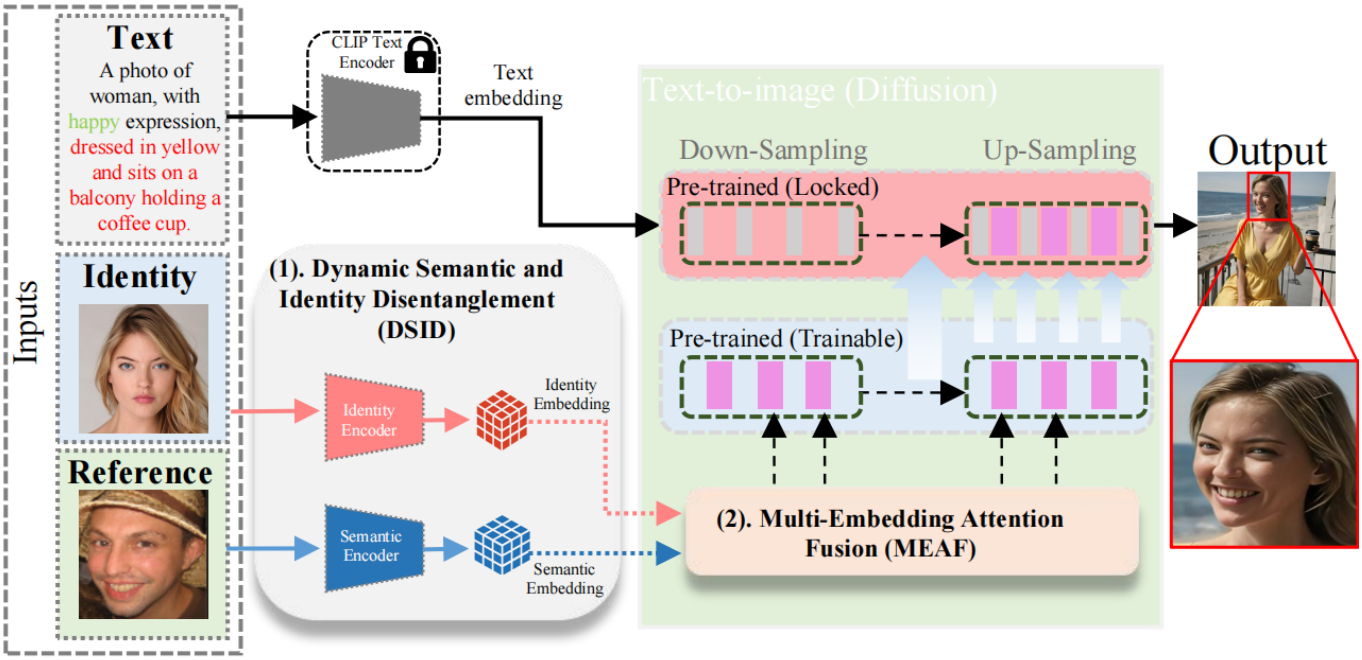
DeepExPo

Facial Expression and Pose Generation via Self-Supervised Disentangled Embeddings Fusion in Text-to-Image Diffusion Models

🔍 Overview

DeepExPo is a framework designed to provide fine-grained control over facial expression and head pose in text-to-image diffusion models, while maintaining identity preservation. Existing diffusion-based personalization methods typically fall short in expression control and pose alignment. DeepExPo overcomes these limitations through a self-supervised disentanglement strategy and adaptive embedding fusion with the U-Net of pre-trained text-to-image models.



Description: DeepExPo generates identity-preserving facial images with controllable expressions and head poses using three inputs: a text prompt for context, an identity image, and a reference image for semantic cues. Identity and semantic features are fused via the MEAF module into a pre-trained diffusion model.

Table 1: Capability comparison of different methods based on identity preservation (ID), expression control (EXP), pose manipulation (POSE), and context integration (CONTEXT). Ratings are derived from quantitative and qualitative metrics reported in the literature, as well as from our own empirical experiments.

