BoLID: Body-Lightning ID Diffusion

Daniil Kazachkov
Moscow Institute of Physics and Technology
Dolgoprudny
kazachkov.di@phystech.edu

Andrei Filatov
Skolkovo Institute of Science and Technology
filatovandreiv@gmail.com

Abstract

Large-scale generative models have demonstrated remarkable performance in producing images from textual descriptions. A significant application domain of this models is the generation of personalized images. Common approaches to enhance personalization typically rely on additional inputs, such as control signals or multiple visual perspectives of the subject. In this work, we propose a method that enables high-quality personalized image generation using only textual and image reference input, without need for any structural control. The core idea lies in training a latent vector within a Variational Autoencoder (VAE) framework to encode information about a person's body shape and composition. Our model, *Body Lightning ID Diffusion* (BoLID), extends the IP-Adapter framework by preserving the integrity of the base diffusion model. Instead of modifying the backbone, we enrich the latent representation prior to training the adapters. These adapters can then be conditioned either on the learned latent vector or directly on an input image. Our approach accomplishes competitive results on standard evaluation metrics such as Fréchet Inception Distance (FID) and Inception Score (IS).

Keywords: Machine Learning, Diffusion Models.

1 Introduction

Significant advances in image generation have been made through architectures such as U-Net [1], DALL-E 2 [2], Imagen [3] и Stable Diffusion [4]. Initially, users could generate images based on textual descriptions, which severely limited the flexibility of the model. The DALL-E 2 model took a step forward by integrating visual representations into the generation process using diffusion models. Further development led to compact adapters such as IP-Adapter [5], which split the *cross-attention* mechanism between textual and visual features. This made it possible to introduce controlled generation elements, similar to ControlNet [6], and to use third-party models pre-trained on the same underlying architecture. The InstantID [7] and PuLID [8] models proposed additional mechanisms to control user identity, including facial features and key points, which significantly improved ID fidelity. However, these approaches focus solely on facial features, ignoring the morphological characteristics of the body.

In this paper, we propose a method that takes into account the user's individual body features by introducing a *Body-ID* representation into the generation process. In addition, our solution allows to discard everything but the textual prompt without degrading the avatar quality. The key challenge is to preserve informative features from reference images in the hidden vector.

To improve ID similarity, we pre-train the model on specially collected datasets grouped by unique user IDs. Each ID is represented by a series of images from different perspectives, which provides a holistic perception of the figure when generated from a single input image.

The proposed solution comprises two main modules:

1. **Body-ID Encoder** — a variational autoencoder (VAE) trained to aggregate features from multiple perspectives weighted by pose confidence extracted by the YOLOv11 [9] model. The result is a compact latent vector reflecting body morphology.

2. **Body-Condition Module** — an adapter that implements a *cross-attention* mechanism between text, face and body features, similar to IP-Adapter [5].

2 Related Works

Modern diffusion models have significantly improved the generation of images from textual descriptions, achieving photorealistic quality. Nevertheless, personalised human image generation remains challenging due to the high variability of anatomical features and limited individual data. This section discusses key approaches to preserve identity and tailor generative models to the individual user.

The authors of **IP-Adapter** [5] proposed separate processing of text and image features by modifying the *cross-attention* mechanism to integrate visual information into the generation process. This approach allowed conditioning the diffusion generation process not only on text but also on image to achieve better consistency with the user's query. However, the spatial characteristics of the reference image are not taken into account, which does not allow to achieve a high-quality generation of the human body.

Further research has focused on extending the idea of lightweight adapters to improve identity and adapting models to individual traits. **InstantID** [7] achieves highly accurate appearance preservation by utilising InsightFace's pre-trained face encoder. The model operates in zero-shot mode and requires only one reference image, making it particularly easy to use, but it is limited to facial identity only and does not take into account the user's body. The **PuLID** [8] model proposes a contrastive learning mechanism for smoothly combining different identities. This is particularly effective in generating images with combined stylisation, e.g. 'user A's face in B's style'. However, the focus also remains solely on the face. The **PhotoMaker** [10] approach offers a *Stacked ID Embedding* strategy that uses multiple photos of the same person from different angles. This allows for a high degree of identity preservation without having to modify the parameters of the underlying diffusion model. However, this method requires more than three images and is not adapted for body pose modelling.

