Supplementary Material for

CaricHarmony: Contrastive Diffusion Paths for Identity-Preserving Caricature Synthesis

A. Algorithm

The steps of our proposed method are summarized in Algorithm 1 for clarity. We use DDIM sampler as an example for simplicity, but it is straightforward to replace it with the DPM++ 2M sampler [5]. $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ is the cumulative product of the noise schedule [3]. It is worth mentioning that, at each time-step, the energy functions should be computed across all the cross-attention blocks and subsequently summed.

Algorithm 1 CaricHarmony

Require: UNET denoiser $\epsilon_\theta(\cdot)$; CLIP-T model $\mathcal{T}(\cdot)$; Word embedding layer $\mathcal{W}(\cdot)$; Text prompts p; Binary sketch image S; Identity image I; Classifier-free guidance scale γ; Guidance rate η; VAE decoder $D(\cdot)$.
Initialization:
Sample latent noise $\hat{z}_T = \hat{z}_T^{\mathrm{i}} = \hat{z}_T^{\mathrm{s}} \sim \mathcal{N}(0, \mathbf{I})$.
Obtain text conditions $C_{\mathrm{txt}} \leftarrow \mathcal{T}(\mathcal{W}(p))$.
Obtain ID conditions $C_{\mathrm{id}} \leftarrow PuLID(I)$.
Obtain shape conditions $C_{\mathrm{s}} \leftarrow Adapter(S)$.
Set joint condition $C_{\mathrm{e}} \leftarrow [C_{\mathrm{txt}}, C_{\mathrm{id}}, C_{\mathrm{s}}]$.
for $t = T$ **to** 0 **do**
  $\mathcal{E}_{\mathrm{b}} \leftarrow 0$.
  $\hat{\epsilon}_t \leftarrow (1+\gamma)\,\epsilon_\theta(\hat{z}_t, t, C_{\mathrm{e}}) - \gamma\,\epsilon_\theta(\hat{z}_t, t, \emptyset)$ {Classifier-Free Guidance}
  Cache intermediate features Q, K_{txt}, V, and O derived from $\epsilon_\theta(\hat{z}_t, t, C_{\mathrm{e}})$.
  if $t^{\mathrm{s}}_{\mathrm{start}} > t > t^{\mathrm{s}}_{\mathrm{end}}$ **then**
    $\hat{z}^{\mathrm{s}}_{t-1}, Q^{\mathrm{s}} \leftarrow DDIMSamplingStep(\hat{z}^{\mathrm{s}}_t, C_{\mathrm{s}}, C_{\mathrm{txt}}, \epsilon_\theta(\cdot), \gamma)$ {A sampling step with Classifier-Free Guidance}
    $\mathcal{E}_{\mathrm{layout}}, \mathcal{E}_{\mathrm{sem}} \leftarrow ComputeShapeEnergy(Q, Q^{\mathrm{s}}, K_{\mathrm{txt}})$ {Equation (3) and Equation (4)}
    $\mathcal{E}_{\mathrm{b}} \leftarrow \mathcal{E}_{\mathrm{b}} + \mathcal{E}_{\mathrm{layout}} + \mathcal{E}_{\mathrm{sem}}$
  end if
  if $t^{\mathrm{i}}_{\mathrm{start}} > t > t^{\mathrm{i}}_{\mathrm{end}}$ **then**
    $\hat{z}^{\mathrm{i}}_{t-1}, [Q^{\mathrm{i}}, O^{\mathrm{i}}] \leftarrow DDIMSamplingStep(\hat{z}^{\mathrm{i}}_t, C_{\mathrm{id}}, C_{\mathrm{txt}}, \epsilon_\theta(\cdot), \gamma)$
    $\mathcal{E}_{\mathrm{id}} \leftarrow ComputeIDEnergy(Q, O, Q^{\mathrm{i}}, O^{\mathrm{i}})$ {Equation (7)}
    $\mathcal{E}_{\mathrm{b}} \leftarrow \mathcal{E}_{\mathrm{b}} + \mathcal{E}_{\mathrm{id}}$
  end if
  if $\mathcal{E}_{\mathrm{b}} \neq 0$ **then**
    $\tilde{\epsilon}_t \leftarrow \hat{\epsilon}_t + \eta \nabla_{\hat{z}_t} \mathcal{E}_{\mathrm{b}}$
  else
    $\tilde{\epsilon}_t \leftarrow \hat{\epsilon}_t$
  end if
  $\hat{z}_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}} \frac{\hat{z}_t - \sqrt{1-\bar{\alpha}_t}\,\tilde{\epsilon}_t}{\sqrt{\bar{\alpha}_t}} + \sqrt{1-\bar{\alpha}_{t-1}}\,\tilde{\epsilon}_t$ {DDIM Sampling}
end for
$\hat{x}_0 \leftarrow D(\hat{z}_0)$
return $\hat{x}_0$

