Svedberg Open
DISSEMINATIONOFKNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/



Research Paper Open Access

Narrative-Driven Composite Facial Sketch Generation Using Dual Semantic and Attribute Conditioning

Aishwarya M[1], Pooja M R[2]

[1]Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Affiliated to VTU, Belagavi, Mysuru, 570002, India.E-mail: aishwarya.m@vvce.ac.in
[2]Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Affiliated to VTU, Belagavi, Mysuru, 570002, India.E-mail: pooja.mr@vvce.ac.in

Abstract
Forensic identification frequently depends on eyewitness accounts when no photograph of a suspect exists. Converting natural-language facial descriptions into composite sketches is traditionally manual, artist-dependent, and inconsistent. Although recent generative models support text-to-image synthesis and photo-to-sketch translation, narrative-driven composite sketch generation remains insufficiently addressed, particularly in terms of explicit facial-attribute preservation. This study presents an attribute-guided conditional generative adversarial network that converts eyewitness narratives into composite facial sketches through dual semantic and attribute conditioning. A frozen CLIP ViT-B/32 encoder provides a semantic representation of the narrative, while a rule-based parser extracts a 40-dimensional facial-attribute vector. Both representations jointly condition the generator, while a matching-aware discriminator evaluates image–condition compatibility. Experiments on Multi-Modal-CelebA-HQ demonstrate an attribute macro-F1 of 0.382, weighted-F1 of 0.520, CLIP narrative consistency of 0.248.Fine-grained analysis shows stronger reproduction of coarse attributes such as gender, age, and beard characteristics, whereas fine and rare attributes exhibit lower recall. The findings demonstrate the feasibility of narrative-to-composite-sketch generation with attribute-level evaluation and identify fine-attribute recall as a key challenge.

Keyword composite facial sketch, narrative-to-image generation, conditional GAN, CLIP, facial attribute preservation, forensic imaging, attribute-level evaluation.

1. Introduction

The facial composite is the image of the individual's face made based on the witness's memory and is widely used in criminal investigations if the photo of the person is not available. Traditionally, the facial composites are generated either by the professionals who are capable of drawing, or through the software that is designed to assist with the feature assembly. Despite being helpful for recognition, these methods may be affected by the skills of the operator, witness's ability to depict facial features, as well as the time needed to make a composition. Thus, the method is prone to being highly subjective and complicated to repeat in various situations.

The natural-language description of a person provides a more direct depiction of the individual's image in the witness's memory. The person may refer to a man as "a young man with a beard and short black hair" without having to select the facial components on the screen. In addition, one can include several facial attributes in one phrase and combine them differently. Thus, the transformation of natural-language description into the facial image becomes possible [19].

On the other hand, sketch generation from narrative is more than just a traditional text to image task. A narrative can contain both high-level semantic information and low-level facial features that need to be incorporated into the generated face. For instance, a description can include age, gender, hair, facial hair, expression, and accessories all at the same time. A face generated through this process can be visually accurate but not necessarily preserving one or more explicitly stated features. The approach, therefore, has to take into account both the meaning conveyed by the description and the facial features mentioned within it.

Facial descriptions also tend to be quite variable in nature. Different witnesses can employ varying language to describe the same feature. Also, a certain feature can be stated explicitly, implicitly, or in connection with other features. A generative model has to convert such variable linguistic descriptions into a fixed representation without any loss of characteristic that will help identify the described face.

Recent advances in multimodal representation learning have enabled models to connect textual and visual information within a shared semantic space. Such representations are useful for understanding the overall meaning of facial descriptions, but semantic similarity alone does not guarantee preservation of individual characteristics. Attributes such as gender, age-related appearance, facial hair, hairstyle, and eyewear may have different levels of importance and may not be adequately represented by a single global embedding.

Explicit attribute representations provide a structured description of the characteristics that a generated face should contain. However, relying exclusively on predefined attributes can restrict the representation of natural-language descriptions because contextual and semantic information may not be captured by independent attribute labels. This motivates the use of complementary global semantic and structured attribute representations.

In this work, we address this issue through dual semantic and attribute conditioning. The proposed framework uses two complementary representations of the input narrative. First, a frozen CLIP ViT-B/32 text encoder produces a 512-dimensional semantic embedding capturing the overall meaning of the description. Second, a rule-based parser extracts a 40-dimensional facial-attribute vector corresponding to predefined facial characteristics. These representations are concatenated and projected into a shared conditioning space before being supplied to the generative network. This enables the model to incorporate both broad narrative semantics and explicit facial properties.

The conditioning information is incorporated into the generator through condition-dependent feature modulation. Rather than providing the representations only at the input stage, the conditioning information influences intermediate feature transformations during generation. The generator progressively transforms a latent representation into a 128 × 128 facial sketch while the conditioning information guides visual features at different stages.

