This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

CaricHarmony: Contrastive Diffusion Paths for Identity-Preserving Caricature Synthesis

Dongyu Wang Dar-Yen Chen Yi-Zhe Song
SketchX, CVSSP, University of Surrey
{dongyu.wang, d.chen, y.song}@surrey.ac.uk
https://dongyuuw.github.io/CaricHarmony/

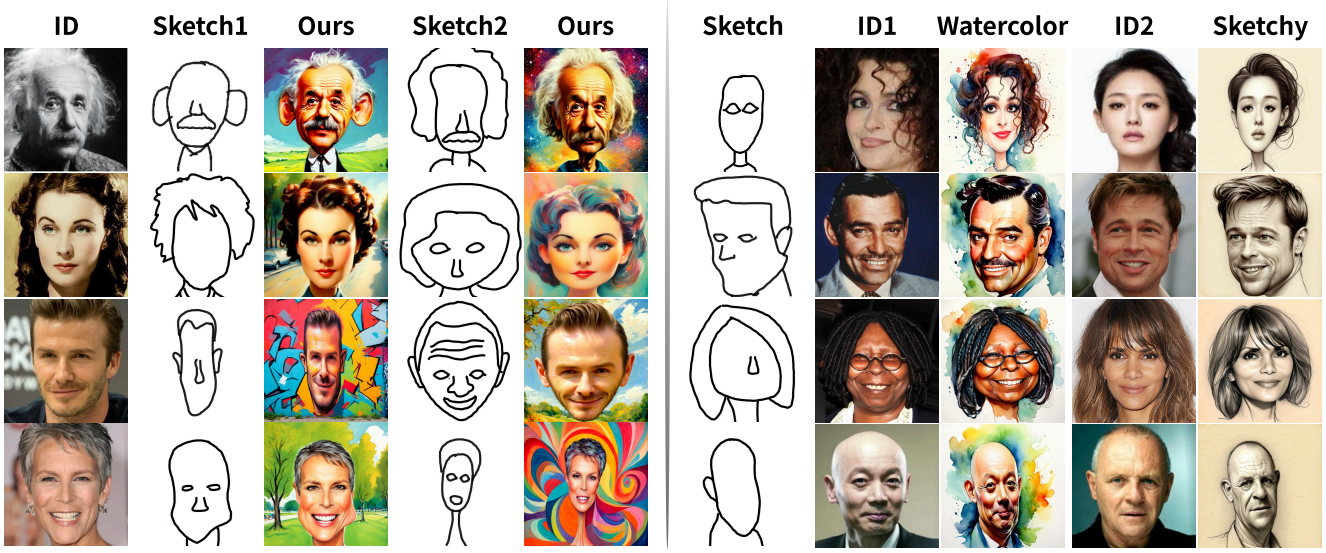


Figure 1. **High-quality caricatures generated by our CaricHarmony. Left:** Results generated with the same identity and different sketches. **Right:** Results generated with the same sketch and different identities and specified styles.

Abstract

Sketch-based caricature synthesis suffers from a fundamental failure mode: when identity and shape conditions are combined in diffusion models, they create destructive interference that causes inevitable collapse toward either bland portraits or unrecognizable distortions. We identify the root cause as condition signal contamination *– competing probability distributions in the denoising trajectory that make balanced generation impossible. We present CaricHarmony, the first training-free method that explicitly resolves this contamination through parallel uncontaminated diffusion paths. During inference, we maintain three paths: $\mathcal{P}^{\text{i}}$ (pure identity), $\mathcal{P}^{\text{s}}$ (pure shape), and $\mathcal{P}^{\text{i+s}}$ (harmonized output). Novel energy functions operating on cross-attention features provide gradient guidance that steers $\mathcal{P}^{\text{i+s}}$ toward optimal balance: $\mathcal{E}_{\text{shape}}$ ensures sketch fidelity through layout and semantic alignment, while $\mathcal{E}_{\text{id}}$ employs token-level correspondence matching robust to extreme distortions. Unlike DemoCaricature requiring 70 seconds per-identity fine-tuning or CaricatureBooth constrained to Bezier curves, CaricHarmony accepts any sketch format and generates in under 16 seconds. Experiments demonstrate state-of-the-art performance: 0.8615 shape CLIP score (vs. 0.8450) under comparable identity consistency score, with 7.81 overall user preference score (vs. 6.06). Our method fundamentally reconceptualizes the ID-shape conflict as conditioning signal contamination for diffusion models, enabling unprecedented creative control while preserving recognition.*

1. Introduction

Caricature synthesis demands a paradox: extreme shape distortions that somehow preserve identity. Artists instinctively navigate this tension, but computational methods sys-

tematically fail. When given a sketch with exaggerated features – a grotesquely elongated nose, impossibly tiny eyes – existing approaches either ignore these creative choices to preserve identity, or follow them faithfully at the cost of recognition. This fundamental *ID-shape conflict* isn't just an implementation detail; it's the core challenge that has prevented sketch-based caricature synthesis from achieving its artistic potential.

We identify the root cause: *condition signal contamination*. When identity conditions C_{id} and shape conditions C_{s} are combined in diffusion models, they create destructive interference in the denoising trajectory. The model oscillates between competing probability distributions – regions near pure identity preservation and pure shape following exhibit higher densities than the balanced intermediate region, causing inevitable collapse toward one extreme. Prior methods [3, 32] fundamentally ignore this problem, attempting means of balancing identity and shape without addressing the underlying contamination.

Table 1 reveals a critical gap in the field. Early synthesis methods [2, 11, 20, 41] lack controllability entirely, performing automatic distortions that ignore user intent. Recent controllable approaches face crippling limitations: DemoCaricature [3] requires expensive per-sample fine-tuning (70 seconds per identity), making it impractical for interactive creation. CaricatureBooth [32] constrains creativity to Bezier curves, requiring large-scale pre-training on synthetic data. Critically, *neither addresses the core signal contamination issue* – they simply shift where the ID-shape conflict manifests.