B. Results in Various Levels of Exaggerations

Figure 1 presents additional results to demonstrate the robustness and compatibility of our model when handling a diverse range of shape conditions. The results are organized into three groups from top to bottom. In each group, the left side takes a celebrity photo as the reference identity image, while the right side takes a synthetically generated face. Even when the sketches are highly exaggerated, the model successfully preserves the intended creative semantics while maintaining strong ID fidelity, highlighting its effectiveness in balancing shape and ID conditions.

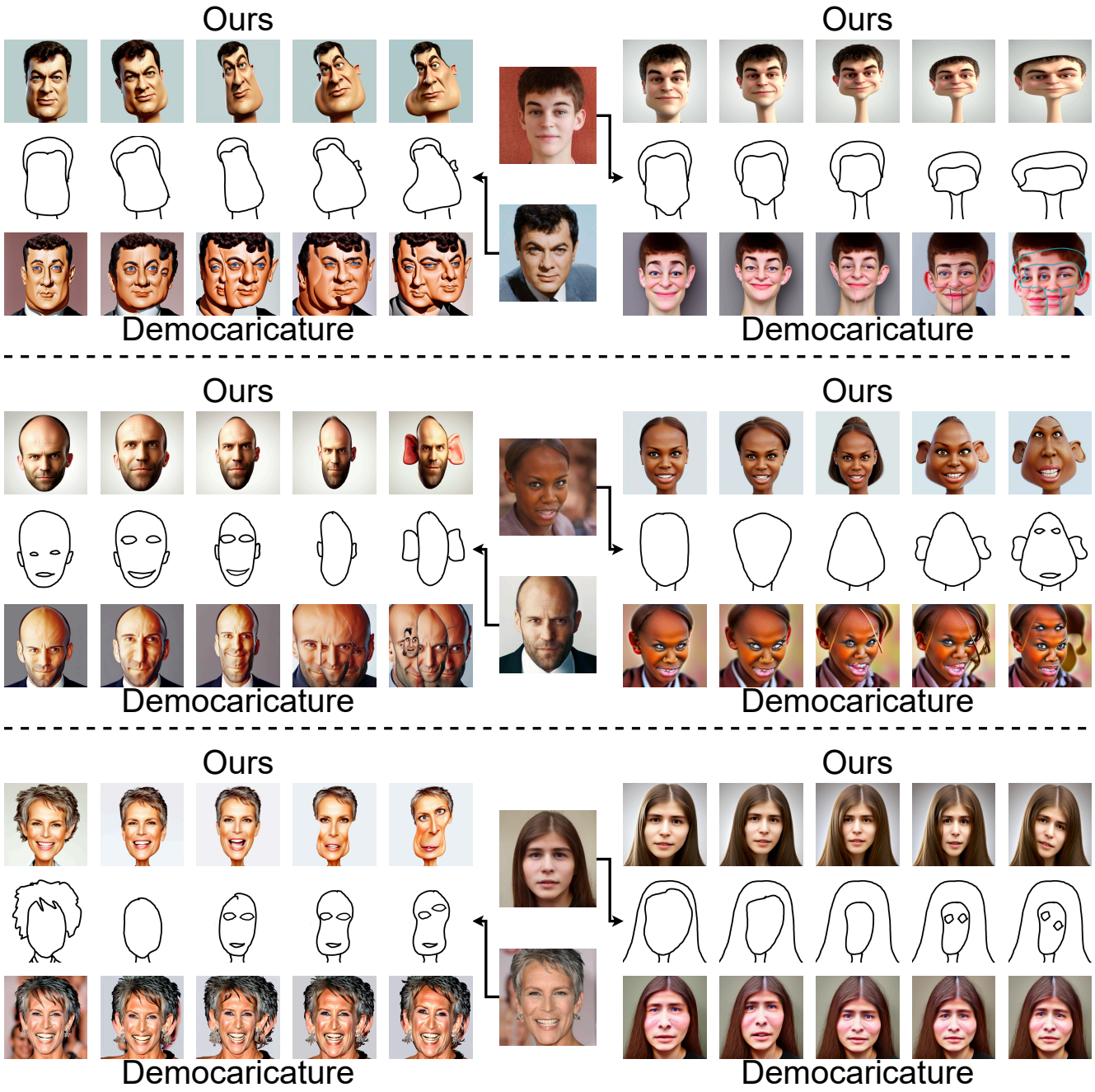

Figure 1. Additional qualitative performance comparison on sketches with varying levels of exaggeration.

C. Additional Results

We present additional results with different combinations of ID and shape conditions and build a comprehensive result matrix as shown in Figure 2. Each cell in the matrix represents a unique pairing of a specific identity and distinct shape conditions. By observing rows and columns where either ID or shape conditions remain fixed while the other varies, it can be seen that the model can preserve ID consistency across shape changes and maintain shape characteristics across different identities. The results highlight the strong robustness of our model in disentangling and recombining ID and shape conditions. Thanks to the robust generalization capability of the PuLID-ID-encoder [2], the model still retains such ability for unknown faces.