The discriminator model also takes into account the conditioning information. In addition to assessing both the real and generated sketches, it gets the corresponding semantic and attributes conditions. The mismatched conditions are intentionally provided during the training process in order to teach the discriminator to discriminate between the correctly paired images/descriptions and mispaired.

The training objective consists of adversarial, reconstruction and semantic terms. The L1 loss is used to encourage the generation of the face similar to the target sketch, and the CLIP-based objective is used for encouraging the semantic alignment with the input narrative. Moreover, the gradient regularisation and gradient clipping methods are used for improving the training process.

There is another problem to deal with - how to check if the sketch generated contains the described attributes. The overall image similarity is not enough to assess the attributes preservation because the image can be quite similar to the target one but can miss some facial attributes. That is why the following aspects will be considered in the evaluation of the generated sketches: structural, pixel level, perceptual, semantic, and attribute level.

However, all of the current works on the generation of facial sketches have focused on the task of generating sketches from photos, sketches from attributes, or other image condition sketch synthesis methods. The present work differs in that it uses natural language descriptions as the input and aims to produce a composite sketch from these. Therefore, a model designed to link natural language descriptions to facial sketches would be appropriate.

Thus, the aim of this research is to explore the possibilities of generating composite facial sketches from natural language narratives by analyzing their ability to preserve facial attributes. This will help us determine if using the combination of global semantics of natural language and structure of facial attributes could help generate sketches under certain conditions.The main contributions of this work are as follows:

1. Narrative to Composite Sketch Model: A conditional generative approach is created to translate unconstrained facial narrations into composite sketches of faces.
2. Semantic and Attribute Conditions: Narrative semantics are extracted from CLIP model globally while an explicit representation of 40 facial attributes is used as condition for sketch generation.
3. Condition Dependent Generation and Discrimination: Generator uses the technique of condition dependent feature modulation while discriminator employs two conditions – a matching one and a mismatching one.
4. Evaluation at Multiple Levels and on Attributes: Generated sketches can be evaluated structurally, perceptually, semantically, and regarding facial attributes.

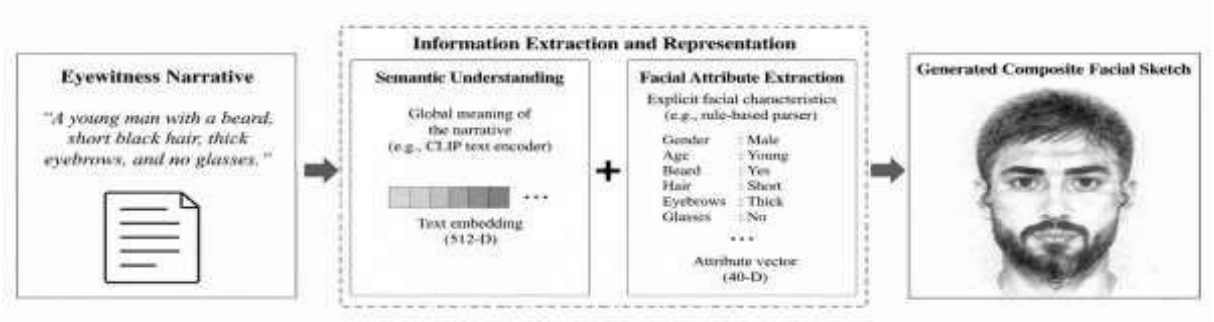


Figure 1. Conceptual overview of narrative-driven composite facial sketch generation

2. Related Work

2.1 Facial Sketch Synthesis

Facial sketch synthesis has been generally considered as an image-to-image translation problem where the input facial photo is transformed into a sketch form. The recent studies have mainly concentrated on enhancing the structural similarity, local facial attributes, maintaining the facial identity and generating different sketch styles. Fan et al. have conducted a detailed investigation into facial sketch synthesis and have proposed the FS2K benchmark along with the evaluation of various existing methods. They found the significance of facial-aware representation and highlighted some challenges of sketch synthesis but the problem still remains image-conditioned not narrative-conditioned [1].

Zhong et al. proposed an unsupervised self-attention-based lightweight framework for photo-to-sketch synthesis, using self-attention to capture facial textures while reducing dependence on paired training data. The approach improves efficiency and generalization across multiple databases, but its input remains a facial image rather than a natural-language description [2].

IsGAN introduced identity-sensitive adversarial learning for photo-sketch synthesis. The method incorporates identity recognition loss and cyclic synthesis constraints to preserve identifiable facial information during cross-domain translation. Although identity preservation is valuable for facial composites, the method assumes that the input face image is already available and therefore does not address the narrative-to-sketch problem [3].

Chen et al. proposed a semi-supervised CycleGAN framework for face photo-sketch translation in unconstrained settings. Their approach reduces dependence on large collections of paired photo-sketch samples by exploiting additional unpaired facial images. The work addresses dataset limitations and cross-domain translation, but it does not interpret textual descriptions or explicitly evaluate whether described facial attributes are preserved [4].