We present CaricHarmony, the first training-free solution that directly resolves condition contamination through *contrastive feature alignment*. Our key insight transforms the problem: instead of mixing contaminated signals, we maintain three parallel diffusion paths during inference – $\mathcal{P}^{\text{i}}$ (pure identity), $\mathcal{P}^{\text{s}}$ (pure shape), and $\mathcal{P}^{\text{i+s}}$ (harmonized output). The uncontaminated reference paths $\mathcal{P}^{\text{i}}$ and $\mathcal{P}^{\text{s}}$ serve as guidance anchors, while specialized energy functions dynamically steer $\mathcal{P}^{\text{i+s}}$ toward optimal balance through gradient guidance at each denoising step.

Our energy functions operate within cross-attention mechanisms to achieve precise control: $\mathcal{E}_{\text{shape}}$ ensures sketch fidelity through layout alignment ($\mathcal{E}_{\text{layout}}$) and semantic consistency ($\mathcal{E}_{\text{sem}}$), while $\mathcal{E}_{\text{id}}$ employs novel token-level correspondence matching that adapts to shape distortions. Unlike conventional identity losses that fail on caricatures due to domain gap, our approach leverages intermediate diffusion features that inherently preserve layout information for localizing facial features. The compositional energy $\mathcal{E}_{\text{b}} = \mathcal{E}_{\text{shape}} + \mathcal{E}_{\text{id}}$ provides continuous gradient guidance throughout denoising, preventing trajectory collapse.

CaricHarmony achieves breakthrough results: generating high-quality caricatures from extreme sketches in 16 seconds (4× faster than DemoCaricature), accepting *any sketch format* without preprocessing (unlike CaricatureBooth's Bezier constraints), and requiring *zero training or fine-tuning* while maintaining state-of-the-art quality. Extensive experiments demonstrate superior performance across CLIP-based shape fidelity metric [3, 33] (0.8615 vs. 0.8450) and aesthetic quality metrics (ImageReward [51] 0.8509 vs. 0.4340 and PickScore [24] 0.2049 vs. 0.1094) with competitive ID consistency. Furthermore, user studies reveal a strong preference for our results, achieving an overall preference score of 7.81 (vs. 6.06).

Our contributions are: (i) *First explicit solution to condition contamination*: We identify signal contamination as the root cause of ID-shape conflict and resolve it through parallel uncontaminated reference paths, fundamentally different from prior diffusion-based approaches. (ii) *Novel cross-attention energy functions*: We introduce specialized alignment mechanisms that operate on intermediate diffusion features, including token-level correspondence matching that adapts to arbitrary shape distortions – impossible with conventional identity losses. (iii) *Practical democratization of caricature creation*: Our training-free framework eliminates all bottlenecks – no pre-training, no per-sample optimization, no sketch format constraints – making professional caricature synthesis accessible to anyone with a rough sketch.

Methods	Training-free	Controllable generation	Free-form conditioning
StyleCariGAN [20]	✗	✗	✓
WarpGAN [41]	✗	✗	✓
AutoToon [11]	✗	✗	✓
DemoCaricature [3]	✗	✓	✓
CaricatureBooth [32]	✗	✓	✗
Ours	✓	✓	✓

Table 1. Our model features user-friendly characteristics beyond dealing with the ID-shape conflict issue. Free-form conditioning: no additional requirements on shape conditions (if applicable) and ID conditions.

2. Related Work

Deep Caricature Synthesis. Caricature synthesis aims at generating a stylized portrait of a human subject with artistic styles and shape exaggerations while preserving the identity [5, 12]. CariGANs [2] utilizes GAN-based networks to generate stylized caricatures under the guidance of deformed facial landmarks. WarpGAN [41] eases the warping process by using automatically generated control points. Several subsequent works include diversifying styles and exaggeration types [12], using SENet [18] to generate warping fields that densely guide the warping process [11], and introducing shape exaggeration blocks [20] based on StyleGAN [23] to achieve shape deformation by manipulating

coarse-scale feature maps. Semantics CariGAN [5] proposes using facial parsing maps to guide the shape transformation in the output. For works realizing the controllability of the shape exaggeration, DemoCaricature [3] first proposes using free-hand sketches as a flexible and expressive medium to guide the shape deformation process. It encapsulates the identity information into textual embeddings and the key-value pathway of cross-attention blocks through per-sample tuning, but lacks consideration of its potential conflict with the shape information, limiting its performance on exaggerated free-hand sketches. CaricatureBooth [32] represents a sketch with a fixed number of Bezier curves [39] and performs TPS deformation [8] to generate synthetic data for training. However, it severely constrains the format of sketches and limits the diversity and creativity of shape conditions.

In contrast to prior works, our approach explicitly addresses the fundamental ID-shape conflict through a novel training-free framework that balances identity preservation and shape exaggeration, overcoming the limitations observed in DemoCaricature. Furthermore, our method accepts free-hand sketches in their natural form, preserving the full expressive potential and creativity of sketch-based control while ensuring robustness to diverse and exaggerated shape conditions.

T2I Personalization. The goal of T2I personalization is to customize T2I models [1, 31, 34, 38] to consider concepts specified by a set of reference images while maintaining their original generalizability. Textual Inversion [10] finds new pseudo-word embeddings to represent the semantic features of referred subjects. DreamBooth [36] fine-tunes pretrained Stable Diffusion [34] and Imagen [38] models to link each specified subject with a unique representative textual identifier. For parameter-efficient fine-tuning, Custom Diffusion [25] only fine-tunes cross-attention layers, while Perfusion [48] presents a Rank-One Model Editing (ROME) method that optimizes only the value-pathway in cross-attention layers. For multi-subject optimization, LatexBlend [21] represents multiple concepts in a latent textual space, enabling multi-concept generation by retrieving concepts from a bank that can be freely combined without tuning. LoRACLR [42] merges multiple pretrained LoRAs into a unified model using contrastive learning to align multiple concepts while preventing interference. For fast personalization without per-sample tuning, IP-Adapter [52] trains lightweight and decoupled cross-attention blocks to embed image prompts into the model in parallel with textual prompts. InstantBooth [40] achieves a tuning-free method by learning a network that can map reference images into aligned textual embeddings during inference. PuLID [13] introduces a tuning-free approach that can output high-fidelity personalized images while mitigating the interruption to the behavior of the original model through contrastive alignment. Imagine yourself [15] proposes a parallel attention architecture with three text encoders and one vision encoder for high-fidelity identity preservation and text alignment. Flux already knows [22] eliminates the need for pre-training and frames subject-driven generation as grid-based image completion with mosaic layouts by utilizing the generalizability of Flux models [1].

