Shiyuan YANG, Zheng GU, Wenyue HAO, Yi WANG, Huaiyu CAI, Xiaodong CHEN, 2025. Few-shot exemplar-driven inpainting with parameter-efficient diffusion fine-tuning. *Frontiers of Information Technology & Electronic Engineering*, 26(8):1428-1440. https://doi.org/10.1631/FITEE.2400395

Few-shot exemplar-driven inpainting with parameter-efficient diffusion fine-tuning

Key words: Diffusion model; Image inpainting; Exemplar-driven; Few-shot fine-tuning

Corresponding author: Xiaodong CHEN
E-mail: xdchen@tju.edu.cn
ORCID: https://orcid.org/0000-0003-1624-2680

Motivation

- Text-to-image diffusion models excel at inpainting but struggle to replicate specific objects with high fidelity from text prompts alone. As the saying goes, “a picture is worth a thousand words.”
- Users often need to inpaint a customized object using a reference (exemplar) image, but existing methods lack high fidelity.

Fig. 1 Text-driven inpainting (top) struggles to accurately describe the object’s details, while exemplar-driven inpainting (bottom) can make it easier. References to color refer to the online version of this figure

Existing Limitations

- **Textual Inversion (TxtInv)** learns a textual embedding from exemplars, but this contains limited parameters and is less expressive. It often produces boundary artifacts from its background blending technique.

- **Paint by Example (PbE)** relies on large-scale datasets, which cannot cover all personalized exemplars and fails to produce high-fidelity results for out-of-dataset objects. Training is also labor-intensive.

Main Idea

To achieve high-fidelity, customized inpainting, this work enhances a pretrained text-driven inpainting model with a lightweight, plug-and-play module that learns a specific object's appearance from just a few examples. To achieve this, we introduce:

- **Lightweight Fine-tuning**, which employs a plug-and-play low-rank adaptation (LoRA) module to efficiently learn an exemplar's concept into model weights, enabling powerful customization while preserving the base model's integrity.
- **GPT-4V Prompting**, which leverages the GPT-4V model to generate expressive text prompts from the exemplar image, guiding the model to better preserve key object details.
- **Prior Noise Initialization**, which initiates the generation process from a "prior noise" containing exemplar information instead of from random noise, significantly improving the final result's fidelity.