Method	ID.	EXP.	POSE.	CONTEXT
GAN-based [7, 8]	High	High	High	None
Textual Inversion [9]	High	None	None	High
DreamBooth [10]	High	None	None	High
DisenBooth [11]	Moderate	Low	None	High
Celeb-Basis [12]	Moderate	Low	None	High
CIP-TIG [13]	Moderate	Low	None	High
MasterWeaver [14]	High	Moderate	None	High
EmojiDiff [15]	Moderate	Moderate	Low	High
DiffSFSR [16]	Moderate	Moderate	Moderate	High
DeepExPo	**High**	**High**	**High**	**High**

Description: Capability comparison of various methods based on identity preservation, expression control, pose manipulation, and context integration. Ratings are based on literature benchmarks and our own experimental analysis.

DeepExPo introduces a two-stage solution:

- **Dynamic Semantic and Identity Disentanglement (DSID)** module
- **Multi-Embedding Attention Fusion (MEAF)** module

This repository includes the full implementation of the **DeepExPo**.

🧠 Key Modules

1. Dynamic Semantic and Identity Disentanglement (DSID) Module

> 📦 `./modules/dsid/`

A self-supervised module designed to disentangle identity and dynamic semantic attributes (e.g., expressions and head pose) from single or paired image frames. This module:

- Leverages consecutive video frame pairs as implicit supervision to separate identity from both static and temporal variations
- Encodes identity and dynamic semantics into two independent latent spaces
- Enables composable conditioning on expressions and pose while preserving identity fidelity

2. Multi-Embedding Attention Fusion (MEAF) Module

The **MEAF** module employs a parallel multi-attention mechanism to fuse **semantic** and **identity embeddings** with intermediate U-Net features from a pre-trained diffusion model. This approach:

- Preserves **identity** and **facial fidelity**

- Avoids degradation typically caused by direct embedding injection
- Enables effective conditioning without disrupting spatial coherence

---

Installation

To use **DeepExPo**, clone the repository and install the required dependencies.

1. Clone the Repository

```
git clone https://github.com/MSAfganUSTC/DeepExPo.git
cd DeepExPo
```

After cloning, the folder contains the following structure:

2. Repository Structure

```
DeepExPo/
│
├── requr/
│   └── config.yaml                    # Configuration file
│
├── DeepExPo_DSID_Module/              # DSID module implementation
├── DeepExPo_MEAF_Module/              # MEAF module implementation
│
├── scripts/
│   └── DeepExPo_Inference.ipynb          # Inference script
│
├── DeepExPo_Weights/                  # Pretrained model weights
│   ├── DSID_Checkpoints/              # Checkpoints for DSID module
│   └── DeepExPo_MEAF_weights/         # Weights for MEAF module
│
├── Images/                            # Figures and illustrations used in
the paper
├── samples/
│   └── subject.jpg                    # Sample input image
│
├── LICENSE                            # License file
└── README.md                          # Project documentation
```

3. Set Up the Environment

This project uses Conda for environment management. Make sure you have Conda installed.

Create the environment from the provided file:

```
conda env create -f configs/environment.yml
conda activate DeepExPo
```

The dependencies include PyTorch, Hugging Face Transformers, and other necessary packages.

4. Usage

Interactive Inference (Jupyter Notebook)

All the steps for loading a model, choosing an expression, and generating images are wrapped in an easy-to-run notebook.

