





HybridBooth: Hybrid Prompt Inversion for Efficient Subject-Driven Generation

Shanyan Guan^{1*}, Yanhao Ge^{1,3*}, Ying Tai^{2[✉]}, Jian Yang²
Wei Li¹, and Mingyu You^{3**[✉]}

¹ vivo Mobile Communication Co., Ltd

² School of Intelligence Science and Technology, Nanjing University

³ College of Electronic and Information Engineering, Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University

{guanshanyan, halege}@vivo.com yingtai@nju.edu.cn myyou@tongji.edu.cn

<https://sites.google.com/view/hybridbooth>

Abstract. Recent advancements in text-to-image diffusion models have shown remarkable creative capabilities with textual prompts, but generating personalized instances based on specific subjects, known as subject-driven generation, remains challenging. To tackle this issue, we present a new hybrid framework called **HybridBooth**, which merges the benefits of optimization-based and direct-regression methods. **HybridBooth** operates in two stages: the Word Embedding Probe, which generates a robust initial word embedding using a fine-tuned encoder, and the Word Embedding Refinement, which further adapts the encoder to specific subject images by optimizing key parameters. This approach allows for effective and fast inversion of visual concepts into textual embedding, even from a single image, while maintaining the model’s generalization capabilities.

Keywords: Customized Generation · Image Inversion · Image synthesis

1 Introduction

Recent advancements in text-to-image diffusion models [25, 34] have demonstrated unprecedented creating capabilities guided by textual prompts, such as creative endeavors in visual arts [28], synthetic dataset build-up [45], and image editing for marketing and advertising [24]. However, creating personalized instantiations based on specific subjects remains a challenging problem, also known as subject-driven generation [9, 10, 30, 36, 46].

To achieve subject-driven generation, recent works [9, 10, 36] explore the inversion of subject images into specific textual embeddings, either through iterative optimization (optimization-based) or direct regression methods. These two paradigms are summarized in Fig. 2. Optimization-based methods, such as

* Equal contributions. Ying Tai and Mingyu You are corresponding authors.

** Mingyu You is also affiliated with State Key Laboratory of Intelligent Autonomous Systems, Frontiers Science Center for Intelligent Autonomous Systems, Shanghai Key Laboratory of Intelligent Autonomous Systems



Fig. 1: We propose HybridBooth for efficient subject-driven generation. Unlike Fast-Composer [48], which targets humans, it cannot affect the Bronze Mask (Top). Compared to optimization-based [19, 36] and direct-regression-based methods [47, 48], our framework generates diverse, precise images while preserving the subject’s identity, requiring only 3 ~ 5 optimization iterations.

Textual Inversion and DreamBooth [9, 37], iteratively optimize the generation model or textual embedding of the subject to better align with given subject images. Constrained by tens of subject images, optimization-based methods can accurately capture the identifying characteristics of the subject. However, the optimization procedure is *computationally expensive* and *slow to converge*. In comparison, direct-regression-based methods [1–3, 7, 10, 39–41, 44, 47] achieve textual concept inversion by tuning a pre-trained model with labeled datasets or concept-specific images. While these methods enable zero-shot subject-driven generation, they suffer from *detail loss and adaptation issues* when combining the target subject with new styles or prompts.

In this paper, we introduce a novel framework called HybridBooth, designed for efficient subject-driven generation using text-to-image diffusion models. The core idea is to synergistically combine the strengths of various methods, creating a hybrid approach that balances quality and speed. As shown in Fig. 2, HybridBooth operates in two stages. The first stage, termed Word Embedding Probe, employs an encoder to generate a robust initial word embedding. This encoder is fine-tuned on a large-scale dataset such as FFHQ [14] with 70,000 images. In the second stage, the domain-agnostic encoder undergoes further fine-tuning to adapt to a specific subject’s image in a word embedding refinement process. This stage focuses on optimizing the most influential parameters of the regressor. These two stages work collaboratively, enhancing each other to achieve precise and efficient subject-driven generation.

Overall, our contributions are summarized as follows:

- Combining strengths of both optimization-based and direct-regression-based paradigms, we propose HybridBooth for precise and efficient subject-driven generation in less than 5 iteration steps.

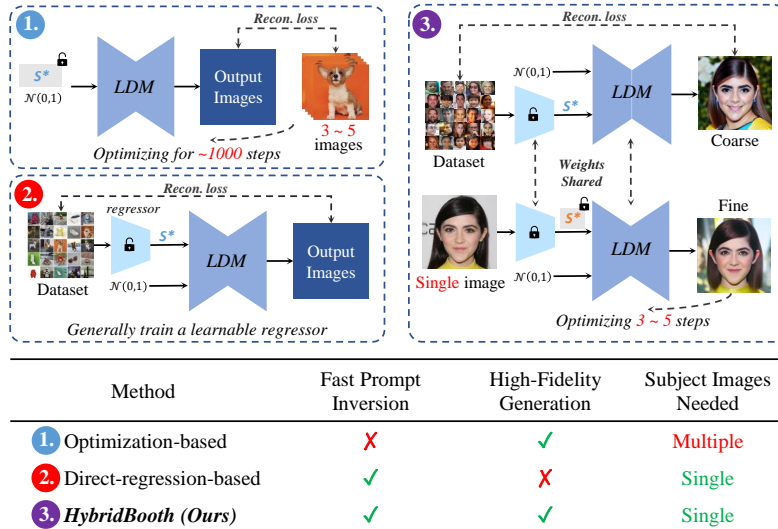


Fig. 2: Comparing with previous frameworks, our method leverages a coarse-to-fine strategy, effectively integrating the advantages of both optimization- and regression-based methods. Consequently, we enable a *fast prompt inversion* with *high-fidelity generation* from just a *single image*.