3. Revisit Text-to-Image Diffusion Models

3.1. Diffusion Models

Diffusion models [17, 31, 34] synthesize images from Gaussian noise by iterating the denoising process. They consist of two key components: a forward diffusion process and a reverse denoising process. The forward process corrupts a data sample z_0 by iteratively adding Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ over T time-steps, producing noisy data z_T. The denoising process recovers z_0 from z_T by iteratively estimating and removing noise using a trained network $\epsilon_\theta(\hat{z}_t, t)$, which takes as input the predicted noisy latent $\hat{z}_t$ and its corresponding time-step t [17, 34]. Typically, $\epsilon_\theta(\cdot)$ follows a UNET [35] architecture and incorporates a series of residual blocks [14], self-attention blocks, and cross-attention blocks [49].

From the continuous perspective [47] of diffusion models, $\epsilon_\theta(\hat{z}_t, t)$ can be explained as a scaled version of the score function $\nabla_{\hat{z}_t} \log q(\hat{z}_t)$ that samples from the probability distribution $q(\hat{z}_t)$ [46] through Langevin dynamics [44, 45]. When the generative process is combined with external conditions C, the corresponding score function $\nabla_{\hat{z}_t} \log q(\hat{z}_t|C)$ can be decomposed into:

$$\nabla_{\hat{z}_t} \log \frac{q(\hat{z}_t)q(C|\hat{z}_t)}{q(C)} = \nabla_{\hat{z}_t} \log q(\hat{z}_t) + \nabla_{\hat{z}_t} \log q(C|\hat{z}_t), \quad (1)$$

where the first term $\nabla_{\hat{z}_t} \log q(\hat{z}_t)$ corresponds to an unconditional denoiser $\epsilon_\theta(\hat{z}_t, t)$, and the second term is the gradient guidance that can be generated by an energy function $-\mathcal{E}(\hat{z}_t; t, C) = \log q(C|\hat{z}_t)$.

By designing energy functions for different generation objectives, such as text-conditioned generation [7], sketch-guided generation [50], mask-guided generation [43], and image editing [9], the model can generate outputs that conform to specified conditions. Inspired by the success of prior works, we address the ID-shape conflict in caricature synthesis by designing tailored energy functions that explicitly guide noisy latent toward achieving an optimal trade-off.

3.2. IP-Adapter and PuLID

IP-Adapter [52] introduces an effective approach to integrate ID features, denoted as C_{id}, into the model through the cross-attention mechanism [34, 49] in parallel with textual features C_{txt}. There are two linear layers W_K^{id} and W_V^{id}

mapping C_{id} into key (K_{id}) and value (V_{id}) matrices. The cross attention is then calculated as follows:

$$\begin{cases} O = \text{Attn}(Q, K_{\text{id}}, V_{\text{id}}) = \text{Softmax}(\frac{Q(K_{\text{id}})^T}{\sqrt{d}})V_{\text{id}} \\ Q = W_Q h_t, K_{\text{id}} = W_K^{\text{id}} C_{\text{id}}, V_{\text{id}} = W_V^{\text{id}} C_{\text{id}}, \end{cases} \tag{2}$$

where $h_t \in R^{n \times d}$ represents hidden feature maps at timestep t flattened along the height and width dimensions, d denotes the latent dimension, and $n = h \times w$ represents the token number and is the product of the height h and width w of the hidden feature maps. The output O is then added to the hidden feature maps. C_{id} is typically extracted using pretrained image encoders, such as the CLIP Image encoder [33] or face recognition models [6]. PuLID [13] employs both encoders [6, 33] jointly and further emphasizes preserving the original model's behavior by carefully controlling how C_{id} is injected using contrastive alignment losses, thereby preventing unintended contamination of the generation process.

3.3. T2I Adapter

T2I Adapter [28] presents a plug-and-play network that enables various spatial condition signals (e.g. sketches, edge maps, pose images, depth images) to guide the image generation process. The lightweight network consists of one convolution layer followed by two residual blocks at each scale. It maps input condition images into multi-scale condition features C_{s}, which capture spatial information at different levels of granularity. These extracted features are then added to the intermediate feature maps of the encoder within the UNET denoiser at each scale.

4. Methodology

The ID–shape conflict refers to the tendency of the diffusion model to favor either identity preservation or shape exaggeration when both conditions are provided, while failing to balance them. As illustrated in Figure 2, the ID-conditioned path $\mathcal{P}^{\text{i}}$ naturally guides the denoising trajectory toward faithful identity preservation, whereas the shape-conditioned path $\mathcal{P}^{\text{s}}$ drives it toward precise caricature geometry. When both conditions C_{id} and C_{s} are applied together, the ideal denoising trajectory should lie between these two extremes. However, since the regions near $\mathcal{P}^{\text{i}}$ and $\mathcal{P}^{\text{s}}$ exhibit higher probability densities than the intermediate region, the trajectory tends to collapse toward one side, overemphasizing either identity or shape.

To address this issue, we proposed CaricHarmony by constructing three denoising trajectories: the two contrastive paths $\mathcal{P}^{\text{i}}$ and $\mathcal{P}^{\text{s}}$, conditioned on C_{id} and C_{s} respectively, and the main path $\mathcal{P}^{\text{i+s}}$ conditioned on both. For clarity, we denote intermediate features from $\mathcal{P}^{\text{i}}$ and $\mathcal{P}^{\text{s}}$ using superscripts i and s, while features from $\mathcal{P}^{\text{i+s}}$ are left without superscripts.

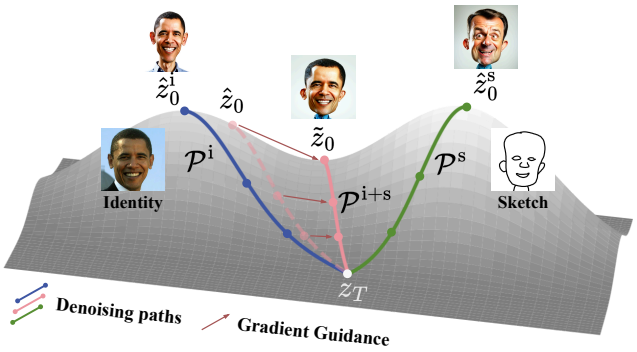


Figure 2. Dynamics of denoising paths in continuous sampling space of score-based diffusion [47]. Brighter color indicates areas where the denoised data is densely distributed. In blue (green) path, only C_{id} (C_{s}) is applied. Both conditions are applied in pink paths. The per-step gradient guidance from the energy function $\mathcal{E}_{\text{b}}$ is represented by red arrows.