Method

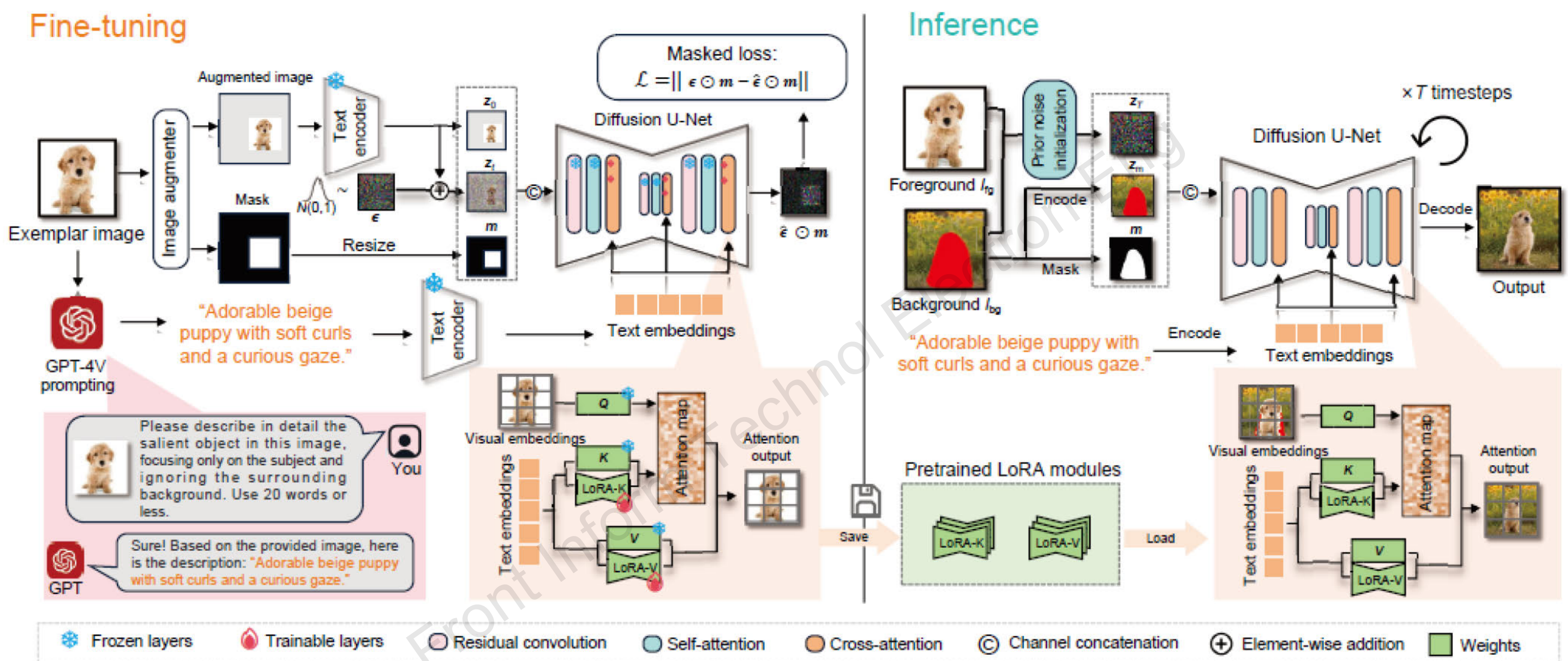


Fig. 2 Pipeline overview of our method. Based on frozen SD-inpaint model, the fine-tuning stage (left) involves fine-tuning learnable LoRA modules on a given exemplar image with GPT-4V generated prompt. The inference stage (right) first loads exemplar-specific pretrained LoRA modules and then samples from the prior noise initialization to facilitate high-fidelity exemplar-driven inpainting within the masked region of the provided background image. References to color refer to the online version of this figure

Method

Two-Stage Pipeline:

Fine-tuning Stage (Few-Shot Learning):

1. A LoRA module is added to a frozen, pretrained Stable Diffusion (SD) inpainting model.
2. The LoRA module learns the specific concept from a few user-provided exemplar images.
3. A detailed text prompt for the exemplar is generated using GPT-4V to improve detail preservation.