- We craft a Regressor-based Word Embedding Probe module to ensure a robust feature initialization while maintaining a focus on general-oriented generation model.
- We introduce a residual refinement strategy to accelerate the fitting process while preventing the collapse of the prior learned during the probe stage.

2 Related Work

2.1 Text-to-image Generation

Text-to-image generation focuses on creating visually accurate and contextually relevant images from textual descriptions. Early approaches, such as Generative Adversarial Networks (GANs) [11, 12, 15, 51], achieved high-quality synthetic images. However, these studies primarily emphasize image generation within specific domains, leading to limitations in extended applications. The progress of large language models [6, 8, 31, 42, 43] has enabled methods based on variational auto-encoders [17, 26] and diffusion models [32, 34, 38] to effectively map word sequences into latent space, guiding image generation from textual prompts. Our method leverages pre-trained text-to-image diffusion models to generate personalized visual content efficiently and effectively.

2.2 Subject-driven Generation

Subject-driven generation is a type of controllable generation task that aims to generate instances of specific subjects in new contexts (*e.g.*, styles, scenarios). Current approaches can be broadly categorized into two types: 1) optimization-based and 2) direct-regression-based methods.

Optimization-based Methods iteratively refine images to align with given textual descriptions. Inversion is a common technique to preserve a subject while modifying its context. For GANs, Pivot Tuning [33] fine-tunes the model using an inverted latent code for robust, ID-preserving generation. Recent work [29] keeps GAN-generated subjects consistent and controllable. However, these methods often struggle with unique subjects and fail to preserve all details. In diffusion models, inversion aims to determine the initial noise that generates the target image through denoising. Methods like DreamBooth and Textual Inversion [9, 19, 36, 37] personalize pre-trained models to generate specific subjects in various contexts but are computationally intensive and slow to converge. Our method uses a hybrid framework for robust initial feature encoding, followed by rapid optimization, achieving superior efficiency.

Direct-regression-based Methods generate subject-specific images directly from textual prompts without iterative optimization, relying on extensive model modifications on large datasets. These methods generate images from scratch based on learned text-visual correlations. Previous GAN-based methods [1, 7, 39, 40, 44] retrain generators or add conditioned constraints for ID-preserving generation but are limited by their training sets. Recent diffusion-based methods [2, 3, 10, 41, 47] tune only the encoder of pre-trained models, offering efficiency but risking overfitting and reduced performance on diverse tasks. To address these limitations, HybridBooth initially refines the encoder and employs a subject-driven residual refinement module. This process requires only 3 to 5 steps, resulting in more detailed and robust images compared to other direct-regression-based methods.

3 Preliminaries

3.1 Text-to-Image Latent Diffusion Models

We implemented our method based on text-to-image Latent Diffusion Models (LDMs) [34] which can understand prompts and generate vivid images. LDMs mainly consist of three components, including a variational auto-encoder (VAE) [26], a UNet, and a CLIP text encoder [31]. The VAE encoder \mathcal{E} compresses \mathbf{x} into a low-dimensional latent code $\mathbf{z} = \mathcal{E}(\mathbf{x})$, and the decoder \mathcal{D} maps the latent code \mathbf{z} back to image space $\hat{\mathbf{x}} = \mathcal{D}(\mathbf{z})$. For the text-guided diffusion process, the latent code \mathbf{z} is perturbed by Gaussian noise ϵ . Then the UNet \mathcal{M}_θ learns to predict the applied noise \mathbf{z}_t with a specified time step t . The denoising

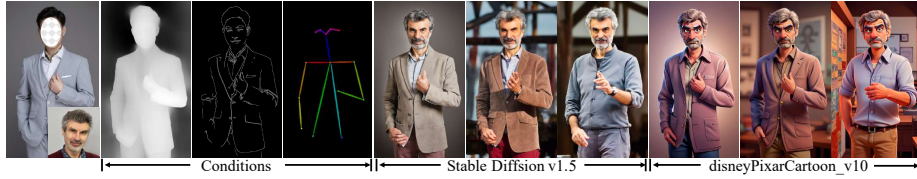


Fig. 3: Showcasing our prompt inversion framework incorporating various conditions (depth, canny, keypoints) and community models¹.

process is conditioned on the text feature \mathbf{c} extracted from text prompts \mathbf{y} by using text encoder τ_θ . The training loss is formulated as:

$$\mathcal{L}_\epsilon = \mathbb{E}_{\mathbf{z} \sim \mathcal{E}(\mathbf{x}), \epsilon \sim \mathcal{N}(0,1), \mathbf{c}, t} \left[\|\epsilon - \mathcal{M}_\theta(\mathbf{z}_t, \mathbf{c}, t)\|_2^2 \right]. \quad (1)$$

During inference, given a prompt \mathbf{y} , the model iteratively denoises \mathbf{z}_t conditioned on the text feature $\mathbf{c} = \tau_\theta(\mathbf{y})$. Finally, \mathbf{z}_0 is decoded to an image $\hat{\mathbf{x}} = \mathcal{D}(\mathbf{z}_0)$.