1. **Activate the environment**

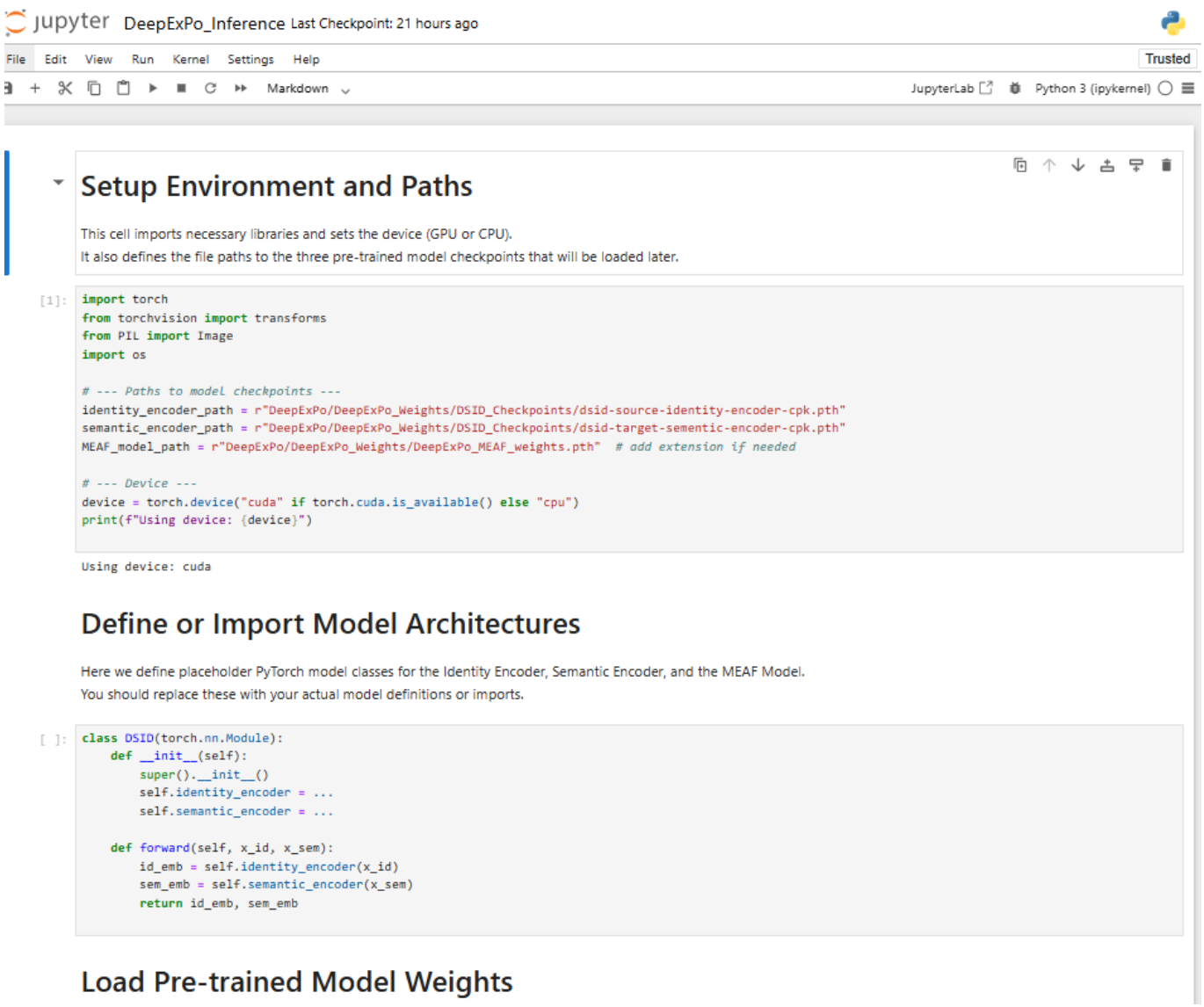
```
conda activate DeepExPo   # or mamba activate DeepExPo
```