Shi and Wan proposed a Transformer-based adversarial network for semi-supervised face sketch synthesis. Adaptive window attention is used to capture both local and long-range facial dependencies, while a Laplacian-based detail extractor improves facial detail representation. The method demonstrates the value of Transformer-based feature modeling for sketch synthesis, but again operates from facial images rather than eyewitness narratives [5].

Chen et al. addressed fine-granularity face sketch synthesis by decomposing the face into multiple facial components and generating them in parallel. A detail-repair network is subsequently used to correct mismatches and deformation between components. This work is particularly relevant to the preservation of individual facial structures, but its conditioning remains image-based and does not use natural-language semantics [6].

A 2025 study introduced facial-structure sampling with contrastive learning for sketch synthesis. The method focuses on improving structural representation and cross-modal feature alignment during photo-to-sketch generation. Although contrastive learning is relevant to semantic alignment, the task is still defined as generating sketches from photographs rather than from narrative descriptions [7].

More recently, multi-level guided latent diffusion has been investigated for generating facial sketches from photographs in multiple sketch styles. The approach exploits diffusion models to improve content and style coverage, addressing limitations of GAN-based single-style synthesis. Its focus, however, remains photo-conditioned sketch generation and does not explicitly address the linguistic interpretation of eyewitness descriptions [8].

2.2 Text-to-Sketch and Narrative-Driven Generation

A particularly relevant study is Text-to-Sketch Synthesis via Adversarial Network, published in 2023. The authors explicitly investigate facial sketch generation from textual descriptions using a structured adversarial network and introduce inverse gamma correction to improve sketch quality. This work establishes that textual facial descriptions can be used as input for sketch generation and therefore represents an important predecessor to the present study. However, its textual representation is based on abstract facial descriptions and does not employ the dual semantic–attribute conditioning adopted in our framework. Furthermore, its evaluation does not provide the same explicit 40-attribute preservation analysis used in this study [9].

Weng et al. proposed a stepwise synthetic face-sketch generation algorithm based on GANs. The method progressively constructs facial sketch representations rather than generating the complete result in a single step. Such decomposition is useful for handling facial structure, but the approach does not focus on interpreting free-form eyewitness narratives through a multimodal semantic representation [10].

Ke et al. introduced SketchDiffusion, a text-guided image-to-sketch diffusion framework designed to produce human-like freehand sketches. The work demonstrates the potential of diffusion models and text-guided control for sketch generation and constructs a large prompt dataset using automatically generated descriptions. However, the task covers general sketch synthesis rather than specifically generating facial composites from eyewitness descriptions, and its evaluation is not centered on explicit facial-attribute preservation [11].

A 2025 study examining generative neural networks for police facial sketches investigated GANs, variational autoencoders, and diffusion models for automated composite-sketch generation. The work highlights the potential of generative AI for police sketch production and text-based facial description interpretation. However, it is primarily an exploration of generative technologies rather than a dedicated dual-conditioned architecture with quantitative attribute-level validation [12].

2.3 Text-to-Face Generation and Attribute Control

Text-to-face synthesis is closely related to narrative-driven sketch generation because both tasks require translating linguistic descriptions into facial characteristics. An embedded method for improving text–face relevance introduced enhanced face-attribute representations and local feature extraction to strengthen the association between textual descriptions and facial regions. Experiments on Multi-Modal-CelebA-HQ showed improvements over several text-to-image baselines. Nevertheless, the target domain is photographic face generation rather than composite facial sketches [13].

Wang et al. investigated text-to-face synthesis through facial landmark prediction. Their work recognizes that textual facial descriptions contain fine-grained appearance information and introduces landmarks as an intermediate structural representation. This is closely related to the idea of using an explicit facial representation, but the method targets photographic face synthesis and does not combine a global CLIP narrative embedding with an explicit facial-attribute vector for sketch generation [14].

Guo and Gu proposed an improved StyleGAN-based TextToFace model using local–global information fusion. Their work addresses the limitation of relying solely on global sentence embeddings, which can cause facial details described in text to be lost. Attention-based semantic mapping and similarity-based classification are used to improve semantic detail and identity consistency. This is particularly relevant to our motivation because it demonstrates the difficulty of preserving fine facial information from textual descriptions. However, the output remains a photorealistic face rather than a composite sketch [15].

Song et al. proposed AttriDiffuser, an adversarially enhanced diffusion framework for text-to-facial-attribute image synthesis. The model introduces attribute-gating cross-attention, global/local feature alignment, and a diversity discriminator to improve the representation of multiple attributes in generated faces. The study is highly relevant to explicit attribute control, but its objective is high-fidelity photographic face synthesis and editing rather than narrative-driven composite-sketch generation [16].