3.2 Prompt Inversion for Subject-Driven Generation

Despite the remarkable general generation capability of LDMs, challenges remain in generating specific subjects across various contexts, known as Subject-Driven Generation [9,36]. One feasible approach is to learn textual concepts from subject-specific images, referred to as Prompt Inversion [3,10,21]. Given several images \mathbf{X} of a subject, the prompt inversion aims to inverse the visual content of the subject to the embedding of words \mathbf{e} :

$$\mathbf{e} \leftarrow \mathcal{I}(\mathbf{X}), \quad (2)$$

where \mathcal{I} represents the inversion process. The primary goal of the inversion process is to ensure that $\mathbf{e} \in \mathbb{R}^{n \times d}$ captures the key identifying characteristics of the subject. The word embedding \mathbf{e} is then encoded into textual features \mathbf{c} using a pre-trained text encoder (*i.e.*, CLIP). Once \mathbf{e} is learned, it can be readily combined with arbitrary text for the creative generation of the subject. Specifically, let ‘S*’ be the text word corresponding to the learned embedding \mathbf{e} , we can combine ‘S*’ with new prompts to realize arbitrary subject-driven image synthesis, *e.g.* ‘S* in snow forest’.

Discussion. Another paradigm involves mapping subject images to image tokens and using an additional cross-attention layer to inject these tokens into the diffusion process. This paradigm significantly alters the generation process of the original Latent Diffusion Models (LDMs), necessitating retraining on millions of images. In contrast, prompt inversion is a more resource-efficient method for subject-driven generation. As shown in Fig. 3, it can seamlessly integrate with various community foundational models and controllers pre-trained on the original stable diffusion, making it a versatile and practical alternative.

¹ disneyPixarCartoon_v10: civitai.com/models/65203/disney-pixar-cartoon-type-a

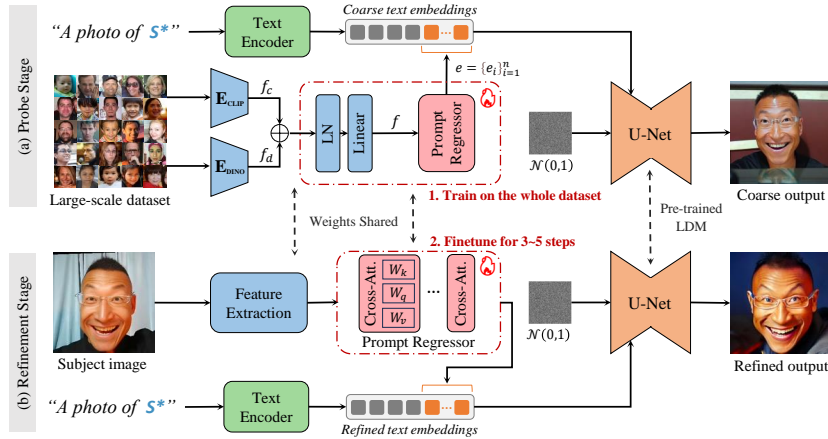


Fig. 4: The framework of HybridBooth consists of two stages: (a) The first *Probe Stage* leverages a prompt regressor estimate the initial word embedding, acting as a probe for a rational word embedding; (b) The second *Refinement Stage* adapts the regressor to the subject by taking only 3-5 iterations to achieve efficient subject generation.

4 Hybrid Prompt Inversion Framework

Given just *one* image x of a specific subject, we aim to efficiently embed the subject into the word embedding space. We combine the strengths of optimization- and direct-regression-based paradigms to achieve precise and efficient subject-driven generation. Thus, we propose a hybrid prompt inversion framework, termed HybridBooth. As shown in Fig. 4, HybridBooth consists of two stages: (1) **Word Embedding Probe Stage**, which learns a domain-agnostic encoder providing an initial word embedding estimate. (2) **Word Embedding Refinement Stage**, which adapts the encoder to the given image of the subject. The key idea is: *A powerful encoder reduces iterative refinement costs, while effective iterative refinement reduces the encoder’s accuracy demands.* Our challenge is to develop a robust encoder for initial embeddings and an effective refinement scheme to reduce the encoder’s precision requirements. Next, we introduce the Word Embedding Probe stage in Sec. 4.1, describe the residual refinement stage in Sec. 4.2, and present the framework details in Sec. 4.3.

4.1 Regressor-based Word Embedding Probe

First, HybridBooth trains a prompt regressor that takes an image of a subject as input and estimates the initial word embedding. This involves extracting multi-grained image features and conducting multiple-word regression.