We then impose inference-time cross-attention alignment guidance between the main path and each contrastive path, encouraging the main trajectory to stay within the region that balances identity and shape cues. The proposed gradient guidance mechanism explicitly steers the denoising direction away from biased extremes and toward the balanced region between $\mathcal{P}^{\text{i}}$ and $\mathcal{P}^{\text{s}}$, effectively resolving the ID–shape conflict without any additional training. The overall architecture is shown in Figure 3.

4.1. Shape Alignments

Without identity conditions, $\mathcal{P}^{\text{s}}$ fully encodes shape exaggeration according to the sketch. To ensure that $\mathcal{P}^{\text{i+s}}$ retains this capability, we align its intermediate cross-attention features with those from $\mathcal{P}^{\text{s}}$ during inference.

Inspired by PuLID [13], we use the query features Q^{s} from each cross-attention block as shape layout target, since they reflect how shape-related textual and sketch conditions are spatially grounded. By aligning Q with Q^{s}, the model can learn to maintain the original layout distribution of Q^{s} and render sufficiently exaggerated caricatures following the shape conditions. Specifically, we use a layout energy function $\mathcal{E}_{\text{layout}}$ to perform alignment between $Q = [q_1, q_2, ..., q_n]$ and $Q^{\text{s}} = [q_1^{\text{s}}, q_2^{\text{s}}, ..., q_n^{\text{s}}]$:

$$\mathcal{E}_{\text{layout}} = \sqrt{\sum_{j=1}^{n} \|q_j - q_j^{\text{s}}\|_2^2 \cdot c_j^{\text{s}}}, \tag{3}$$

where $c_j^{\text{s}} \in [0, 1]$ denotes a confidence weight indicating how strongly each token is associated with the sketch strokes. To compute weights, we first rasterize S into a binary stroke map, where pixels on sketch strokes are assigned value 1 and background pixels 0. We then resize this map to the spatial resolution of the query features to obtain C. This selectively preserves the shape cues tied to informative sketch strokes while suppressing irrelevant regions.

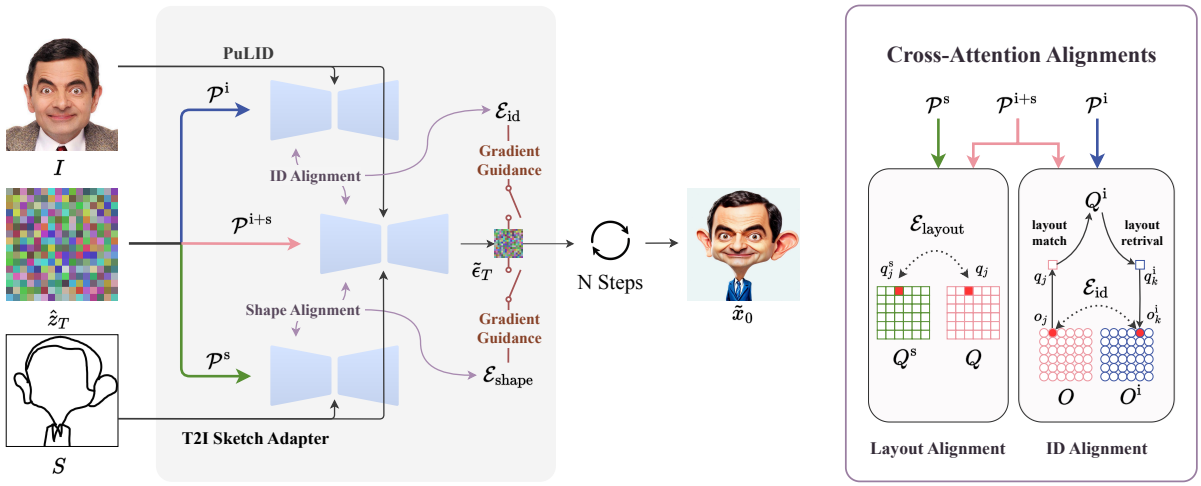


Figure 3. **Left: Inference pipeline of CaricHarmony**. The output path $\mathcal{P}^{\mathrm{i+s}}$ is jointly guided by two contrastive paths $\mathcal{P}^{\mathrm{s}}$ and $\mathcal{P}^{\mathrm{i}}$. $\mathcal{P}^{\mathrm{s}}$ is only fed the binary sketch map S as shape conditions and guides the shape alignment through $\mathcal{E}_{\mathrm{shape}}$, which comprises $\mathcal{E}_{\mathrm{layout}}$ and $\mathcal{E}_{\mathrm{sem}}$. $\mathcal{P}^{\mathrm{i}}$ is only fed the ID conditions I and guides the ID alignment through $\mathcal{E}_{\mathrm{id}}$. **Right: Cross-Attention Alignments.** Layout alignment obtains $\mathcal{E}_{\mathrm{layout}}$ by computing the token-wise differences between the query matrices from $\mathcal{P}^{\mathrm{i+s}}$ and $\mathcal{P}^{\mathrm{s}}$. ID alignment obtains $\mathcal{E}_{\mathrm{id}}$ by computing the token-wise differences between the ID cross attention outputs from $\mathcal{P}^{\mathrm{i+s}}$ and $\mathcal{P}^{\mathrm{i}}$. The associations between the output tokens for alignment are built by computing the similarities of corresponding query tokens.