Dong et al. proposed a text-guided multi-attribute embedded diffusion model for face personalization at fine-grained levels. The proposed model directly addresses the issue of loss of facial attributes during text-to-image generation and makes use of multi-attribute embeddings for better controllability. It highlights the significance of fine-grained attribute representations; however, it only applies to personalized photographic faces and does not address sketch-based generation based on eyewitness narratives [17].

A model that was proposed in 2025 based on landmarks and diffusion, used landmarks along with heat map guided denoising to facilitate controllable text-to-face generation. Structural guidance was employed for better control over facial features during diffusion. Although the model highlights the significance of using both semantic and structural features for controlling the generated images, it applies to the field of photographic faces [18].

2.4 Research Gap

Existing facial sketch-generation research mainly focuses on photo-to-sketch translation or text-to-face synthesis, with limited attention to narrative-driven composite sketch generation. While text-to-sketch methods demonstrate the feasibility of using textual descriptions, they generally do not jointly model global narrative semantics and explicit facial attributes. Global embeddings capture overall meaning but may miss specific characteristics, whereas attribute representations provide structured information but may overlook narrative context. Furthermore, existing evaluations emphasize image quality and semantic similarity rather than verifying whether stated facial attributes are preserved. This work addresses these gaps through dual semantic-attribute conditioning and explicit attribute-level evaluation. Table 1 positions representative methods against our task requirements. No prior method simultaneously accepts a natural-language narrative, outputs a facial sketch, and provides explicit attribute-level control and evaluation for composite generation.

Table 1. Positioning against representative methods.

Method	Text in	Face gen	Sketch out	Attr. control	Composite
AttnGAN / DF-GAN	Yes	Yes	No	Weak	No
TediGAN / StyleCLIP	Yes	Yes	No	Latent	No
AttGAN / StarGAN	Attr.	Yes	No	Yes	No
APDrawingGAN	No	Photo-in	Yes	No	No
Stable Diffusion	Yes	Yes	Weak	Prompt	No
This work	Narrative	Yes	Yes	Explicit	Yes

3. Dataset and Narrative Construction

3.1 Dataset and Attribute Taxonomy

Experiments use Multi-Modal-CelebA-HQ (MM-CelebA-HQ), comprising 30,000 high-resolution face images, each aligned with natural-language captions and the 40 binary CelebA facial-attribute annotations. For analysis we group the 40 attributes into three families whose difficulty we will later show to differ systematically: structural / demographic attributes (e.g. Male, Young, Chubby, Oval_Face) with a strong global signature; appearance attributes (hair colour/style, facial hair, expression, makeup) that are moderately local; and accessory / rare attributes (eyeglasses, hat, necklace, earrings) that are spatially small and infrequent. Table 2 summarises the statistics.

Table 2. Dataset statistics.

Property	Value
Images	30,000
Captions per image	~10 narratives
Attributes	40 (CelebA)
Working resolution	128 x 128
Sketch targets	synthesised from images
Train / Val / Test	27,000 / 1,500 / 1,500

3.2 Data Preprocessing

Images are resized to 128 x 128, converted to single-channel sketch targets with a line-drawing operator, and normalised to the range [-1, 1] to match the generator's hyperbolic-tangent output. Because MM-CelebA-HQ does not distribute sketches, the synthesised sketches define the output domain and serve as the real samples for the discriminator, while attributes are encoded as a 40-dimensional multi-hot vector.

3.3 Narrative Construction

Each narrative is the image's MM-CelebA-HQ caption, an attribute-rich sentence such as 'This young woman has wavy hair and high cheekbones and is wearing lipstick.' A deterministic parser maps the narrative onto A(N): gender is inferred from pronoun and noun statistics, and the remaining attributes from curated lexical patterns covering hair colour and type, facial hair, eyeglasses, expression, and accessories. Because the captions are themselves attribute-conditioned, the parser is highly reliable; the distinction between such constructed narratives and spontaneous eyewitness speech is examined explicitly in Section 9.

3.4 Train / Validation / Test Split

The corpus is partitioned into 27,000 training, 1,500 validation and 1,500 test images with a fixed random seed; because CelebA-HQ images correspond to distinct identities, the partitioning avoids trivial identity leakage across sets. Reported results use 1,000 held-out test narratives.

4. Proposed Methodology

The proposed framework generates composite facial sketches from natural-language facial narratives by jointly modeling the semantic content of the description and explicitly identified facial attributes. The framework consists of five main stages: facial-sketch and narrative preparation, semantic representation using a frozen CLIP encoder, extraction of a 40-dimensional facial-attribute representation, joint semantic-attribute conditioning, and conditional adversarial sketch generation. A condition-aware discriminator is employed to distinguish realistic sketches from generated sketches while also learning the compatibility between a sketch and its associated narrative condition.

4.1 Overview

The framework (Figure 1) comprises a frozen text encoder, a deterministic attribute parser, a conditionally-modulated generator, and a matching-aware discriminator, trained under a composite objective balancing adversarial realism, discriminator regularisation, structural reconstruction and cross-modal semantic alignment.