Multi-grained Image Feature Merging. Existing regressor-based methods exclusively utilize CLIP features f_c , which often contain global semantic information (*e.g.*, subject’s category) [13, 31, 36]. Hence, we incorporate DINOv2 [27]

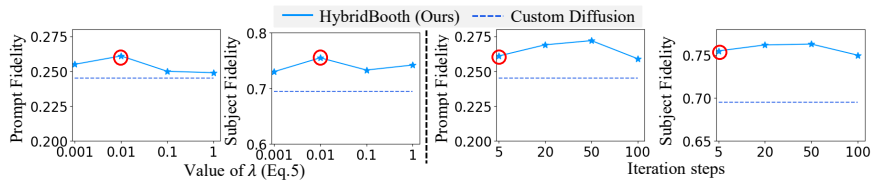


Fig. 5: Impact of hyper-parameters in the residual refinement stage (on the Dream-Booth dataset [36]). Default parameters are marked with red circles. Custom Diffusion [19] serves as the reference. Prompt Fidelity and Subject Fidelity are CLIP-I and DINO-T, respectively (see Sec. 5.1).

features \mathbf{f}_d as a supplement, offering detailed pixel-level information. Taking inspiration from previous practices [13, 48], we employ a Multilayer Perceptron (MLP) to project DINOv2 features \mathbf{f}_d , then concatenate the output with CLIP feature \mathbf{f}_c . Finally we align the feature dimension using a Linear layer [35]. The merging process is formulated as:

$$\mathbf{f} = \text{Linear}([\mathbf{f}_c, \text{MLP}(\mathbf{f}_d)]), \quad (3)$$

where $[\cdot, \cdot]$ represents the concatenation operation, and the structure of MLP is `LyaerNorm-Linear-GELU` [49]-`Linear`. As verified in our experiments, this simple module yields significant improvements.

Multiple-word Regression. The merged feature \mathbf{f} is then mapped to word embeddings \mathbf{e} . Typically, a single word represents a specific concept, *e.g.* ‘man’ conveys the subject’s category but may lack details like hairstyle. Thus, we suggest projecting the feature \mathbf{f} to multiple-word embeddings to provide a more thorough description of the subject:

$$\mathbf{e} = \mathcal{R}(\mathbf{f}), \quad (4)$$

where $\mathbf{e} = \{\mathbf{e}_i\}_{i=1}^n$, n is the number of word embedding $\mathbf{e}_i \in \mathbb{R}^d$, and d is the dimension of word embedding.

4.2 Residual Refinement

Although the regressor-based probe is fast, precise alignment between word embeddings and key subject features remains challenging without image feature feedback. Thus, our subsequent stage focuses on refinement, aiming to adapt the learned regressor to target subject for better alignment. Striking a balance is challenging due to the *single* image input, as it requires avoiding overfitting to a specific image while capturing essential subject characteristics effectively.

To tackle this issue, we enhance the learned regressor with a residual refinement strategy that adjusts only critical parameters (\mathbf{W}_ϕ) for faster adaptation while preserving the learned prior. The residual refinement is formulated as:

$$\mathbf{W}'_\phi = \mathbf{W}_\phi + \lambda \Delta \mathbf{W}_\phi, \quad (5)$$

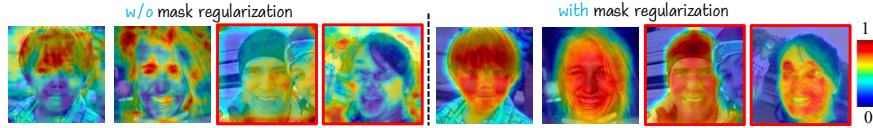


Fig. 6: Showcases of the cross-attention map for the subject word ‘S*’. Applying mask regularization \mathcal{L}_M prevents attention leakage to irrelevant areas, even with other subjects present (highlighted in red boxes).

where $\Delta\mathbf{W}_\phi$ is the learned residual parameters and a hyperparameter λ is to control the scale of $\Delta\mathbf{W}_\phi$. With residual refinement, it only takes 3-5 iterations to achieve high-fidelity subject generation while preserving editability.

Discussing the overfitting problem. Overfitting is a common challenge in finetune-based methods, especially when tuning the prompt regressor on a single image. To mitigate this, we refine only *partial* parameters (see Eq. (5)). Intuitively understanding, \mathbf{W}_ϕ acts as an anchor, preserving the learned prior and stabilizing updates. Zeroing out \mathbf{W}_ϕ collapses its prior (see the last column in Fig. 10), and we need to adjust hyper-parameters more carefully. As shown in Fig. 5, by using residual refinement, λ and iteration steps can vary widely yet still yield good results. Fig. 7 and Fig. 9 demonstrate that HybridBooth can generate novel subject images with diverse expressions, postures, styles, and scenarios.

4.3 Framework Details

Prompt Regressor Design. Previous studies have used either an MLP or a UNet as the prompt regressor. HybridBooth is flexible and can integrate any network mapping image features \mathbf{f} to word embeddings \mathbf{e} . We employ the state-of-the-art regressor PromptNet [50], which uses a block structure similar to LDM and can be initialized with LDM pretrained on large-scale image datasets. In the refinement stage, we focus on refining the cross-attention layer parameters (key, query, and value matrices $\mathbf{W}_k, \mathbf{W}_q, \mathbf{K}_v$) based on layer importance experiments. Fine-tuning the entire regressor with 100 images, we calculated importance scores for each layer based on the mean cosine distance of parameters before and after refinement: cross-attention (56.3), self-attention (43.9), and other layers (*e.g.* convolution) (12.4). The cross-attention layer had the highest importance. Similar experiments by Custom Diffusion [20] and E4T [10] focused on the denoising UNet within LDMs.