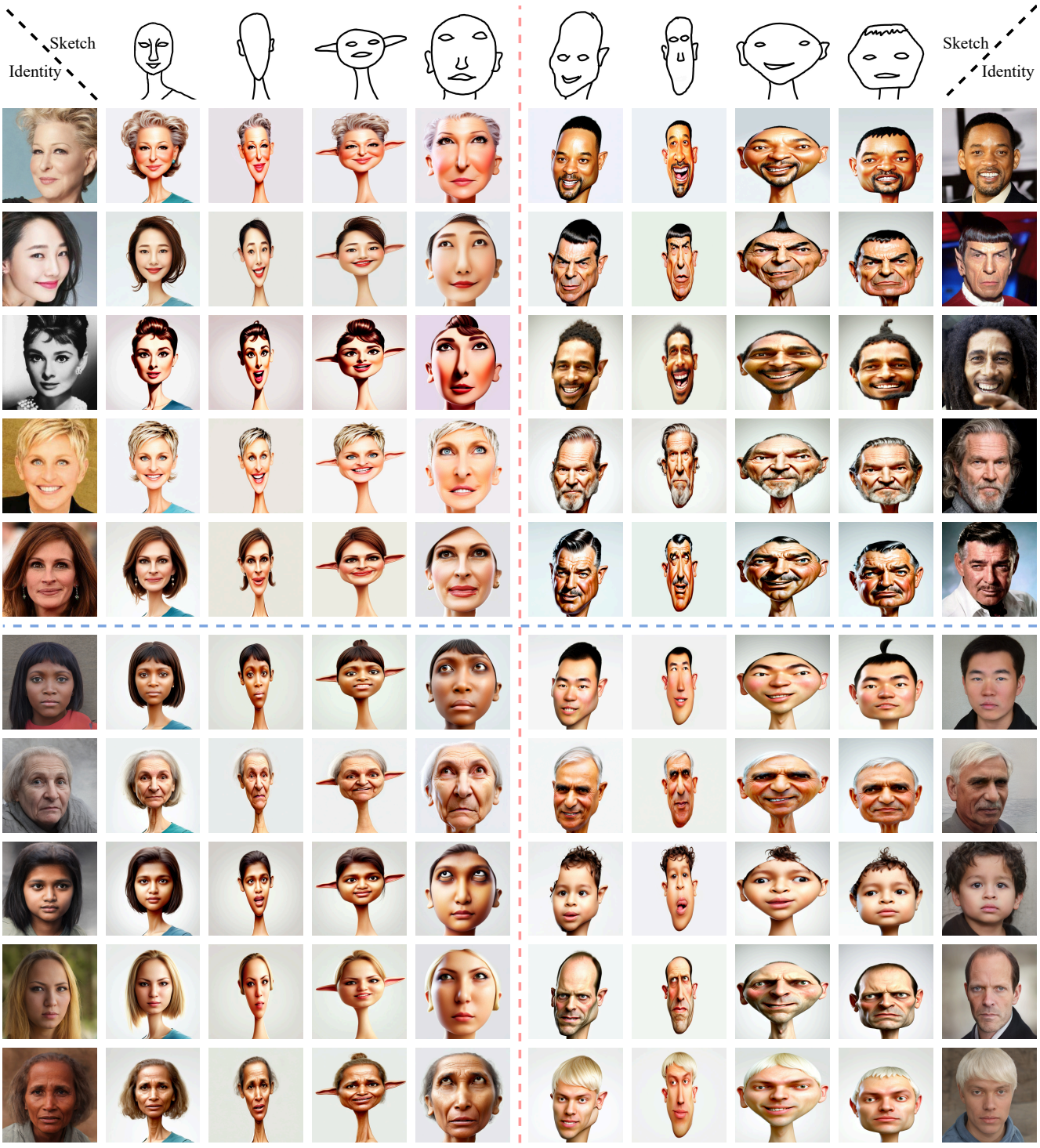

Figure 2. Additional results of our model with various identity images and shape conditions. Left of the pink line: female faces. Right of the pink line: male faces. Above the blue line: celebrity faces. Below the blue line: unknown faces.

D. Failure Cases

We discuss the failure cases of our model as shown in Figure 3. One issue arises when the features presented in the sketch are significantly different from those in the reference identity: the model will not ignore but signify them in the generated

results, causing degradation of ID fidelity. For example, in the first row of Figure 3, the sketch depicts a character with pigtails, whereas the identity has short golden hair. To handle the contradiction, the model tends to overly follow the shape conditions (i.e., the pigtails), while also incorporating irrelevant features (such as hair color) from $\mathcal{P}^s$ into the output. This leads to an attribute shift (a change of hair color from golden to white) that harms ID fidelity. Another issue stems from our model's dependency on the T2I-Sketch-Adapter [6]. Suppose the sketch is highly ambiguous and cannot be interpreted by the original T2I-Sketch-Adapter correctly. In that case, our model will follow inappropriate guidance from $\mathcal{P}^s$ and produce undesired results. In the second row of Figure 3, we expect the model to generate a caricature with a square-shaped face. However, the T2I-Sketch-Adapter misinterprets the shape conditions as a backdrop, leading to failure to produce the desired result.