4.2 Narrative and Attribute Encoding

The narrative is tokenised and mapped by a frozen CLIP ViT-B/32 encoder to a 512-dimensional embedding $e = E(N)$. Freezing E preserves a stable, transferable semantic space, reduces the number of trainable parameters, and allows the identical encoder to serve as an impartial evaluator of semantic consistency at test time. In parallel the parser yields the attribute vector

$$A(N) = [a_1, a_2, ..., a_{40}], \quad a_i \in \{0, 1\}, \qquad (1)$$

each dimension a named CelebA attribute. The conditioning code c = [e ; A] is projected to a 256-dimensional vector shared by generator and discriminator.

4.3 Generator with Deep Feature Modulation

The generator maps z to a 4 x 4 tensor via a learned projection and upsamples it to 128 x 128 through a stack of residual blocks. Within each block l the condition modulates the features by a learned affine transformation before convolution,

$h_l' = \gamma_l(c) \odot h_l + \beta_l(c),$ (2)

where $\gamma_l(\cdot)$ and $\beta_l(\cdot)$ are small networks producing per-channel scale and shift and $\odot$ denotes the Hadamard product. Applying (2) at every resolution — deep text-image fusion — injects both global semantics and explicit attributes throughout synthesis rather than only at the input, which is essential for controllable, attribute-consistent generation.

4.4 Matching-Aware Discriminator

The discriminator D downsamples the sketch to a 4 x 4 feature map, concatenates the spatially broadcast conditioning code, and outputs a single scalar D(S, c) scoring realism and narrative matching jointly. During training D additionally receives real sketches paired with a mismatched condition c^- as negatives, so a sketch is penalised both for being unrealistic and for being inconsistent with its description.

4.5 Training Objective

With S a real sketch, $\hat{S} = G(z, c)$ a generated one, and c^- a mismatched condition, the discriminator minimises the hinge loss

$L_D = E[\max(0, 1 - D(S, c))] + \tfrac{1}{2} E[\max(0, 1 + D(\hat{S}, c))]$ (3)
$+ \tfrac{1}{2} E[\max(0, 1 + D(S, c^-))],$

and the generator's adversarial term is

$L_{adv} = - E[D(\hat{S}, c)].$ (4)

A matching-aware gradient penalty on the real-matching pair, applied at every discriminator update, stabilises the game:

$L_{MAGP} = E[\, \| \nabla_{(S,c)} D(S, c) \|_2^2 \,].$ (5)

An L1 term anchors coarse structure and a CLIP term aligns semantics:

$L_{L1} = E \| \hat{S} - S \|_1, \quad L_{CLIP} = E[\, 1 - \cos(\Phi_{img}(\hat{S}), \Phi_{txt}(N)) \,].$ (6)

The full generator objective is

$L_{total} = L_{adv} + \lambda_{MAGP} L_{MAGP} + \lambda_{L1} L_{L1} + \lambda_{CLIP} L_{CLIP},$ (7)

with $\lambda_{MAGP} = 2.0$, $\lambda_{L1} = 0.1$ and $\lambda_{CLIP} = 1.0$.

4.6 Optimisation and Stabilisation Analysis

We optimise with Adam under a two time-scale update rule (learning rates 1e-4 for G and 4e-4 for D, $\beta = (0.0, 0.9)$), batch size 32, at 128 x 128 for 30 epochs on a single NVIDIA A2 GPU, checkpointing every epoch with exact resume. Two stabilisation findings are of theoretical interest and are reported as controlled observations (a formal ablation is future work, Section 7.4). First, the frequency of the gradient penalty is critical: applying MA-GP only intermittently under-regularises the discriminator, allowing its effective Lipschitz constant to grow so that its gradients to the generator vanish, driving G to the trivial minimiser (a blank sketch); applying (7) at every step restores bounded gradients and a healthy equilibrium. Second, an over-weighted L1 term interacts pathologically with the predominantly white sketch background: the fastest descent direction for the L1 term is the constant white image, a degenerate mode that adversarial and semantic pressure cannot escape until λ_{L1} is reduced to 0.1. Gradient clipping suppresses early-training penalty transients. These observations identify the discriminator's regularisation schedule and the reconstruction weight as the two decisive stability controls for sketch-domain conditional GANs.

5. Experimental Setup

5.1 Implementation Details

Table 3. Implementation details.