Training with Mask Regularization. The diffusion loss \mathcal{L}_ϵ (Eq. (1)) is commonly used for training the prompt regressor. However, as shown in Fig. 6 (Left), the cross-attention maps of the subject’s word embedding often leak into the irrelevant background. To address this, we use the subject’s mask to constrain the cross-attention maps of subject words, ensuring focus on the subject region. Let $\mathbf{A}_{\mathbf{e}_i} \in \mathbb{R}^{h \times w}$ be the attention map for \mathbf{e}_i , and $\mathbf{M} \in \mathbb{R}^{H \times W}$ be the segmentation

mask for the subject. We use a mask regularization loss to align \mathbf{A}_{e_i} with \mathbf{M} :

$$\mathcal{L}_M = \frac{1}{n} \sum_{i=1}^n \text{mean}(\mathbf{A}_{e_i} \cdot (\mathbf{1} - \mathbf{M})) - \text{mean}(\mathbf{A}_{e_i} \cdot \mathbf{M}), \quad (6)$$

where n is the number of word embeddings. This loss minimizes attention outside the mask and maximizes it inside, ensuring focus on the subject. The interpolation of \mathbf{M} to match \mathbf{A}_{e_i} is omitted for simplicity. The final training objective of the regressor is:

$$\mathcal{L} = \mathcal{L}_\epsilon + \alpha_M \mathcal{L}_M, \quad (7)$$

where α_M is the weight of \mathcal{L}_M . During refinement, \mathcal{L} is used to prevent diverging update directions, focusing on fine-tuning the cross-attention parameters.

5 Experiments

5.1 Experimental Setup

Datasets. We mainly conduct experiments on facial images. Following prior research practices, we train HybridBooth on the aligned FFHQ dataset [14], comprising 70,000 in-the-wild facial images. Constrained by computational resources, we conducted testing on the first 1,000 images from the CelebA-HQ test split. In addition, we also evaluate HybridBooth on other categories to demonstrate its versatility. With the same setting of ELITE [47], we train HybridBooth on the test split of OpenImagesv6 [18], and test on the DreamBooth Dataset. All images are cropped and resized to 512×512 pixels.

Training Details. We use Stable Diffusion v1.5 [34] as our Text-to-Image model, initializing the prompt regressor with the UNet from Stable Diffusion. The DINOv2-Large model [27] extracts the DINO feature, and the vision model of clip-vit-large-patch14 [31] extracts the CLIP image feature. The segmentation mask \mathbf{M} (Eq. (6)) is obtained using InSPyReNet [16]. We use the same hyperparameters for both facial and non-facial experiments. For the word embedding probe stage, we use the AdamW optimizer [23] with a learning rate of $2e-5$ and a batch size of 8. Training on a single Nvidia A100 GPU takes 40 hours. In the refinement stage, we fine-tune the cross-attention parameters of the regressor for 5 steps using AdamW, with a learning rate of $2e-5$ and a weight decay of $1e-2$. Parameters α_M (Eq. (7)) and λ (Eq. (5)) are set to $1e-3$ and $1e-2$, respectively.

Baselines. We compare the proposed HybridBooth with two categories of methods: (1) Optimization-based methods: DreamBooth [36], Textual Inversion [9], and Custom Diffusion [19]. (2) Direct-regression-based methods: ELITE [47] and FastComposer [48]. Optimization-based methods typically require 3~5 images as input, which are not available in CelebA-HQ (our test set). To ensure a fair

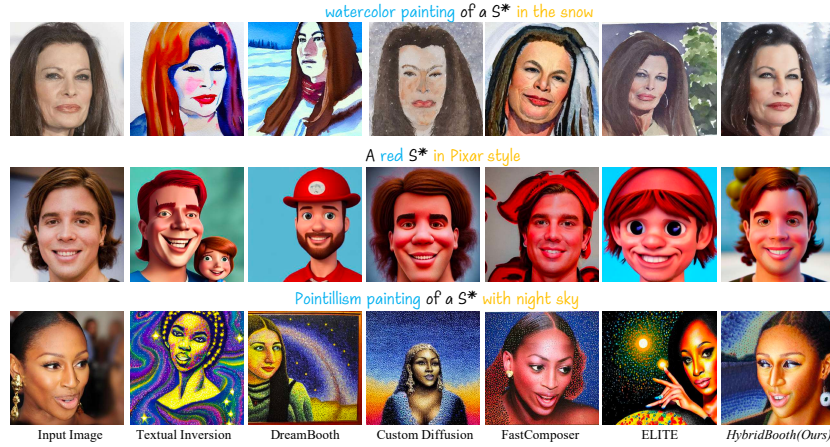


Fig. 7: Qualitative Comparison Results on CelebA-HQ (facial dataset). Columns 2-4: Optimization-based methods. Columns 5-6: Direct-regression-based methods.

comparison, we augment each image to generate 5 images for these methods using affine transformations and background changes. More details on the augmentation process are provided in the Supplementary. To clarify the difference between HybridBooth and HyperDreamBooth [37], HyperDreamBooth updates weights in a low-rank space for rapid model tuning but overlooks the importance of training a powerful encoder initially. Since the official code for HyperDreamBooth was unavailable during our submission, we could only make qualitative comparisons with examples from their paper (see Fig. 8).