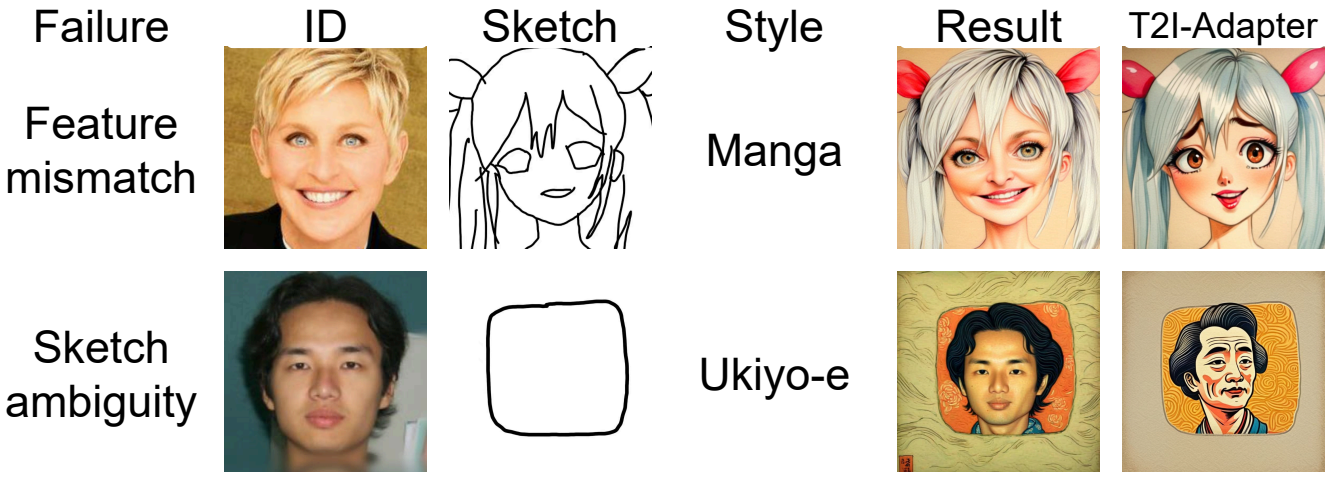

Figure 3. Failure cases of our model. Results under the column T2I-Adapter are those generated from $\mathcal{P}^s$.

E. Additional Comparison against Modern Models

We further compare our model with CaricatureBooth [8] and QWEN-Image-Edit-Max[1] [9] in Figure 4. CaricatureBooth is the latest caricature synthesis model based on SDXL [7]. QWEN-Image-Edit-Max is a 20B multi-modal model for high-fidelity image editing that supports multi-image input, allowing us to provide both the identity image and sketch condition simultaneously. We annotate hand-drawn sketches as Bezier curves representing the chin, lips, eyes, and nose to make them compatible with CaricatureBooth. However, CaricatureBooth cannot sufficiently capture sketch shapes due to its restrictive input format with a fixed number of control points, resulting in poor representation of additional facial features and highly curved regions. QWEN-Image-Edit-Max, on the other hand, tends to overly follow the sketch shape while neglecting the subject's identity. Our model can effectively harmonize both conditions via the introduced balancing mechanism while supporting flexible sketch formats. We use the same data used in the user study for quantitative evaluation. As shown in Table 1, our model achieves the best ID and shape consistency score, as well as the best aesthetic score according to ImageReward.

Methods	I-CLIP ↑	S-CLIP ↑	PickScore ↑	ImageReward ↑
Qwen-Image-Edit-Max	0.5487	0.8374	0.1274	-0.3983
CaricatureBooth	0.5615	0.7637	**0.5141**	0.4827
Ours	**0.6472**	**0.8500**	0.3584	**0.8493**

Table 1. Quantitative comparison with concurrent models.