Setting	Value
Framework	PyTorch
Image size	128 x 128 (grayscale)
Batch	32
Optimiser	Adam, TTUR (1e-4 / 4e-4)
Epochs	30
GPU	NVIDIA A2

Test evaluated	1,000 narratives
Text encoder	CLIP ViT-B/32 (frozen)

6. Evaluation Metrics

6.1 Attribute-Level Evaluation

An independent CNN attribute predictor f, trained on real sketches with the parsed labels, is applied to each generated sketch; its multi-label predictions are compared to the ground-truth attributes. For attribute k with true/false positives and negatives (TP, FP, FN),

$P_k = TP_k / (TP_k + FP_k), \quad R_k = TP_k / (TP_k + FN_k), \quad (8)$

$F1_k = 2 P_k R_k / (P_k + R_k),$

aggregated as macro-F1 = $(1/K) \Sigma_k F1_k$, weighted-F1 = $\Sigma_k (n_k/N) F1_k$ with support n_k, and micro-F1 from pooled (TP, FP, FN). Macro-F1 weights every attribute equally and is therefore sensitive to rare-attribute failure, whereas weighted-F1 reflects frequency-weighted performance; reporting both exposes the imbalance in attribute difficulty.

6.2 Cross-Modal Semantic Consistency

CLIP consistency measures narrative alignment as the mean cosine similarity between the CLIP image embedding of the generated sketch and the CLIP text embedding of its narrative,

$C = (1/M) \Sigma_i \cos(\Phi_img(\hat{S_i}), \Phi_txt(N_i)). \quad (9)$

7. Results

7.1 Overall Quantitative Results

The results of 1,000 held-out narratives are presented in Table 4, and an example of the generated sketch is provided in Figure 2. The model generates meaningful composite sketches with global semantics that match the narrative with CLIP consistency 0.248 and macro-F1 0.382. The difference between macro-F1 (0.382) and weighted-F1 (0.520)

Table 4. Overall quantitative results (proposed model).

Metric	Value
Attribute macro-F1	0.382
Attribute weighted-F1	0.520
CLIP consistency	0.248
LPIPS	0.363
SSIM (suppl.)	0.268
PSNR (suppl.)	15.34 dB



Figure 2. Composite sketches generated from held-out narratives.

Table 5. Per-attribute Precision / Recall / F1 (ordered by F1).

Attribute	P	R	F1	#pos
No_Beard	0.94	0.93	0.93	857
Young	0.88	0.97	0.92	796
Male	0.68	1.00	0.81	374
Bags_Under_Eyes	0.55	0.87	0.67	294
Wearing_Lipstick	0.98	0.49	0.65	563

Sideburns	0.69	0.61	0.65	88
Bushy_Eyebrows	0.84	0.49	0.62	201
Big_Nose	0.58	0.65	0.61	321
Arched_Eyebrows	0.79	0.46	0.58	360
Attractive	0.86	0.43	0.57	586
Mouth_Slightly_Open	0.97	0.40	0.57	474
Smiling	0.99	0.39	0.56	465
Goatee	0.57	0.55	0.56	73
High_Cheekbones	0.96	0.38	0.55	451
Wearing_Earrings	0.73	0.41	0.53	253
Chubby	0.44	0.62	0.52	64
Black_Hair	0.34	0.95	0.50	241
Brown_Hair	0.41	0.53	0.46	244
Heavy_Makeup	0.99	0.29	0.45	472
Receding_Hairline	0.32	0.57	0.41	93
Wavy_Hair	0.96	0.23	0.37	365
Gray_Hair	0.53	0.26	0.35	35
Eyeglasses	1.00	0.19	0.31	43
Double_Chin	0.50	0.19	0.28	62
Big_Lips	0.69	0.12	0.21	373
Straight_Hair	0.65	0.10	0.17	209
Bald	0.67	0.08	0.14	25
Wearing_Necktie	0.60	0.04	0.07	81
Narrow_Eyes	0.75	0.02	0.05	121
Pointy_Nose	0.62	0.01	0.03	335
Rosy_Cheeks	1.00	0.01	0.02	119
Bangs	**0.73**	**0.73**	**0.73**	**0.73**
Oval_Face	0.82	0.82	0.82	0.82
Wearing_Necklace	**0.68**	**0.68**	**0.68**	**0.68**
Wearing_Hat	0.78	0.78	0.78	0.78
Blond_Hair	**0.80**	**0.74**	**0.74**	**0.78**
Pale_Skin	0.81	0.81	0.81	0.81