Evaluation Metrics. We evaluate the model’s performance in two aspects:

- **Prompt Fidelity:** This measures how well the generated images match the textual prompt, quantified by the average cosine similarity of CLIP embeddings between the prompt and the generated images (CLIP-T). *Higher CLIP-T indicates better textual editability.*
- **Subject Fidelity:** This assesses the preservation of subject details during generation. We use CLIP-I and DINO-I metrics, which are the cosine similarities between the embeddings of the generated images and the original concept images. *Higher values indicate better subject fidelity.*

5.2 Qualitative Comparison

In Fig. 7, we present qualitative comparison results. Similar to previous works, ‘S*’ denotes the word embedding learned from the input subject image. For a fair comparison, optimization-based methods use 5 images for optimization. Our setting is more complex, specifying both artistic style (e.g., ‘watercolor painting’) and environmental characteristics (e.g., ‘in the snow’). From the results, two key

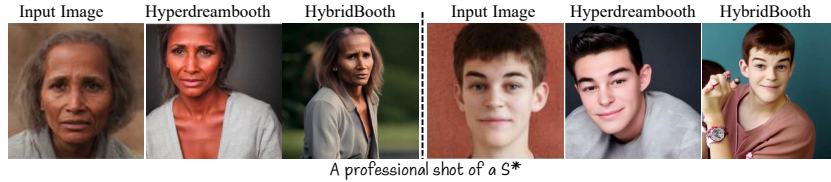


Fig. 8: Qualitative Comparison with Hyperdreambooth [37].



Fig. 9: Qualitative Comparison Results on non-facial images.

observations emerge: (1) Optimization-based methods (Columns 2-4), using multiple image inputs and finely tuned hyper-parameters (e.g., total iteration steps), generally outperform direct-regression-based methods (Columns 5-6) in prompt alignment and subject similarity. (2) Our proposed HybridBooth outperforms all compared methods in text alignment and subject similarity. For example, in the third row, other methods either fail in text similarity (e.g., FastComposer) or in preserving facial similarity, whereas HybridBooth successfully aligns the subject with the editing prompt, demonstrating its superiority.

Qualitative comparisons with HyperDreamBooth [37] are shown in Fig. 8. HyperDreamBooth trains a hyper-network to predict low-rank weights of the image diffusion model and fine-tunes these weights on one subject image. However, low-rank weight updates limit personalized representation, resulting in poorer subject feature preservation, as seen in Fig. 8. This supports our core idea that a more powerful encoder reduces the cost of iterative refinement.

As shown in Fig. 9, we evaluate HybridBooth on non-facial images, including animals (Line 1) and complex buildings (Line 2). FastComposer [48] is excluded as it is designed for facial images. In Line 1, most methods preserve the texture of the subject well. However, in Line 2, only HybridBooth maintains the structural information while adhering to the complex prompt. This is due to the hybrid framework, which enables quick feature learning while retaining the prompt understanding capability of the Stable Diffusion model.

Table 1: Quantitative Evaluation on the CelebA-HQ. **First Block:** Optimization-based methods. **Second Block:** Direct-regression-based methods. **Third Block:** HybridBooth and its ablations.

Method	CelebA-HQ			DreamBooth-dataset			Iter. Step [↓]
	CLIP-T [↑]	CLIP-I [↑]	DINO-I [↑]	CLIP-T [↑]	CLIP-I [↑]	DINO-I [↑]	
Textual Inversion [9]	0.164	0.612	0.236	0.183	0.663	0.462	5000
DreamBooth [36]	0.251	0.564	0.376	0.251	0.785	0.674	1000
Custom Diffusion [19]	0.237	0.675	0.398	0.245	0.801	0.695	200
ELITE [47]	0.169	0.592	0.311	<u>0.255</u>	0.762	0.652	1
FastComposer [48]	0.201	0.782	<u>0.581</u>	-	-	-	1
HybridBooth	<u>0.246</u>	0.865	0.644	0.261	0.865	0.755	5
- w/o Refinement	0.177	<u>0.842</u>	0.568	0.192	0.835	0.645	-
- w/o Probe	0.153	0.408	0.068	0.142	0.412	0.050	5
- w/o DINO Feature	0.161	0.837	0.453	0.184	<u>0.841</u>	0.596	5
- w/o CLIP Feature	0.182	0.734	0.510	0.206	0.752	0.630	5
- w/o Mask Regularization	0.203	0.831	0.625	0.233	0.833	<u>0.705</u>	5



Fig. 10: We conducted ablations on HybridBooth to evaluate its influence on learning subject-specific word embeddings. DreamBooth and FastComposer are utilized here for comparison. ‘A photo of S*’ is used for generating images.

5.3 Quantitative Comparison

The quantitative results in Tab. 1 include the necessary iteration steps for assessing time efficiency. Optimization-based methods generally outperform direct-regression-based methods but require more iteration steps. Our HybridBooth significantly exceeds other methods in both prompt fidelity and subject fidelity. Notably, our DINO-T metric is over 10% higher than FastComposer, a key metric in the DreamBooth framework [36].

5.4 Ablation Analysis

Ablations on the Hybrid Framework. We evaluate our hybrid framework, which includes a word embedding stage for initial estimation and a residual refinement stage for precise alignment. Ablations, termed ‘w/o Probe’ and ‘w/o Refinement’, are shown in Fig. 10 and Tab. 1. The prompt regressor is initialized using the pre-trained UNet from Stable Diffusion in both settings. We assess subject fidelity using the prompt ‘A photo of S*’, comparing it with DreamBooth

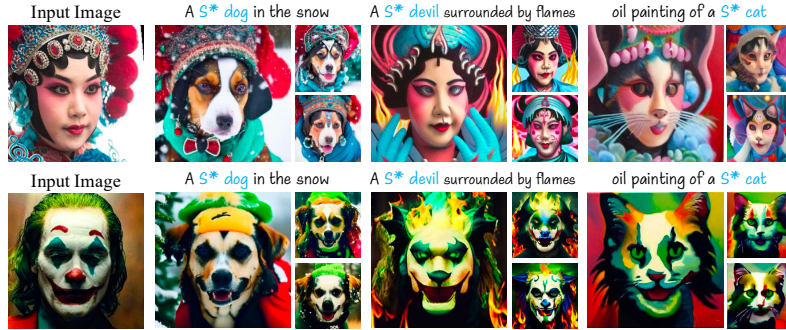


Fig. 11: The learned words from human beings can be applied to cross-species subjects, which verifies that HybridBooth manages to learn intrinsic subject features.