If the face shapes of two generated caricatures are closely matched, they should localize textual condition signals similarly in cross-attention blocks. The attention map $\mathrm{Attn}(K_{\mathrm{txt}}, Q^{\mathrm{s}}, Q^{\mathrm{s}})$ of query Q to the textual key K_{txt} is hence used as additional signals to guide $\mathcal{P}^{\mathrm{i+s}}$ with semantic information:

$$\mathcal{E}_{\mathrm{sem}} = \|\mathrm{Attn}(K_{\mathrm{txt}}, Q, Q) - \mathrm{Attn}(K_{\mathrm{txt}}, Q^{\mathrm{s}}, Q^{\mathrm{s}})\|_2. \quad (4)$$

The combined shape energy is

$$\mathcal{E}_{\mathrm{shape}} = \mathcal{E}_{\mathrm{layout}} + \mathcal{E}_{\mathrm{sem}}. \quad (5)$$

4.2. ID Alignment Guidance

Solely following the shape guidance from $\mathcal{P}^{\mathrm{s}}$ can lead to a significant degradation of ID fidelity. To improve ID consistency, previous works typically employ specialized loss functions [4, 6, 30], which compare ID features between generated and reference images using feature extractors trained solely on real-domain photographs. However, the exaggerated shapes of caricatures introduce a significant domain gap, thereby reducing the effectiveness of these methods. Additionally, the ID alignment guidance should be performed across intermediate time-steps, where the predicted images are often noisy and flawed.

To overcome the issue, we design a novel ID energy function tailored for caricature synthesis. We guide $\mathcal{P}^{\mathrm{i+s}}$ using the ID-conditioned cross-attention outputs from $\mathcal{P}^{\mathrm{i}}$. Let $O = [o_1, ..., o_n]$ and $O^{\mathrm{i}} = [o^{\mathrm{i}}_1, ..., o^{\mathrm{i}}_n]$ be the outputs of cross-attention layers, where O^{i} serves as the target. However, direct alignment is inefficient, since O may present a significantly different layout distribution compared to O^{i} when conditioned on the exaggerated shape C_{s}. Using O^{i} to guide O naively will lead to a layout mismatch and confuse the model. Therefore, we build associations based on the relevance between query tokens. For each o_j, we find its corresponding o^{i}_k through

$$k = \arg\max_l \Phi(q_j, q^{\mathrm{i}}_l), \quad (6)$$

where $\Phi(\cdot,\cdot)$ is the cosine similarity between two vectors. The query-output pairs from $\mathcal{P}^{\mathrm{i}}$ behave like a ground-truth dictionary, where query tokens are "keys", and output tokens are "values". When a query token q_j arrives, we refer to the dictionary, look up q^{i}_k that best matches q_j, and retrieve its corresponding output token o^{i}_k to provide guidance.

The ID energy is computed with a confidence weight $c^{\mathrm{i}}_j \in [0, 1]$ as

$$\mathcal{E}_{\mathrm{id}} = \sqrt{\sum_{j=1}^{n} \|o_j - o^{\mathrm{i}}_k\|_2^2 \cdot c^{\mathrm{i}}_j}, \quad (7)$$

The confidence weight c^{i}_j aims to adapt the guidance strength:

$$c^{\mathrm{i}}_j = \frac{\Phi(q_j, q^{\mathrm{i}}_k) - \phi}{1 - \phi}, \quad \phi = \min_l \Phi(q_j, q^{\mathrm{i}}_l). \quad (8)$$

It helps weaken the guidance strength when the level of match between q_j and q^{i}_k is not distinguishable from the pool of target query tokens.

4.3. Gradient Guidance with Balance Conditions

$\mathcal{E}_{\mathrm{shape}}$ and $\mathcal{E}_{\mathrm{id}}$ are in opposition to each other. When the noisy latent is biased toward either side, the guidance

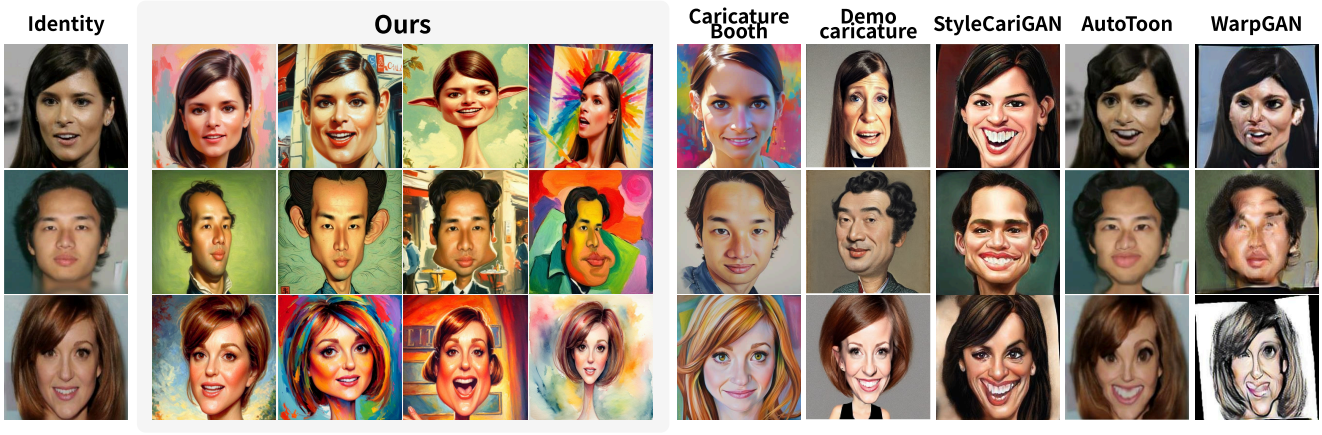


Figure 4. Qualitative performance comparison against caricature synthesis models.

strength from the other side increases, pulling the latent back to a balanced point. When combining both, the compositional energy function provides additional balance conditions C_{b} and helps the model find a sweet point on harmonizing ID and shape conditions with explicit gradient guidance.