Таблица 1: Comparative characteristics of personalised generation methods.

Method	**References**	**Saving ID similarity**	**Modality**
DreamBooth	3–5	High	Only text
IP-Adapter	1	Medium	Image + Text
PhotoMaker	3+	Very high	Image + Text
InstantID	1	High	Image + Text
PuLID	1	Very high	Image + Text

Our approach extends the ideas of **IP-Adapter** by introducing additional control over the pose and shape of the user's body. **The key difference of our approach is that only a textual promt is sufficient for generation - the model retains high quality of body reconstruction even in the absence of an input image**. If an image is still used, the model does not require a set of perspectives or multiple images: the spatial and kinematic characteristics of the body are reconstructed by training a latent vector extracted by the VAE encoder. This approach makes the generation robust to limitations on the number or variety of reference images and increases the versatility of the system in real-world applications.

3 Preliminary

3.1 Diffusion models. DDPM

The DDPM [11] operation consists of two processes: forward and reverse. The forward process is the successive denoising of the input image x_0 in T steps, where x_t is calculated using the following formula:

$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\varepsilon, \tag{1}$$

where $\varepsilon \sim \mathcal{N}(0, I)$, and β_t is a hyperparameter chosen so that each successive image x_t is more heavily noisy,

$$x_t|x_{t-1} \sim \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t I). \tag{2}$$

When $T \to \infty, x_T \to \mathcal{N}(0, I)$, that is, a Gaussian noise is obtained at the last iteration step.

Let $\alpha_t = 1 - \beta_t, \overline{\alpha_t} = \prod_{s=1}^{t} \alpha_s$. Then

$$x_t = \sqrt{\overline{\alpha_t}} x_0 + \sqrt{1 - \overline{\alpha_t}} \varepsilon, \tag{3}$$

where $\varepsilon \sim \mathcal{N}(0, I)$,

$$x_t | x_0 \sim \mathcal{N}(\sqrt{\overline{\alpha_t}} x_0, (1 - \overline{\alpha_t}) I). \tag{4}$$

During the inverse process, the original image is reconstructed from the noise. We know $x_T \sim \mathcal{N}(0, I)$. The sampling is iterative:

$$\hat{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\hat{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \overline{\alpha_t}}} \hat{\epsilon}_\theta(\hat{x}_t, t) \right) + \hat{\sigma}_t z, \tag{5}$$

where $\hat{x}_t$ is the reconstructed image at iteration t, and $\hat{x}_T = x_T$; $\hat{\epsilon}_\theta(\hat{x}_t, t)$ — is the noise reconstruction obtained by the model $\hat{\epsilon}_\theta$ for $\hat{x}_t$; z — noise to generate different images, with $z \sim \mathcal{N}(0, I)$ if $t > 1$, otherwise $z = 0$.

3.2 Conditional generation: Classifier-free Guidance

The classifier-free guidance[12] method increases the degree to which the model is guided by the class identifier c. During sampling, the prediction is obtained by linearly combining the predictions of the conditioned and unconditioned models:

$$\hat{\epsilon}_\theta(x_t, c, t) = (w + 1)\epsilon_\theta(x_t, c, t) - w\epsilon_\theta(x_t, t), \tag{6}$$

where w is the weight coefficient, $t \in [0, T]$ is the time step of the diffusion process, x_t is the noisy image at step t.

This model is based on the pre-trained diffusion text-to-image model $\hat{x}_\theta$, whose loss function is defined as:

$$\mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)} w_t \| \hat{x}_\theta(\alpha_t \mathbf{x} + \sigma_t \epsilon, \mathbf{c}) - \mathbf{x} \|^2, \tag{7}$$

where $\mathbf{x}$ is latent representation of the original image, $\mathbf{c} = \Gamma(P)$ is condition vector obtained with text encoder Γ and text prompt P, $t \in [0, T]$ denotes the time step of the diffusion process; α_t, σ_t, w_t are predetermined functions from t defining the diffusion process. The original diffusion model is pre-trained on several input images of a single object paired with a text prompt containing the name of the class to which the object belongs. Data $\mathbf{x_{pr}} = \hat{x}(z, \mathbf{c_{pr}})$ is generated using a sampler based on a pre-trained diffusion model with random initial noise $z \sim \mathcal{N}(0, I)$ condition vector $\mathbf{c_{pr}} := \Gamma(f(\text{"a [class noun]"}))$, where f is a tokeniser. The loss function takes the following form:

$$\mathbb{E}_{\epsilon, \epsilon'} [w_t \| \hat{x}_\theta(\alpha_t \mathbf{x} + \sigma_t \epsilon, \mathbf{c}) - \mathbf{x} \|^2 + \lambda w_{t'} \| \hat{x}_\theta(\alpha_{t'} \mathbf{x_{pr}} + \sigma_{t'} \epsilon', \mathbf{c_{pr}}) - \mathbf{x_{pr}} \|^2], \tag{8}$$

where λ is a weighting factor. Image generation is done by embedding a unique identifier in a text prompt in the form: "a [identifier] [class noun]".

3.3 Stable Diffusion

Stable Diffusion is a text-to-image model that takes a textual prompt as input. CLIP converts it into an embedding that directs the generation. Next, random noise in latent space is generated. The U-Net model removes the noise in a reverse diffusion process, at each step considering the textual embedding so that the image matches the description. Once this step is complete, the VAE decoder translates the latent representation into an image of the original size.

3.4 Adapting the image via IP-Adapter

The IP-Adapter consists of two parts: an image encoder to extract image features from the prompt and adapted modules with an isolated cross-attention mechanism to embed image features into a pre-trained text-to-image model.

Let the features of the image $\mathbf{c_i}$ and the text $\mathbf{c_t}$ be given. Since the method is based on the idea of using text and image separately, we get the following formula for cross-attention

$$\mathbf{Z}^{new} = softmax(\frac{\mathbf{Q}\mathbf{K}^\mathbf{T}}{\sqrt{d}})\mathbf{V} + \lambda \cdot softmax(\frac{\mathbf{Q}(\mathbf{K}')^\mathbf{T}}{\sqrt{d}})\mathbf{V}',$$

where λ is image weighting; $\mathbf{Q} = \mathbf{Z}\mathbf{W_q}, \mathbf{K} = \mathbf{c_t}\mathbf{W_k}, \mathbf{V} = \mathbf{c_t}\mathbf{W_v}$ are queries, keys and values matrices of attention mechanism for textual features, respectively, and $\mathbf{W_q}, \mathbf{W_k}, \mathbf{W_v}$ are the corresponding weight

matrices; $\mathbf{K}' = \mathbf{c_i}\mathbf{W}'_\mathbf{k}, \mathbf{V}' = \mathbf{c_i}\mathbf{W}'_\mathbf{v}$ are queries, keys and values matrices of attention mechanism for image features, respectively, and $\mathbf{W}'_\mathbf{k}$, $\mathbf{W}'_\mathbf{v}$ are the corresponding weight matrices. Since the UNet model is frozen, only $\mathbf{W}'_\mathbf{k}$ and $\mathbf{W}'_\mathbf{v}$ are trainable parameters.

The following loss function is minimised in the training process:

$$\mathcal{L}(\epsilon, \epsilon_\theta) = \mathbb{E}_{\mathbf{x_0},\epsilon,\mathbf{c_t},\mathbf{c_i},t}\|\epsilon - \epsilon_\theta(\mathbf{x_t}, \mathbf{c_t}, \mathbf{c_i}, t)\|^2, \tag{9}$$

where $\mathbf{x_t}$ is the noisy image at step t.

In order to engage classifier-free guidance in the inference phase, image conditions are randomly discarded during training:

$$\hat{\epsilon}_\theta(\mathbf{x_t}, \mathbf{c_t}, \mathbf{c_i}, t) = w\epsilon_\theta(\mathbf{x_t}, \mathbf{c_t}, \mathbf{c_i}, t) + (1 - w)\epsilon_\theta(\mathbf{x_t}, t) \tag{10}$$

If the image condition is discarded, the embedding of the corresponding image is nullified.

Thus, the Stable Diffusion model with IP-Adapter allows us to train the compact module on images and prompts while keeping the rest of the model fixed.

4 Method

This section describes in detail the proposed *BoLID* (Body Language Identity Diffusion) architecture and its training procedure. The main goal of the method is to integrate the user's body features into the process of text-oriented image generation without losing the quality of the modelled figure.