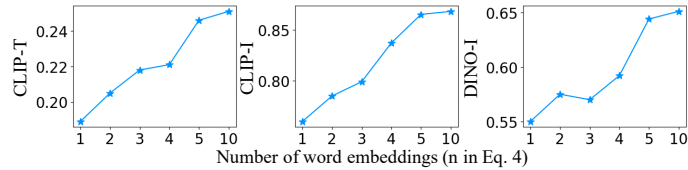


Fig. 12: Impact of increasing number of word embeddings (on CelebA-HQ).

and FastComposer. As shown in Fig. 10, subject embedding from residual refinement alone leads to lower subject similarity. Without refinement, word embedding only crudely captures key features. Our HybridBooth framework, integrating both probe and refinement stages, outperforms all ablations and compared methods, effectively restoring subject characteristics.

Ablations on Multi-grained Image Feature Merging. Previous works [47, 48] typically use CLIP features as input for the regressor. As discussed in Sec. 4.1, we use CLIP features for global semantic information and DINO features for detailed pixel-level information. We ablate the feature extraction strategy to verify this approach. As shown in Tab. 1, relying solely on DINO or CLIP features results in suboptimal performance. However, integrating both features significantly enhances performance. Notable improvements in prompt fidelity and subject fidelity are observed, even with just a single layer normalization and linear layer.

Analysis on the Mask Regularization Loss \mathcal{L}_M . We propose a mask regularization loss \mathcal{L}_M (Eq. (6)) to ensure that the estimated word embeddings align with the subject area. As shown in the last line of Tab. 1, both subject fidelity and prompt fidelity decrease significantly without the mask regularization loss.

Ablations on the Number of Word Embedding. Recalling that the prompt regressor estimates multiple word embeddings for enhanced description of key



Fig. 13: Showcases of limitations in HybridBooth.

subject features, we adjust the number of embeddings. Quantitative results in Fig. 12 indicate a consistent improvement in metrics related to Prompt Fidelity and Subject Fidelity with an increasing number of words. To balance efficiency and computation cost, we set the number of word embeddings we estimate to 5.

Can the Learned Subject Concept be Cross-species Adapted? The well-aligned training data in the *Probe Stage* results in the encoder developing a consistent semantic tendency in attention maps, as shown in Fig. 6. The subsequent *Refinement Stage* further enhances the model’s generalization ability. Fig. 11 illustrates that species with similar semantic structures, such as humans, dogs, and cats, consistently yield robust results, even with complex prompts.

6 Conclusion

In this paper, we address the challenge of achieving fast prompt inversion while preserving the generalization capabilities of text-to-image models. By leveraging both optimization-based and direct-regression-based methods, we propose a two-stage hybrid prompt inversion framework, termed HybridBooth. This framework ensures robust and efficient subject-driven generation while maintaining the model’s general creative ability. Experimental results demonstrate the superior performance of HybridBooth compared to existing state-of-the-art methods.

Limitation. As shown in Fig. 13, HybridBooth cannot perform precise semantic editing, such as adjusting expressions and age. In addition, HybridBooth inherits the weakness from StableDiffusion that fails to generate fine-grained details of the subject, such as fingers. To address these issues, we believe possible solutions are incorporating language-and-vision models [22] for improved understanding and planning, and utilizing more 3D structural information [4, 5].

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62073244 and 61825303, 62088101, and the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0100) and the Fundamental Research Funds for the Central Universities. We thank Yixuan Li for his assistance with the illustrations.

References

1. Alaluf, Y., Patashnik, O., Cohen-Or, D.: Only a matter of style: Age transformation using a style-based regression model. *ACM TOG* **40**(4) (2021) [2](#), [4](#)
2. Alaluf, Y., Richardson, E., Metzger, G., Cohen-Or, D.: A neural space-time representation for text-to-image personalization (2023) [2](#), [4](#)
3. Arar, M., Gal, R., Atzmon, Y., Chechik, G., Cohen-Or, D., Shamir, A., Bermano, A.H.: Domain-agnostic tuning-encoder for fast personalization of text-to-image models (2023) [2](#), [4](#), [5](#)
4. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Wagnerspack, W.N. (ed.) *SIGGRAPH*. pp. 187–194 (1999) [14](#)
5. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: *ECCV*. pp. 561–578. Springer (2016) [14](#)
6. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020) [3](#)
7. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: *CVPR* (2018) [2](#), [4](#)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: *North American Chapter of the Association for Computational Linguistics* (2019) [3](#)
9. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: *ICLR* (2022) [1](#), [2](#), [4](#), [5](#), [9](#), [12](#)
10. Gal, R., Arar, M., Atzmon, Y., Bermano, A.H., Chechik, G., Cohen-Or, D.: Encoder-based domain tuning for fast personalization of text-to-image models. *ACM TOG* **42**(4), 1–13 (2023) [1](#), [2](#), [4](#), [5](#), [8](#)
11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: *NeurIPS*. pp. 2672–2680 (2014) [3](#)
12. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: *CVPR* (2017) [3](#)
13. Jiang, D., Liu, Y., Liu, S., Zhang, X., Li, J., Xiong, H., Tian, Q.: From clip to dino: Visual encoders shout in multi-modal large language models. *CoRR*:2310.08825 (2023) [6](#), [7](#)
14. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: *CVPR*. pp. 4401–4410 (2019) [2](#), [9](#)
15. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: *CVPR*. pp. 4401–4410 (2019) [3](#)
16. Kim, T., Kim, K., Lee, J., Cha, D., Lee, J., Kim, D.: Revisiting image pyramid structure for high resolution salient object detection. In: *Proceedings of the Asian Conference on Computer Vision*. pp. 108–124 (2022) [9](#)
17. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: *ICLR* (2014) [3](#)
18. Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Veit, A., et al.: Openimages: A public dataset for