[1] https://www.alibabacloud.com/help/en/model-studio/qwen-image-edit-api

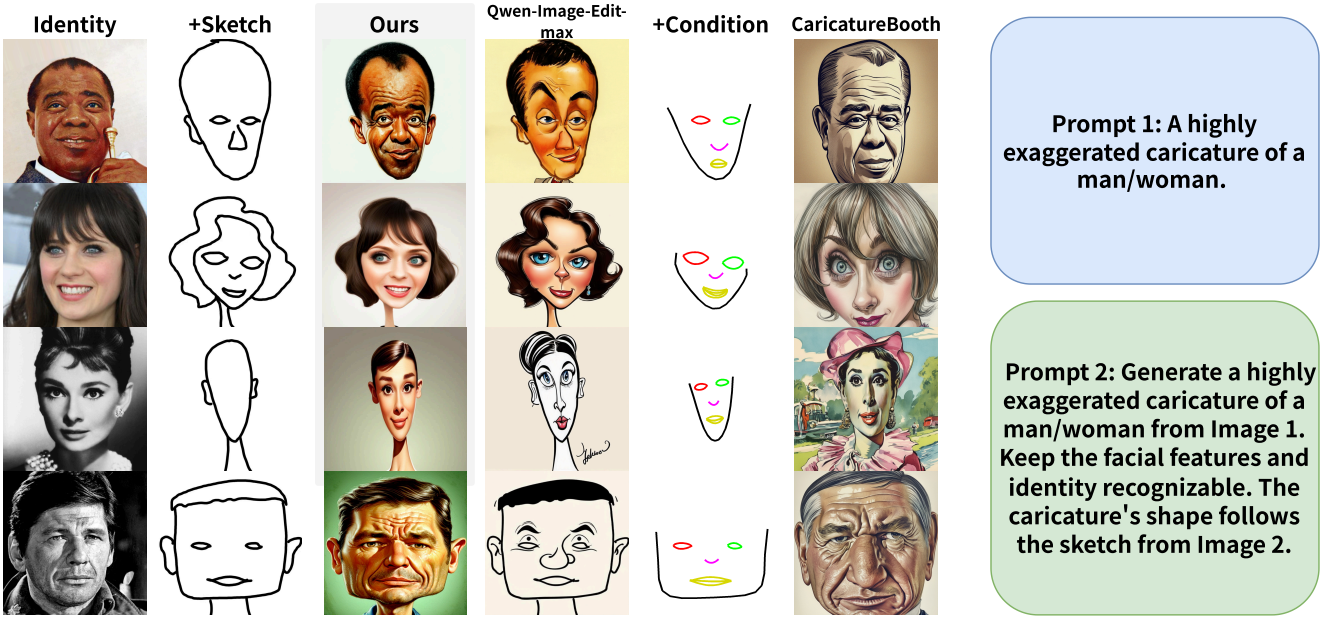

Figure 4. Performance comparison against modern generative models. We use Prompt 1 for our model and CaricatureBooth, and Prompt 2 for Qwen-Image-Edit-Max to fit the image editing task.

Methods	I-CLIP ↑	ArcFace-sim ↑
DemoCaricature	**0.8182**	0.5405
CaricatureBooth	0.6450	0.4989
Ours	0.8121	**0.6929**

Table 2. Results of face reconstruction.

F. Face Reconstruction Capability

To evaluate the identity preserving capability of our model, we perform face reconstruction for all identities in WebCaricature [4], using the lowest-indexed photo as the ID image and extracting its edge map or landmarks as shape conditions. We use the cosine similarity of ArcFace [1] features as an additional metric. As shown in Table 2, our method achieves a significantly higher ArcFace similarity and a competitive I-CLIP score compared to modern caricature synthesis models, demonstrating that our method can effectively preserve facial features of identities.

References

[1] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 5

[2] Zinan Guo, Yanze Wu, Chen Zhuowei, Peng Zhang, Qian He, et al. Pulid: Pure and lightning id customization via contrastive alignment. In *NeurIPS*, 2024. 2

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 1

[4] Jing Huo, Wenbin Li, Yinghuan Shi, Yang Gao, and Hujun Yin. Webcaricature: a benchmark for caricature recognition. In *BMVC*, 2018. 5

[5] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 2025. 1

[6] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, 2024. 4

[7] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 4

[8] Zhiyu Qu, Yunqi Miao, Zhensong Zhang, Jifei Song, Jiankang Deng, and Yi-Zhe Song. Caricaturebooth: Data-free interactive caricature generation in a photo booth. In *CVPR*, 2025. 4

[9] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 4