2. **Launch Jupyter and open the notebook**

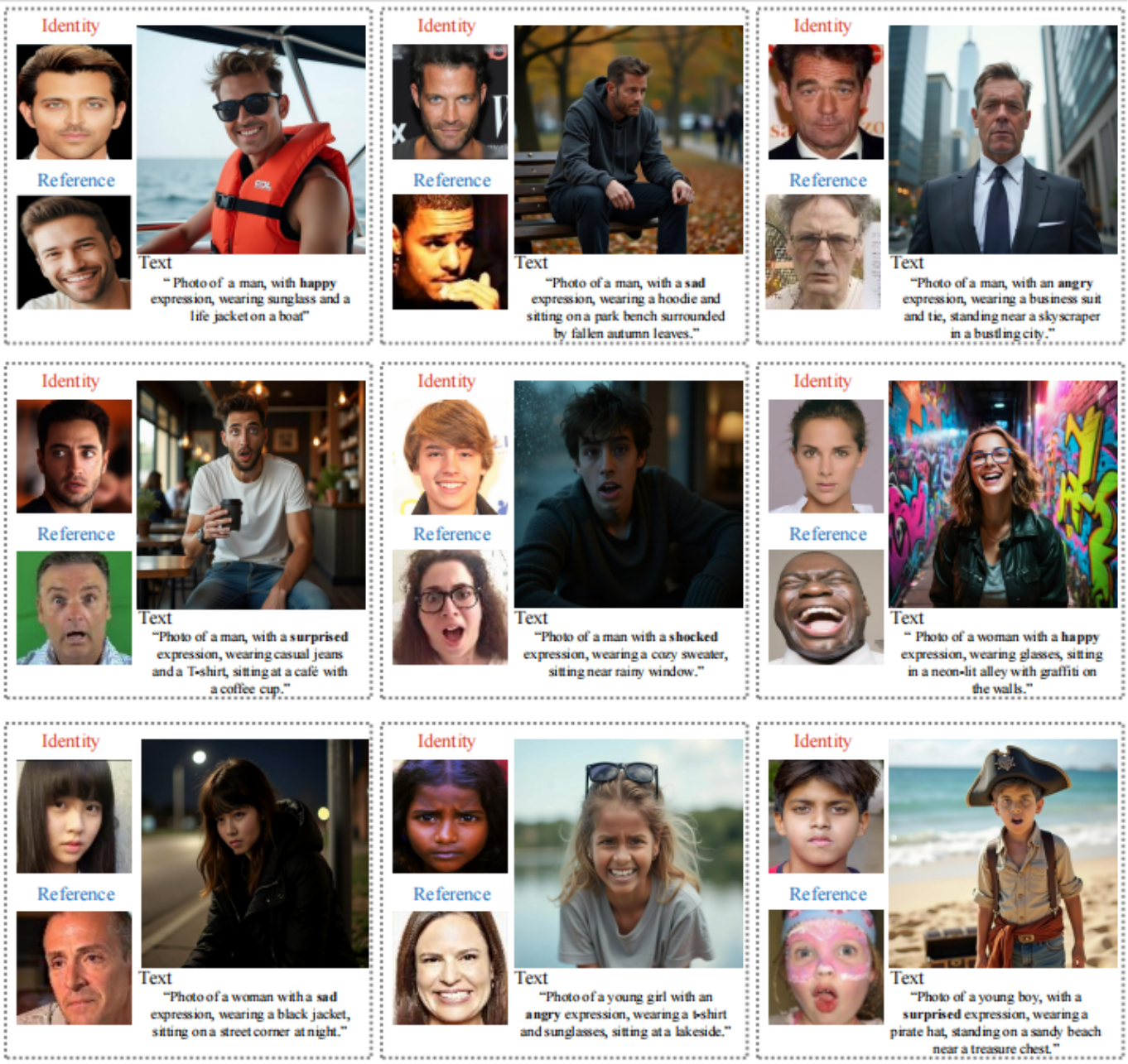
```
jupyter notebook scripts/DeepExPo_Inference.ipynb
```

3. **Run the cells from top to bottom**

- The first cell lets you set paths like `MODEL_PATH` to load DSID checkpoints (both identity encoder and semantic encoder).
- The second cell asks for identity and reference images to extract embeddings, and lets you specify the `output_dir`.
- Later cells handle loading the model, running inference, and saving the results.
- Generated images will appear inside the notebook and in the output directory you specified. Below is a snapshot of the Jupyter Notebook interface used for inference:



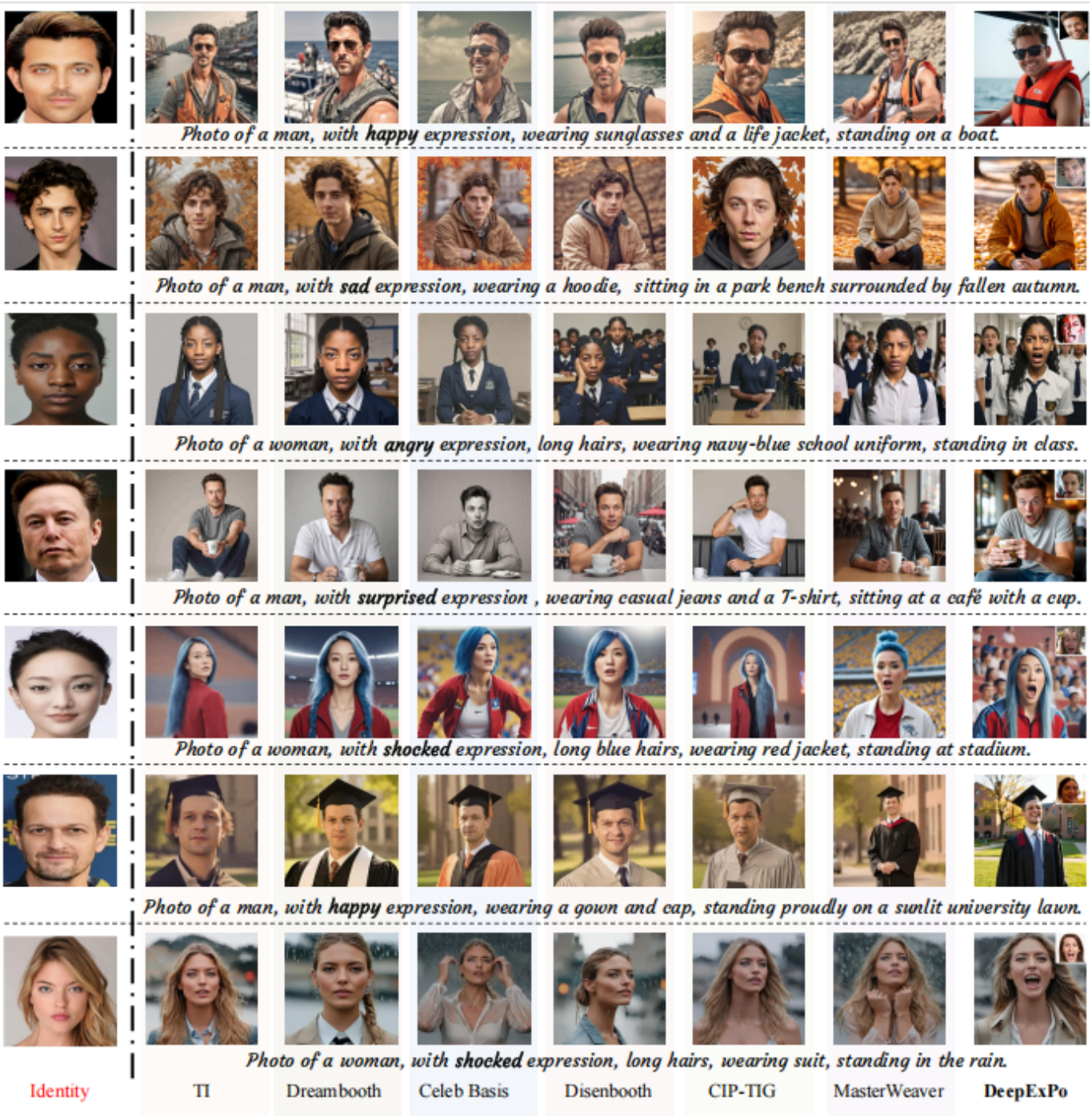
🖼️ Inference Results



Description: DeepExPo inference samples showing accurate expression and pose synthesis while maintaining subject identity and contextual fidelity.

