when the target image first enters the top ten rankings, irrespective of textual input. The target image is highlighted within a red box. It is evident from the figure that our method efficiently identifies the target images with minimal strokes, thereby offering users an optimal retrieval experience.

## IV. CONCLUSION

The SLFIR framework is designed to retrieve the target image swiftly using the minimal number of strokes. Given the challenges associated with stroke identification in SLFIR, we have developed a model (SLME) that accurately rec-

ognizes effective strokes. The proposed model offers stable feedback throughout the sketching process. Extensive experimental results demonstrate that our method exhibits strong early retrieval performance and further ensures stable output across multiple modalities. However, there are critical issues that need addressing, including: (1) The issue of diminished retrieval performance due to sparse strokes in the early stages of sketching; (2) The need to enhance the interaction between humans and machines for improved efficiency.

## REFERENCES


[1] Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song, "What can human sketches do for object detection?," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 15083–15094.

[2] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song, "Text-to-image diffusion models are great sketch-photo matchmakers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16826–16837.

[3] Ayan Kumar Bhunia, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song, "Sketch less for more: On-the-fly fine-grained sketch-based image retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9779–9788.

[4] Zhengyu Huang, Yichen Peng, Tomohiro Hibino, Chunqi Zhao, Haoran Xie, Tsukasa Fukusato, and Kazunori Miyata, "dualface: Two-stage drawing guidance for freehand portrait sketching," *Computational Visual Media*, vol. 8, pp. 63–77, 2022.

[5] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song, "Stylemeup: Towards style-agnostic sketch-based image retrieval," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8504–8513.

[6] Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song, "Scenetrilogy: On human scene-sketch and its complementarity with photo and text," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10972–10983.

[7] Hmrishav Bandyopadhyay, Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain, Tao Xiang, and Yi-Zhe Song, "What sketch explainability really means for downstream tasks?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10997–11008.

[8] Dar-Yen Chen, Ayan Kumar Bhunia, Subhadeep Koley, Aneeshan Sain, Pinaki Nath Chowdhury, and Yi-Zhe Song, "Democaricature: Democratising caricature generation with a rough sketch," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8629–8639.

[9] Dawei Dai, Yutang Li, Liang Wang, Shiyu Fu, Shuyin Xia, and Guoyin Wang, "Sketch less face image retrieval: A new challenge," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[10] Yingge Liu, Dawei Dai, Kenan Zou, Xiufang Tan, Yiqiao Wu, and Guoyin Wang, "Prior semantic-embedding representation learning for on-the-fly fg-sbir," *Expert Systems with Applications*, p. 124532, 2024.

[11] Weikang He, Yunpeng Xiao, Tun Li, Rong Wang, and Qian Li, "Interest hd: An interest frame model for recommendation based on hd image generation," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[12] Kwan Yun, Kwanggyoon Seo, Chang Wook Seo, Soyeon Yoon, Seongcheol Kim, Soohyun Ji, Amirsaman Ashtari, and Junyong Noh, "Stylized face sketch extraction via generative prior with limited data," in *Computer Graphics Forum*. Wiley Online Library, 2024, p. e15045.

[13] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song, "How to handle sketch-abstraction in sketch-based image retrieval?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16859–16869.

[14] 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 *International conference on machine learning*. PMLR, 2023, pp. 19730–19742.

[15] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5549–5558.

[16] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[17] Yongming Rao, Wenliang Zhao, Yansong Tang, Jie Zhou, Ser Nam Lim, and Jiwen Lu, "Hornet: Efficient high-order spatial interactions with recursive gated convolutions," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10353–10366, 2022.

[18] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby, "Multimodal contrastive learning with limoe: the language-image mixture of experts," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9564–9576, 2022.

[19] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi, "Neighborhood attention transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6185–6194.

[20] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi, "Oneformer: One transformer to rule universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2989–2998.

[21] Tianlong Chen, Xuxi Chen, Xianzhi Du, Abdullah Rashwan, Fan Yang, Huizhong Chen, Zhangyang Wang, and Yeqing Li, "Adamv-moe: Adaptive multi-task vision mixture-of-experts," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17346–17357.

[22] Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, et al., "On the representation collapse of sparse mixture of experts," *Advances in Neural Information Processing Systems*, vol. 35, pp. 34600–34613, 2022.

[23] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al., "Mixtral of experts," *arXiv preprint arXiv:2401.04088*, 2024.

[24] 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 *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[25] 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," *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021.

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12888–12900.

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

[28] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales, "Deep spatial-semantic attention for fine-grained sketch-based image retrieval," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5551–5560.

[29] Liang Wang, Dawei Dai, Shiyu Fu, and Guoyin Wang, "Multi-granularity representation learning for sketch-based dynamic face image retrieval," *arXiv preprint arXiv:2401.00371*, 2023.

[30] Dawei Dai, Yingge Liu, Shiyu Fu, and Guoyin Wang, "Multimodal image-text representation learning for sketch-less facial image retrieval," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.

[31] 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 *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.