Inference Stage:

1. The fine-tuned, exemplar-specific LoRA module is loaded.
2. The denoising process starts from a “prior noise initialization” derived from a composite image (exemplar pasted into the background) instead of random noise.
3. The model inpaints the object into the masked region, guided by the GPT-4V prompt.

Technical Details

1. Parameter-Efficient Fine-tuning with LoRA:

• The LoRA module is lightweight and is added only to the cross-attention layers of the U-Net, learning the exemplar concept directly into the model weights.
• This provides a larger parameter space and stronger fitting capability compared to just optimizing text embeddings (like TxtInv).
• Base model weights remain frozen, preserving original text-driven inpainting capabilities.

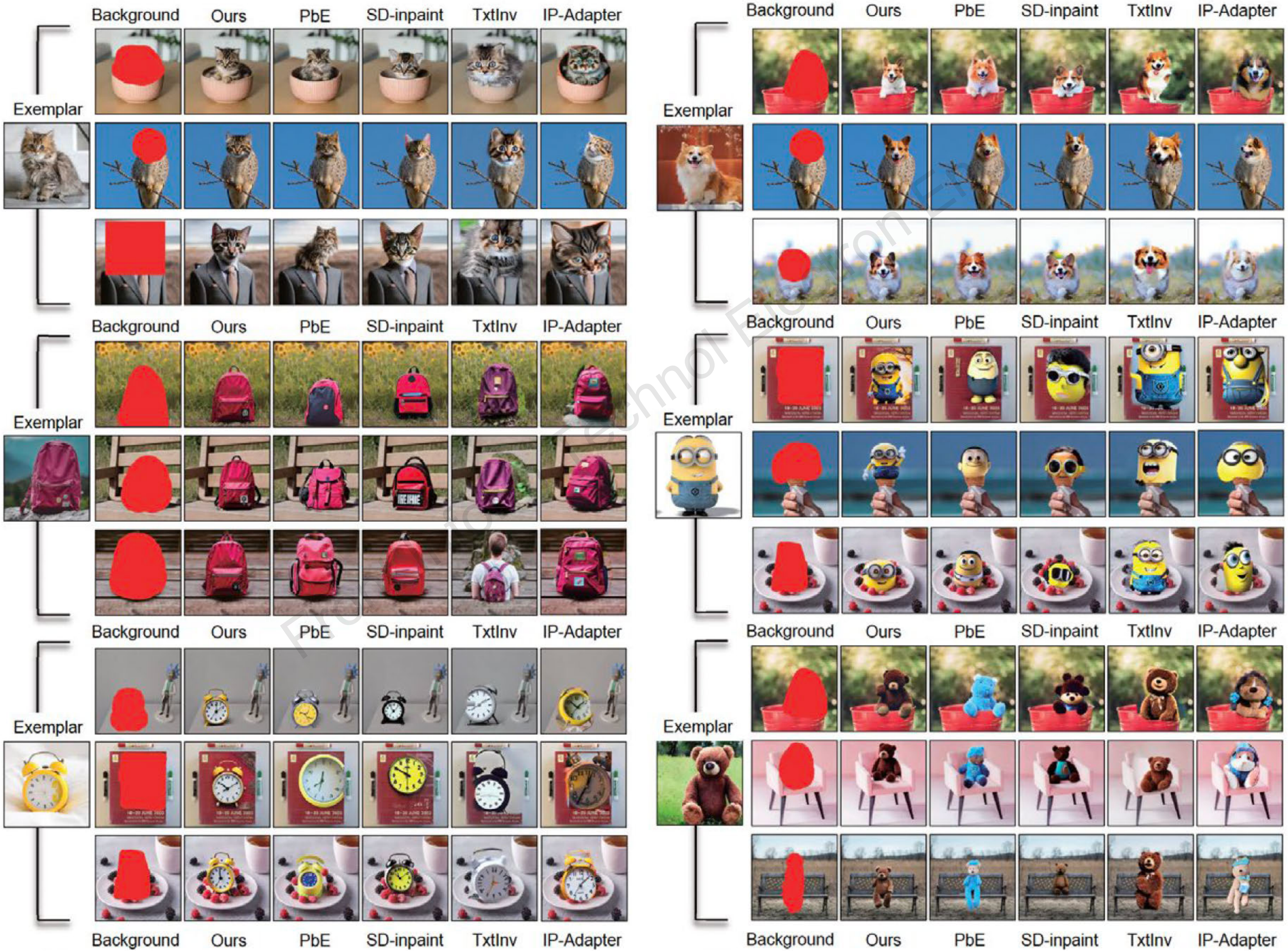
2. GPT-4V Prompting:

• Instead of simple prompts (e.g., “a photo of <x>”), we use the GPT-4V model to generate expressive, detailed descriptions of the exemplar image.
• High-quality prompts are crucial for model performance and help retain more details from the exemplar.

3. Prior Noise Initialization:

• Standard diffusion models start inference from random Gaussian noise.
• Our method applies one-step forward noising to a composite image, where the exemplar is pasted onto the masked background.
• This ensures that the initial latent input retains prior information from the exemplar, bridging the gap between training and inference and improving the fidelity.

Results



Results

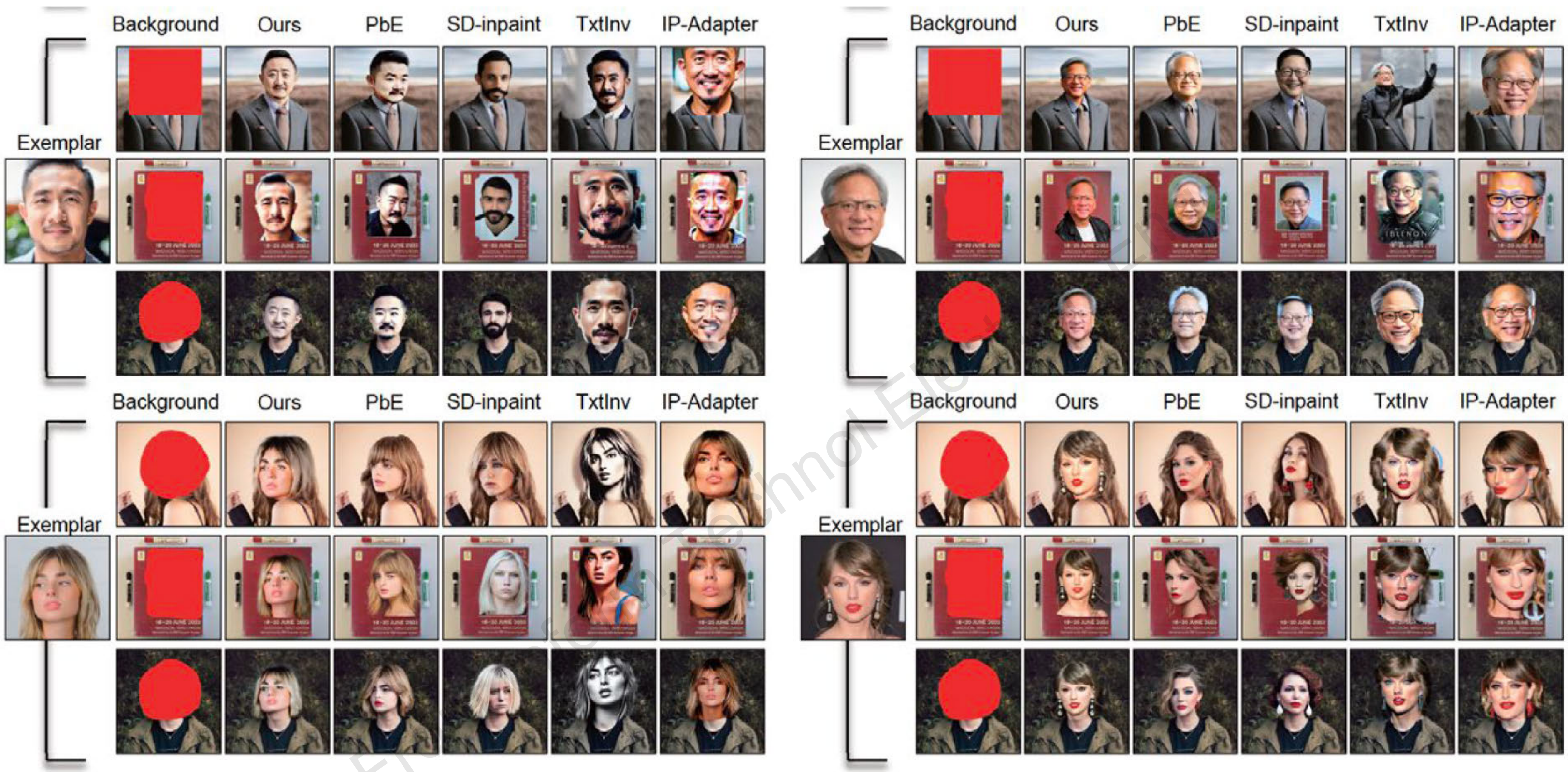


Table 1 Qualitative comparison with baselines

Method	I2I score	T2I score	FID score	Aesthetic score
SD-Inpaint (Rombach et al., 2022)	65.59	28.10	149.62	5.0836
TxtInv (Gal et al., 2022)	70.86	**29.45**	146.89	5.0765
PbE (Yang BX et al., 2023)	69.47	28.79	150.45	**5.0864**
IP-Adapter (Ye et al., 2023)	64.77	27.74	152.45	4.9558
Ours	**72.05**	29.31	**146.14**	5.0854

The best results are in bold

Ablations

GPT-4V Prompting: Using detailed GPT-4V prompts significantly improves the retention of object details.

Fig. 5 Visual comparison between using plain prompting (blue box) vs. GPT-4V prompting (orange box). References to color refer to the online version of this figure

Prior Noise Initialization: Sampling from prior noise leads to more stable and consistent outputs.

Fig. 6 Batched visual comparison between sampling with (w/) or without (w/o) prior noise initialization

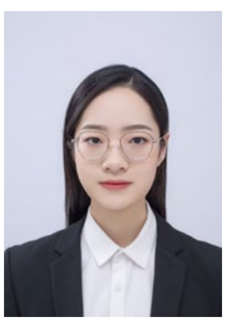
Conclusions

We presented a novel approach for few-shot exemplar-driven inpainting using a parameter-efficient fine-tuned diffusion model.

- Our method upgrades a pretrained SD-inpaint model with a lightweight LoRA module to learn customized concepts efficiently.
- We introduced GPT-4V prompting and prior noise initialization to significantly enhance the fidelity and detail preservation of the inpainting results.
- The method outperforms state-of-the-art baselines both qualitatively and quantitatively, offering a more adaptable and effective solution for customized image editing.

Shiyuan YANG received the B.S. and Ph.D. degrees in opto-electronic information science and engineering and optical engineering from Tianjin University in 2019 and 2025, respectively. Currently, he is a researcher at Tencent. His research interests include image/video generation and editing and multimodal understanding.

Zheng GU received the B.S. and Ph.D. degrees in computer science from Nanjing University in 2017 and 2024, respectively. Now he is an assistant professor at the Department of Computer Science, Shenzhen University. His research interests include few-shot learning, image generation, and computer vision.



Wenyue HAO received the B.S. and M.S. degrees in opto-electronic information science and engineering and optical engineering from Tianjin University in 2021 and 2024, respectively. Her research interests include medical image processing and computer vision.

Yi WANG received the Ph.D. degree in optical engineering from Tianjin University. She is an associate professor and currently works with the Key Laboratory of Optoelectronics Information Technology, Ministry of Education, Tianjin University. Her research interests include opto-electronic imaging, detecting, and processing.

Huaiyu CAI received the Ph.D. degree in optical engineering from Tianjin University. She is currently a professor with the School of Precision Instruments and Optoelectronic Engineering, Tianjin University. She has led the National Thirteenth Five-Year Scientific and Technological Research Projects and enterprise cooperative projects in recent years. She was also awarded the Second and the Third Prize of Tianjin Science and Technology Progress, separately. She is the author of one book and more than 80 articles. Her research interests include photoelectric imaging and detection technology, information optics, and image processing.

Xiaodong CHEN received the Ph.D. degree in optical engineering from Tianjin University. He is a professor and doctorial tutor with the School of Precision Instruments and Optoelectronic Engineering, Tianjin University. In recent years, he has undertaken or completed more than 20 projects including the National Thirteenth Five-Year Scientific and Technological Research Projects and the National 863 Project. He has published more than 270 papers in academic journals and international conferences. His research interests include photoelectric detection, medical image processing, and computer vision.