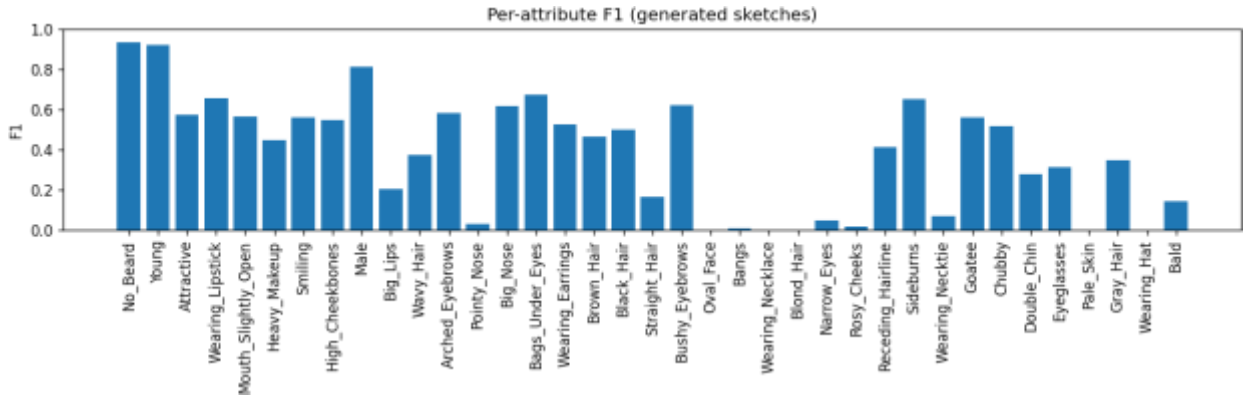


Figure 3. Per-attribute F1-score on generated sketches.

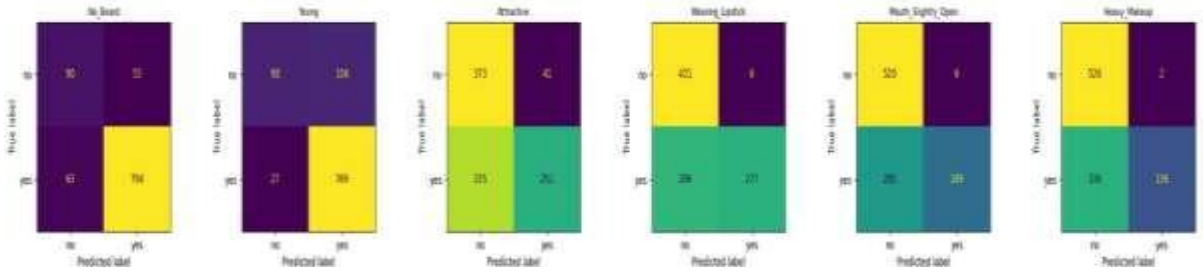


Figure 4. Confusion matrices.

8. Discussion

The results show that the proposed framework is capable of creating composite sketches of faces from natural language descriptions through the integration of global semantic information along with explicit facial attributes. Through the implementation of the dual-conditioning method, the model becomes able to take into account both the context of the witness description and the facial attributes, such as the person's age, gender, facial hair, hairstyle and hair properties, face shape, expression, and accessories. The better performance of attributes like No Beard, Young, and Male proves that the prominent features of the person can be effectively captured in the generated composite sketches, while the results of semantic and perceptual evaluation show that the generated sketches are semantically relevant to their natural language descriptions.

The importance of the proposed system can be found in its capability to transform an unstructured natural language description into a visual sketch, which can then be used for identification purposes. The necessity of manual transformation of a witness description into a sketch is bypassed by the proposed method since it provides an automatic process for the generation of a sketch. The sketches generated by the proposed framework can then be used for the identification process and

9. Conclusion

This experiment has proven that it is feasible to generate composite sketches from narrative descriptions of faces through a dual semantic and attribute conditioning method. In this process, the approach utilized not only the semantic information in the narrative description but also specific facial attributes to create a facial sketch of the person in question. Through this approach, there has been a computational connection established between the information that has been narrated by an eyewitness and the process of creating a facial sketch. Facial composites obtained through this process can be compared with the photographic database of the person to identify the individual or at least get a possible suspect. Thus, this approach provides a way of translating unstructured facial descriptions into machine-generated facial composites..

10. References

[1] D.-P. Fan, Z. Huang, P. Zheng, H. Liu, X. Qin, and L. Van Gool, “Facial-sketch Synthesis: A New Challenge,” Machine Intelligence Research, vol. 19, pp. 257–287, 2022, doi: 10.1007/s11633-022-1349-9.

[2] K. Zhong, Z. Chen, C. Liu, Q. M. J. Wu, and S. Duan, “Unsupervised Self-Attention Lightweight Photo-to-Sketch Synthesis with Feature Maps,” Journal of Visual Communication and Image Representation, vol. 90, 103747, 2023, doi: 10.1016/j.jvcir.2022.103747.

[3] L. Yan, W. Zheng, C. Gou, and F.-Y. Wang, “IsGAN: Identity-Sensitive Generative Adversarial Network for Face Photo-Sketch Synthesis,” Pattern Recognition, vol. 119, 108077, 2021, doi: 10.1016/j.patcog.2021.108077.