Let the existing conditions be $C_{\mathrm{e}} = [C_{\mathrm{txt}}, C_{\mathrm{shape}}, C_{\mathrm{id}}]$. The score function in this work can be expressed as:

$$\nabla_{\hat{z}_t} \log q(\hat{z}_t|C_{\mathrm{b}}, C_{\mathrm{e}}) = \nabla_{\hat{z}_t} \log \frac{q(\hat{z}_t|C_{\mathrm{e}})q(C_{\mathrm{b}}|\hat{z}_t, C_{\mathrm{e}})}{q(C_{\mathrm{b}}|C_{\mathrm{e}})} \tag{9}$$

$$= \nabla_{\hat{z}_t} \log q(\hat{z}_t|C_{\mathrm{e}}) + \nabla_{\hat{z}_t} \log q(C_{\mathrm{b}}|\hat{z}_t, C_{\mathrm{e}}). \tag{10}$$

We apply classifier-free guidance [16] to the first term and compute gradients of $\mathcal{E}_{\mathrm{b}} = \mathcal{E}_{\mathrm{shape}} + \mathcal{E}_{\mathrm{id}}$ for the second. The final guided noise prediction $\tilde{\epsilon}_t$ is:

$$\hat{\epsilon}_t \leftarrow (1+\gamma)\epsilon_\theta(\hat{z}_t, t, C_{\mathrm{e}}) - \gamma\epsilon_\theta(\hat{z}_t, t, \emptyset), \tag{11}$$

$$\tilde{\epsilon}_t \leftarrow \hat{\epsilon}_t + \eta\nabla_{\hat{z}_t}\mathcal{E}_{\mathrm{b}}(\hat{z}_t, t, C_{\mathrm{b}}, C_{\mathrm{e}}), \tag{12}$$

where γ controls classifier-free guidance and η controls balance strength.

4.4. Timestep-Constrained Guidance

$\mathcal{E}_{\mathrm{shape}}$ guides layout distribution [13] at a coarse scale, while $\mathcal{E}_{\mathrm{id}}$ guides detailed ID information at a relatively fine-grained scale. To fit the coarse-to-fine nature of the denoising process [29], we only activate $\mathcal{E}_{\mathrm{shape}}$ and $\mathcal{E}_{\mathrm{id}}$ when t falls in $[t^{\mathrm{s}}_{\mathrm{start}}, t^{\mathrm{s}}_{\mathrm{end}}]$ and $[t^{\mathrm{i}}_{\mathrm{start}}, t^{\mathrm{i}}_{\mathrm{end}}]$. We encourage the guidance of $\mathcal{E}_{\mathrm{shape}}$ at early time intervals, setting $t^{\mathrm{s}}_{\mathrm{start}} = T = 1000$ and $t^{\mathrm{s}}_{\mathrm{end}} = 700$. $\mathcal{E}_{\mathrm{id}}$ starts relatively late at $t^{\mathrm{i}}_{\mathrm{start}} = 900$ and ends at $t^{\mathrm{i}}_{\mathrm{end}} = 400$ to prevent harming the quality of results fine-grained details [27].

5. Experiments

Implementation Details. Our model is built on SDXL [31] with the weights of Juggernaut-XL-v9 [37]. We use a pre-trained ID Encoder from PuLID v1.1 [13] and T2I-Sketch-Adapter-SDXL [28] to inject ID and shape conditions. During inference, the model is given textual prompts: "A highly exaggerated and detailed caricature of a man/woman." The output image resolution is 768×768. We use DPM++ 2M sampler [26] to accelerate inference speed, setting the number of inference steps N to 40 and using a classifier-free guidance scale γ of 7. The guidance rate η is set to 0.4. The experiments are performed on a single RTX4090, and the code is implemented with PyTorch. All the other settings follow the original PuLID-ID-Encoder and T2I-Sketch-Adapter.

Dataset. We use the WebCaricature dataset [19] to evaluate the performance of different models. Specifically, we consistently select the photo with the least sequence number for each identity as the reference ID image I and extract the edge map from each caricature associated with the identity as the sketch image S. The total number of samples is 1216.

Metrics. We use CLIP [33] scores to evaluate shape and ID consistency [3]: S-CLIP measures shape consistency by comparing edge maps of generated and ground-truth caricatures, while I-CLIP measures identity consistency between generated caricatures and identity images. To assess overall quality, we further compare our model with the other caricature synthesis models using ImageReward [51] and PickScore [24] as evaluation metrics.

Baselines. The models for comparison include warping-based networks StyleCariGAN [20], WarpGAN [41], and AutoToon [11], also diffusion-based approaches DemoCaricature [3], and CaricatureBooth [32].

5.1. Qualitative Evaluation

Figure 4 illustrates the performance of our model against several caricature synthesis models. StyleCariGAN [20], WarpGAN [41], and AutoToon [11] cannot explicitly control the shape distortion. DemoCaricature [3], similar to this work, introduces explicit shape control by leveraging free-hand sketches. CaricatureBooth [32] guides the shape

distortion with a fixed number of Bezier curves. Compared with the baselines, our model strikes a balance between preserving the facial features from identity images and ensuring a high level of creativity.

Figure 5 evaluates our method against several state-of-the-art diffusion-based models. DemoCaricature [3] is the first diffusion-based model tailored for caricature synthesis. PuLID [13] is a tuning-free personalization model. We plug in a T2I-Sketch-Adapter [28] to provide additional shape control. Our model can generate high-quality caricatures in 16 seconds, which is approximately 4× faster than DemoCaricature without per-sample tuning. Compared to PuLID, our results align substantially better with the provided shape conditions after balancing the shape and ID information.

5.2. Quantitative Evaluation

It is worth mentioning that CaricatureBooth [32] only accepts a fixed number of Bezier curves as input and cannot be applied to free-hand sketches and edge maps, which makes it incompatible with the experiment. The results are shown in Table 2.

When $\mathcal{E}_{\text{shape}}$ and $\mathcal{E}_{\text{id}}$ are disabled in turn, the model is biased toward ID preservation and shape exaggeration, and achieves the highest I-CLIP and S-CLIP score, respectively. By combining both, our final model finds a balanced point and generates results that respect both conditions (see Figure 7). As discussed before, custom evaluation models, including CLIP [33], cannot extract accurate facial features of caricatures because of the domain gap, causing I-CLIP to be highly sensitive to the level of shape distortion. Although the final model gets a slightly lower I-CLIP score compared to DemoCaricature, the results still show that our model can sufficiently preserve ID information under highly creative shape conditions (see Figure 5). Additionally, all our variants consistently outperform counterparts in ImageReward and PickScore due to SDXL's strong generative capabilities, while their similar scores confirm that our gradient guidance preserves aesthetic quality.