- large-scale multi-label and multi-class image classification. Dataset available from <https://github.com/openimages> **2(3)**, 18 (2017) **9**
19. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: CVPR (2023) **2, 4, 7, 9, 12**
 20. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: CVPR. pp. 1931–1941 (2023) **8**
 21. Li, D., Li, J., Hoi, S.C.H.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In: NeurIPS (2023) **5**
 22. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: CVPR. pp. 26296–26306 (2024) **14**
 23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2018) **9**
 24. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. CoRR:2211.09794 (2022) **1**
 25. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML. pp. 8162–8171 (2021) **1**
 26. van den Oord, A., Vinyals, O., kavukcuoglu, k.: Neural discrete representation learning. In: NeurIPS. p. 6306–6315 (2017) **3, 4**
 27. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. CoRR:2304.07193 (2023) **6, 9**
 28. Pan, X., Qin, P., Li, Y., Xue, H., Chen, W.: Synthesizing coherent story with auto-regressive latent diffusion models (2022) **1**
 29. Pan, X., Tewari, A., Leimkühler, T., Liu, L., Meka, A., Theobalt, C.: Drag your gan: Interactive point-based manipulation on the generative image manifold. In: SIGGRAPH. pp. 1–11 (2023) **4**
 30. Peng, X., Zhu, J., Jiang, B., Tai, Y., Luo, D., Zhang, J., Lin, W., Jin, T., Wang, C., Ji, R.: Portraitbooth: A versatile portrait model for fast identity-preserved personalization. In: CVPR. pp. 27080–27090 (2024) **1**
 31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021) **3, 4, 6, 9**
 32. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML (2021) **3**
 33. Roich, D., Mokady, R., Bermano, A.H., Cohen-Or, D.: Pivotal tuning for latent-based editing of real images. ACM TOG (2021) **4**
 34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022) **1, 3, 4, 9**
 35. Rosenblatt, F.: The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review **65(6)**, 386 (1958) **7**
 36. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR. pp. 22500–22510 (2023) **1, 2, 4, 5, 6, 7, 9, 12**
 37. Ruiz, N., Li, Y., Jampani, V., Wei, W., Hou, T., Pritch, Y., Wadhwa, N., Rubinstein, M., Aberman, K.: Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In: CVPR. pp. 6527–6536 (2024) **2, 4, 10, 11**
 38. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, S.K.S., Lopes, R.G., Ayan, B.K., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS (2022) **3**

39. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of gans for semantic face editing. In: CVPR (2020) [2](#), [4](#)
40. Shen, Y., Yang, C., Tang, X., Zhou, B.: Interfacegan: Interpreting the disentangled face representation learned by gans. TPAMI (2020) [2](#), [4](#)
41. Tewel, Y., Gal, R., Chechik, G., Atzmon, Y.: Key-locked rank one editing for text-to-image personalization. SIGGRAPH (2023) [2](#), [4](#)
42. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models (2023) [3](#)
43. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models (2023) [3](#)
44. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for stylegan image manipulation. CoRR:2102.02766 (2021) [2](#), [4](#)
45. Voetman, R., Aghaei, M., Dijkstra, K.: The big data myth: Using diffusion models for dataset generation to train deep detection models (2023) [1](#)
46. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A.: Instantid: Zero-shot identity-preserving generation in seconds. CoRR preprint CoRR:2401.07519 (2024) [1](#)
47. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In: ICCV. pp. 15943–15953 (2023) [2](#), [4](#), [9](#), [12](#), [13](#)
48. Xiao, G., Yin, T., Freeman, W.T., Durand, F., Han, S.: Fastcomposer: Tuning-free multi-subject image generation with localized attention. CoRR:2305.10431 (2023) [2](#), [7](#), [9](#), [11](#), [12](#), [13](#)
49. Zhang, Q., Wang, C., Wu, H., Xin, C., Phuong, T.V.: Gelu-net: A globally encrypted, locally unencrypted deep neural network for privacy-preserved learning. In: IJCAI. pp. 3933–3939 (2018) [7](#)
50. Zhou, Y., Zhang, R., Sun, T., Xu, J.: Enhancing detail preservation for customized text-to-image generation: A regularization-free approach. CoRR:2305.13579 (2023) [8](#)
51. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017) [3](#)