[4] C. Chen, W. Liu, X. Tan, and K.-Y. K. Wong, "Semi-Supervised Cycle-GAN for Face Photo-Sketch Translation in the Wild," Computer Vision and Image Understanding, vol. 235, 103775, 2023, doi: 10.1016/j.cviu.2023.103775.
[5] J. Shi and L. Wan, "Transformer-Based Adversarial Network for Semi-Supervised Face Sketch Synthesis," 2024.
[6] Y. Chen, Y. Wang, Y. Zhang, R. Feng, T. Zhang, X. Lu, and S. Gao, "Fine-Granularity Face Sketch Synthesis," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4035–4039, 2024, doi: 10.1109/ICASSP48485.2024.10447315.
[7] K. Du, J. Zhang, L. Cao, Y. Guo, and W. Sun, "A Facial Structure Sampling Contrastive Learning Method for Sketch Facial Synthesis," Scientific Reports, vol. 15, 16056, 2025, doi: 10.1038/s41598-025-00574-6.
[8] D. Lu, Z. Chen, C. Liu, Y. Hu, and Q. M. J. Wu, "Facial Sketch Synthesis with Multi-Level Guided Latent Diffusion Model," Neurocomputing, 2026.
[9] J. E. Martis, S. M. Shetty, M. R. Pradhan, U. Desai, and B. Acharya, "Text-to-Sketch Synthesis via Adversarial Network," Computers, Materials & Continua, vol. 76, no. 1, pp. 915–938, 2023, doi: 10.32604/cmc.2023.038847.
[10] L. Weng, C. Li, and H. Xu, "Stepwise Synthetic Face Sketch Generation Algorithm Based on GAN," Journal of Computer-Aided Design & Computer Graphics, vol. 35, no. 9, pp. 1363–1373, 2023.
[11] A. Ke, Y. Huang, J. Yang, and B. Cai, "Text-Guided Image-to-Sketch Diffusion Models," Knowledge-Based Systems, vol. 304, 112441, 2024, doi: 10.1016/j.knosys.2024.112441.
[12] N. Sa´daba-Campo and H. Go´mez-Moreno, "Exploration of Generative Neural Networks for Police Facial Sketches," Big Data and Cognitive Computing, vol. 9, no. 2, 42, 2025, doi: 10.3390/bdcc9020042.
[13] "Embedded Method for Improving Text-Face Relevance," Pattern Recognition / Related text-to-face synthesis literature, 2022.
[14] K. Wang et al., "Text-to-Face Synthesis Based on Facial Landmarks Prediction," Machine Vision and Applications, 2025.
[15] Q. Guo and X. Gu, "An Improved StyleGAN-Based TextToFace Model with Local-Global Information Fusion," Expert Systems with Applications, vol. 250, 123698, 2024, doi: 10.1016/j.eswa.2024.123698.
[16] W. Song, Z. Ye, M. Sun, X. Hou, S. Li, and A. Hao, "AttriDiffuser: Adversarially Enhanced Diffusion Model for Text-to-Facial Attribute Image Synthesis," Pattern Recognition, vol. 163, 111447, 2025, doi: 10.1016/j.patcog.2025.111447.
[17] J. Dong, P. Xu, Q. Zhang, H. Fang, G. Schaefer, R. Liu, P. Yi, and X. Fang, "Fine-Grained Face Personalisation Using a Text-Guided Multi-Attribute Embedded Diffusion Model," Expert Systems with Applications, 129015, 2026, doi: 10.1016/j.eswa.2025.129015.
[18] X. Wang et al., "A Landmarks-Assisted Diffusion Model with Heatmap-Guided Denoising Loss for High-Fidelity and Controllable Facial Image Generation," Image and Vision Computing, vol. 159, 105545, 2025, doi: 10.1016/j.imavis.2025.105545.
[19] M. Aishwarya and M. R. Pooja, "Advancements in Face Sketch Recognition: Techniques and Challenges," in Proc. 2024 IEEE Conf. Eng. Informatics (ICEI), Melbourne, Australia, Nov. 2024.
[20] W. Wan, Y. Yang, and H. J. Lee, "Generative adversarial learning for detail-preserving face sketch synthesis," Neurocomputing, vol. 438, pp. 107–121, 2021.
[21] C. Navuluri, S. Jukanti, and R. R. Allapuram, "Semantic neural model approach for face recognition from sketch," arXiv preprint arXiv:2305.01058, 2023.
[22] S. Bae, N. U. Din, H. Park, and J. Yi, "Face photo-sketch recognition using bidirectional collaborative synthesis network," in Proc. 16th Int. Conf. Ubiquitous Inf. Manage. Commun. (IMCOM), 2022, pp. 1–8.
[23] M. Guo, M. Xiong, J. Huang, X. Hu, and T. Peng, "Face photo-sketch portraits transformation via generation pipeline," The Visual Computer, vol. 41, no. 2, pp. 1183–1196, 2025.
[24] K. Du, H. Zhou, L. Cao, Y. Guo, and T. Wang, "MHGAN: Multi-Hierarchies Generative Adversarial Network for high-quality face sketch synthesis," IEEE Access, vol. 8, pp. 212995–213011, 2020, doi: 10.1109/ACCESS.2020.3041284.
[25] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 11, pp. 1955–1967, 2009, doi: 10.1109/TPAMI.2008.222.