Methods	I-CLIP ↑	S-CLIP ↑	ImageReward ↑	PickScore ↑
StyleCariGAN	0.5228	-	0.4340	0.0637
WarpGAN	0.6634	-	-0.2588	0.1033
AutoToon	<u>0.7628</u>	-	-0.4978	0.1094
DemoCaricature	0.7591	0.8450	0.2871	0.0949
Ours (full)	0.7512	<u>0.8615</u>	<u>0.8509</u>	<u>0.2049</u>
w/o $\mathcal{E}_{\text{shape}}$	**0.7747**	0.8296	**0.8967**	0.2038
w/o $\mathcal{E}_{\text{id}}$	0.7381	**0.8698**	0.8443	**0.2199**

Table 2. Quantitative Comparison.

5.2.1. User study

We conduct a user study to evaluate the model's performance in terms of: (1) identity consistency with identity images, (2) shape consistency with sketches, and (3) overall quality. The study involved 20 identities (10 men and 10 women) and 20 exaggerated hand-drawn sketches (10 for men and 10 for women), resulting in a pool of 200 caricature samples. Each of the 16 volunteers received 10 samples, each asking them to rate the caricatures generated by different models on a discrete scale from 1 to 10 across the three aspects. The results are presented in Table 3. Our model achieves the highest ID consistency score, benefiting from humans' ability to capture facial features under shape deformations. By addressing the conflict with shape conditions, it also achieves the highest shape score, demonstrating overall superior performance.

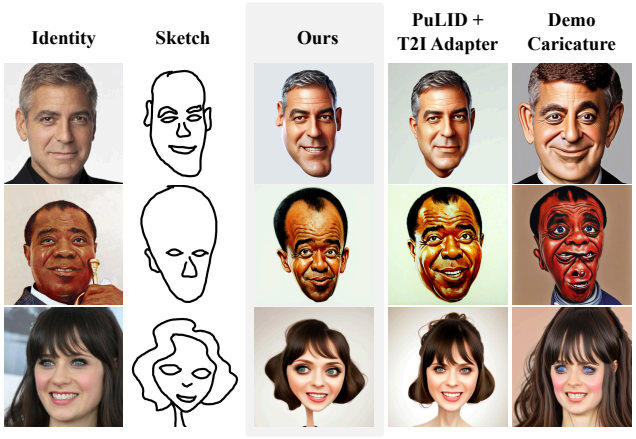


Figure 5. Qualitative performance comparison against diffusion-based models.

Methods	ID ↑	Shape ↑	Overall ↑
StyleCariGAN	4.59	-	4.91
WarpGAN	4.81	-	4.62
AutoToon	6.73	-	5.50
DemoCaricature	6.03	5.51	6.06
Ours	**6.83**	**8.08**	**7.81**

Table 3. Results of the user study. Each value represents the model's average score under a specified criterion.

6. Ablation Study

6.1. Impact of Different Guidance Terms

Figure 7 shows the results under different gradient guidance settings. Without $\mathcal{E}_{\text{shape}}$, the model preserves facial structure well but fails to introduce sufficient exaggeration and lacks proper alignment with the target shape conditions, resulting in outputs that appear more like conventional portraits rather than expressive caricatures. Without $\mathcal{E}_{\text{id}}$, the model is biased toward the shape conditions without respecting ID information, leading to outputs with diminished recognizability and less fine-grained textural details. In the absence of TCG, the model produces suboptimal results with less fine-grained details and generates unwanted artifacts that compromise the overall image quality.



Figure 6. Interpolation of the scaling factors of $\mathcal{E}_{\text{id}}$ and $\mathcal{E}_{\text{shape}}$. From left to right, the scaling factor of $\mathcal{E}_{\text{id}}$ linearly decreases from 2 to 0, while that of $\mathcal{E}_{\text{shape}}$ linearly increases from 0 to 2. The results in the middle match our default setting.

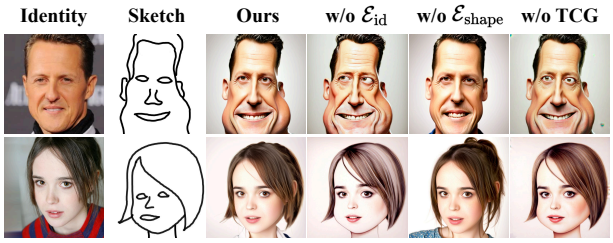


Figure 7. We disable different guidance terms respectively to demonstrate their functionalities. When timestep-constrained guidance (TCG) is disabled, $\mathcal{E}_{\text{shape}}$ and $\mathcal{E}_{\text{id}}$ are activated across all timesteps.

With the full version incorporating all guidance terms, the model achieves a balance where the generated caricatures faithfully adhere to the shape conditions while effectively preserving the identity-specific characteristics, demonstrating the competing yet complementary nature of each energy function in resolving the ID-shape conflict.

6.2. Impact of Different Guidance Scales

Figure 6 presents the outputs produced when $\mathcal{E}_{\text{id}}$ and $\mathcal{E}_{\text{shape}}$ are weighted by different scaling factors. Increasing the scaling factor of $\mathcal{E}_{\text{id}}$ while decreasing that of $\mathcal{E}_{\text{shape}}$ yields results that preserve more identity information while following shape conditions less strictly, and vice versa. This smooth transition between identity-dominant and shape-dominant modes enables users to adjust the balance according to their specific preferences, demonstrating the flexibility of our model.

6.3. Extreme Shape Exaggerations

Our model shows robust compatibility with extreme shape conditions. As illustrated in Figure 8, it can effectively resolve the conflict between ID preservation and shape exaggeration, producing caricatures with high ID fidelity even under highly exaggerated shape conditions. Notably, our approach maintains stable performance and generates high-quality outputs with recognizable identity characteristics

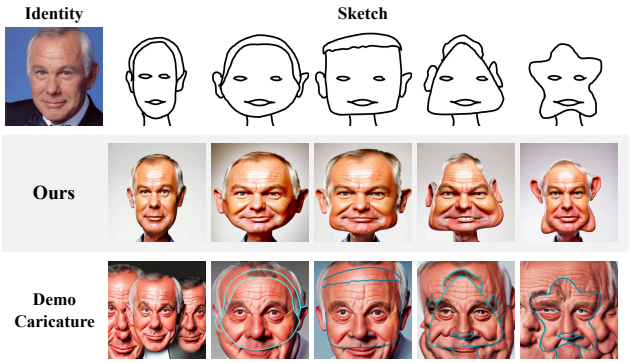


Figure 8. Performance comparison when sketches with different levels of exaggeration are used.

even when the facial structures undergo drastic deformations. In contrast, the prior work [3] struggles to interpret the spatial information from sketches when they are highly exaggerated because of the interruption of the conditioning signals. More examples are provided in Appendix B.