Consider a dataset $\mathfrak{D} = \{(I_i, \tau_i) \mid i = 1, \ldots, n\}$, where I_i is the user's image and τ_i is the corresponding text prompt. It is assumed that each I_i contains both face and body information of the user. Let ϵ_θ be a model from the class of diffusion models.

4.1 Gaussian weighting

We will use Gaussian weighting for each key body point. This will assign a high weight to the parts of the image close to the key points and effectively ignore the background.

For a key point with coordinates (x_i, y_i), $i = \overline{1, 17}$, the weight of a pixel with coordinates (x, y) is determined by a Gaussian function:

$$w_i(x, y) = \exp\left(-\frac{(x - x_i)^2 + (y - y_i)^2}{2\sigma^2}\right), \tag{11}$$

where σ is a parameter controlling the radius of influence of the point. The larger σ is, the wider the area of influence.

Suppose we have K keypoints, the final weight for pixel (x, y) is the maximum of the weights from all points:

$$W(x, y) = \max_{i=1,\ldots,K} w_i(x, y) \tag{12}$$

To keep the weights in the range $[0, 1]$ and not distort the intensity of the image, we normalise:

$$W_{\text{norm}(x,y)} = \frac{W(x, y)}{\max(W(x, y))} \tag{13}$$

Then we define the image weighting by the formula $I_w(x, y) = I(x, y) \cdot \big(\alpha + (1 - \alpha)W_{\text{norm}(x,y)}\big)$, where $\alpha \in [0, 1]$ allows the image background to be completely unweighted.

4.2 Latent knowledge vector

The first step is to extract the vector $c_{\text{body}} \in \mathbb{R}^{756}$. Each image I_j from the input set of views $\{I_j\}_{j=1}^{100}$ is reduced to size 1024×1024 and then processed by the Γ_{body} model to extract structural control elements $\{s_j\}_{j=1}^{100}$, $s_j \in \mathbb{R}^{17}$. In our case Γ_{body} is a YOLO11-pose model, if some point of the body is not visible, the corresponding value of the vector component is nullified. To select image fragments, we apply Gaussian weighting, obtaining $\{I_w^j\}_{j=1}^{100}$. All I_w^j images are combined into a single tensor, which is fed as input to a VAE encoder whose output is the vector $c_{\text{body}} \in \mathbb{R}^{756}$.

4.3 Body-Condition Module

The conditional generation module takes as input the vector c_{cbody}, the text description τ_i and the features of the reference image I_{ref} if available. Next, we need to train the adapters integrating them into the generative process. For this purpose, a module similar to IP-Adapter [5] is used, extended to accommodate multiple sources of information simultaneously:

$$\mathbf{Z}' = \text{Attention}(\mathbf{Q}, \mathbf{K}_t, \mathbf{V}_t) + \lambda_{\text{ref}} \, \text{Attention}(\mathbf{Q}, \mathbf{K}_{\text{ref}}, \mathbf{V}_{\text{ref}}) + \lambda_b \, \text{Attention}(\mathbf{Q}, \mathbf{K}_b, \mathbf{V}_b),$$

where

- $(\mathbf{K}_t, \mathbf{V}_t)$ are key and value pairs from the text encoder $\Gamma_\tau(\tau)$;
- $(\mathbf{K}_{\text{ref}}, \mathbf{V}_{\text{ref}})$ are key and value pairs from the image encoder for the reference image;
- $(\mathbf{K}_b, \mathbf{V}_b)$ are key and value pairs from the Body-ID Encoder.

Varying the hyperparameters $\lambda_{\text{ref}}, \lambda_b$ gives additional degrees of freedom. This will allow the model to place more emphasis on either the face or the body during the generation process. Note that in the absence of a reference image, the λ_{ref} summand is nullified. This approach allows the visual representation of the body to be decomposed through aggregated features of other images, making generation from a single textual description possible. Given an image as input, the model uses it as a direct source of body embedding without requiring multiple perspectives, as the structural information is already encoded in the hidden VAE space.

4.4 Full objective

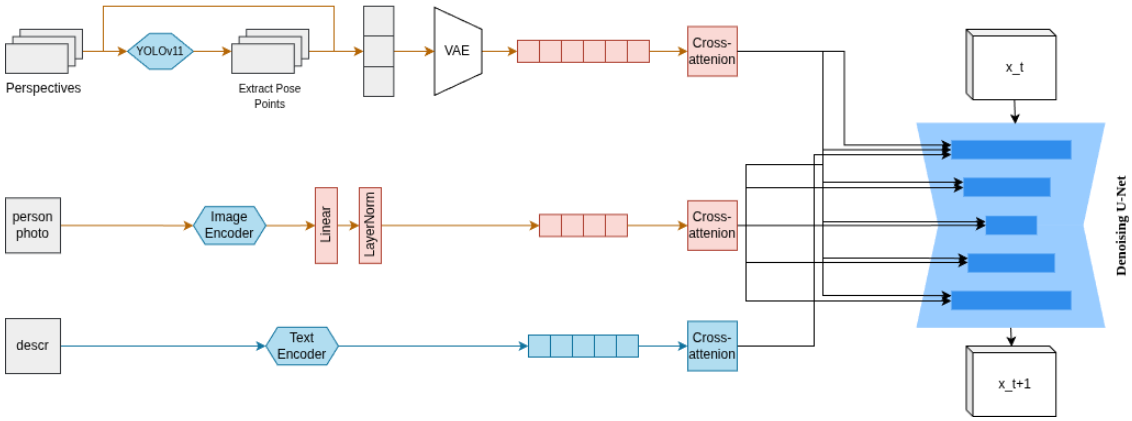

Рис. 1: *BoLID* architecture diagram.

The entire system is trained by minimising the weighted sum of losses:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \alpha_{\text{id}} \, \mathcal{L}_{\text{id}} + \beta_{\text{CLIP}} \, \mathcal{L}_{\text{CLIP}}, \tag{14}$$

where

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\epsilon} \left\| \epsilon - \epsilon_\theta(x_t, \mathbf{c}_\tau, \mathbf{c}_{\text{ref}}, \mathbf{c}_{\text{body}}, t) \right\|^2,$$

$$\mathcal{L}_{\text{id}} = \sum_{k=1}^{17} \left\| p_k^{\text{gen}} - p_k^{\text{ref}} \right\|^2,$$

$$\mathcal{L}_{\text{CLIP}} = 1 - \cos\left(\text{CLIP}(x_{\text{gen}}), \, \text{CLIP}(\tau)\right).$$

5 Metrics

The following metrics are proposed to evaluate the quality of the *BoLID* model:

- **FID (Fréchet Inception Distance)**[13] measures the distance between the distributions of real and generated images in the InceptionV3 feature space:

$$FID = ||\mu_p - \mu_q||^2 + Tr(\mathbf{\Sigma_p} + \mathbf{\Sigma_q} - 2(\mathbf{\Sigma_p}\mathbf{\Sigma_q})^{1/2}) \tag{15}$$

  where μ_p and μ_q are the mean feature values in the real and generated images, respectively, $\mathbf{\Sigma_p}$ and $\mathbf{\Sigma_q}$ are the covariance matrices for the feature distributions in the real and generated images, respectively.
- **IS (Inseption Score)** [14]

6 Dataset

Since our goal is to embed the user's body, it is important to pre-train the model on multiple angles of a single person. Therefore, high quality (at least 1000×1000 px resolution) celebrity photos were chosen as the dataset. For each person, a package of at least 100 photos from theplace using a written parser.

The processing of the dataset consisted of removing images with watermarks and more than one person in the image. The watermark-detection model was used to detect watermarks, and the YOLOv11n-face-detection model was used to check the number of people.

7 Computational experiment

8 Conclusion

Список литературы

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[2] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[3] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

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

[5] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023.

[6] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[7] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

[8] Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, Peng Zhang, and Qian He. Pulid: Pure and lightning id customization via contrastive alignment. In *Advances in Neural Information Processing Systems*, 2024.

[9] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024.

[10] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding, 2023.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[14] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[15] Xiao Yang Shanchuan Lin, Anran Wang. Sdxl-lightning: Progressive adversarial diffusion distillation. 2024.

[16] 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. 2023.

[17] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. 2023.

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 23–29 Jul 2023.

[19] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.