7. Conclusion

We present a caricature synthesis model that achieves a remarkable performance improvement by explicitly handling the ID-shape conflict. The core of our approach is a training-free framework that aligns hidden features in the cross-attention mechanism with those from two contrastive paths, enabling dynamic control over identity preservation and shape exaggeration. Tailored energy functions are designed to provide the model with additional information about the degree of deviation from either ID or shape conditions, pushing it to strike a balance on harmonizing both. By strategically harnessing the manifold of pretrained T2I models, our approach circumvents the data scarcity problem that hinders the development of caricature synthesis. The resulting framework possesses the capability to output high-quality caricatures that can not only preserve high ID fidelity but also retain the shape exaggeration defined by free-hand sketches.

References

[1] Black Forest Labs. Flux.1-dev. https://github.com/black-forest-labs/flux, 2024. 3

[2] Kaidi Cao, Jing Liao, and Lu Yuan. Carigans: Unpaired photo-to-caricature translation. *arXiv preprint arXiv:1811.00222*, 2018. 2

[3] 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 *CVPR*, 2024. 2, 3, 6, 7, 8

[4] Li Chen, Mengyi Zhao, Yiheng Liu, Mingxu Ding, Yangyang Song, Shizun Wang, Xu Wang, Hao Yang, Jing Liu, Kang Du, et al. Photoverse: Tuning-free image customization with text-to-image diffusion models. *arXiv preprint arXiv:2309.05793*, 2023. 5

[5] Wenqing Chu, Wei-Chih Hung, Yi-Hsuan Tsai, Yu-Ting Chang, Yijun Li, Deng Cai, and Ming-Hsuan Yang. Learning to caricature via semantic shape transform. *IJCV*, 2021. 2, 3

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 4, 5

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 3

[8] Jean Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. In *Constructive theory of functions of several variables: Proceedings of a conference held at Oberwolfach April 25–May 1, 1976*, 2006. 3

[9] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. In *NeurIPS*, 2023. 3

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 3

[11] Julia Gong, Yannick Hold-Geoffroy, and Jingwan Lu. Autotoon: Automatic geometric warping for face cartoon generation. In *WACV*, 2020. 2, 6

[12] Zheng Gu, Chuanqi Dong, Jing Huo, Wenbin Li, and Yang Gao. Carime: Unpaired caricature generation with multiple exaggerations. *IEEE TMM*, 2021. 2

[13] Zinan Guo, Yanze Wu, Chen Zhuowei, Peng Zhang, Qian He, et al. Pulid: Pure and lightning id customization via contrastive alignment. In *NeurIPS*, 2024. 3, 4, 6, 7

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[15] Zecheng He, Bo Sun, Felix Juefei-Xu, Haoyu Ma, Ankit Ramchandani, Vincent Cheung, Siddharth Shah, Anmol Kalia, Harihar Subramanyam, Alireza Zareian, et al. Imagine yourself: Tuning-free personalized image generation. *arXiv preprint arXiv:2409.13346*, 2024. 3

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop*, 2021. 6

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

[18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 2

[19] Jing Huo, Wenbin Li, Yinghuan Shi, Yang Gao, and Hujun Yin. Webcaricature: a benchmark for caricature recognition. In *BMVC*, 2018. 6

[20] Wonjong Jang, Gwangjin Ju, Yucheol Jung, Jiaolong Yang, Xin Tong, and Seungyong Lee. Stylecarigan: caricature generation via stylegan feature map modulation. *ACM TOG*, 2021. 2, 6

[21] Jian Jin, Zhenbo Yu, Yang Shen, Zhenyong Fu, and Jian Yang. Latexblend: Scaling multi-concept customized generation with latent textual blending. In *CVPR*, 2025. 3

[22] Hao Kang, Stathi Fotiadis, Liming Jiang, Qing Yan, Yumin Jia, Zichuan Liu, Min Jin Chong, and Xin Lu. Flux already knows–activating subject-driven image generation without training. *arXiv preprint arXiv:2504.11478*, 2025. 3

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2

[24] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *NeurIPS*, 2023. 2, 6

[25] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 3

[26] 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. 6

[27] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. In *ICLR*, 2024. 6

[28] 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, 6, 7

[29] Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. In *ICCV*, 2023. 6

[30] Xu Peng, Junwei Zhu, Boyuan Jiang, Ying Tai, Donghao Luo, Jiangning Zhang, Wei Lin, Taisong Jin, Chengjie Wang, and Rongrong Ji. Portraitbooth: A versatile portrait model for fast identity-preserved personalization. In *CVPR*, 2024. 5

[31] 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. 3, 6

[32] 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. 2, 3, 6, 7

[33] 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. 2, 4, 6, 7

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3

[36] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 3

[37] RunDiffusion. Juggernaut XL v9 + RunDiffusion Photo v2. https://huggingface.co/RunDiffusion/Juggernaut-XL-v9, 2024. 6

[38] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 3

[39] Lejun Shao and Hao Zhou. Curve fitting with bezier cubics. *Graphical models and image processing*, 1996. 3

[40] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *CVPR*, 2024. 3

[41] Yichun Shi, Debayan Deb, and Anil K Jain. Warpgan: Automatic caricature generation. In *CVPR*, 2019. 2, 6

[42] Enis Simsar, Thomas Hofmann, Federico Tombari, and Pinar Yanardag. Loraclr: Contrastive adaptation for customization of diffusion models. In *CVPR*, 2025. 3

[43] Jaskirat Singh, Stephen Gould, and Liang Zheng. High-fidelity guided image synthesis with latent diffusion models. In *CVPR*, 2023. 3

[44] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 3

[45] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019. 3

[46] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *NeurIPS*, 2020. 3

[47] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 3, 4

[48] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH*, 2023. 3

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[50] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH*, 2023. 3

[51] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *NeurIPS*, 2023. 2, 6

[52] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arxiv:2308.06721*, 2023. 3