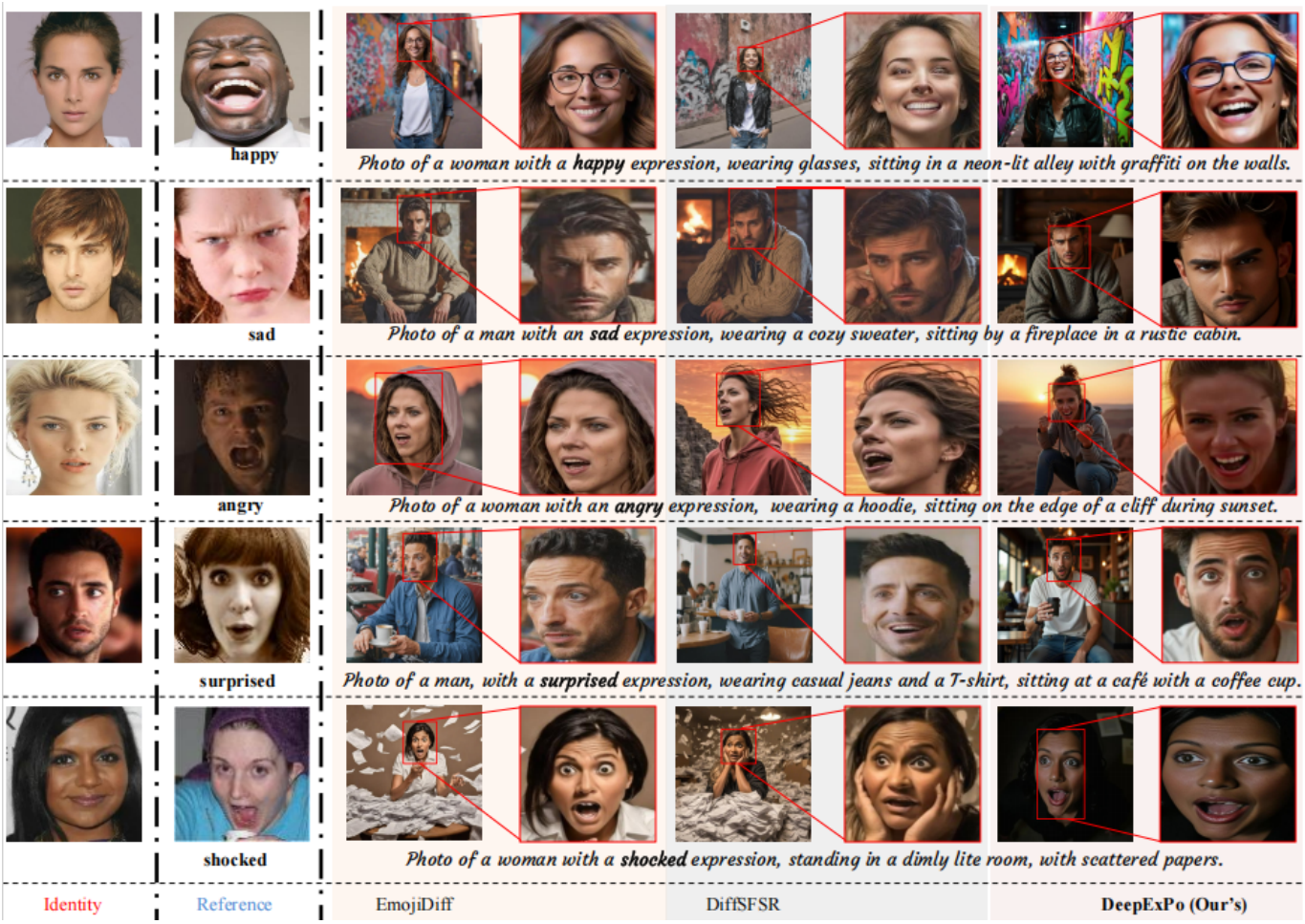
🔁 Comparison with Baselines

🧑‍🎨 **Personalized Generation Baselines**



Description: Qualitative comparison against personalized generation models. **DeepExPo** maintains strong identity consistency while effectively transferring facial semantics and adhering to contextual prompts.

😮 Expression Generation Baselines



Description: DeepExPo achieves better identity preservation and precise semantic transfer (e.g., mouth and eye region alignment) compared to other expression generation techniques.

📊 Quantitative Evaluation

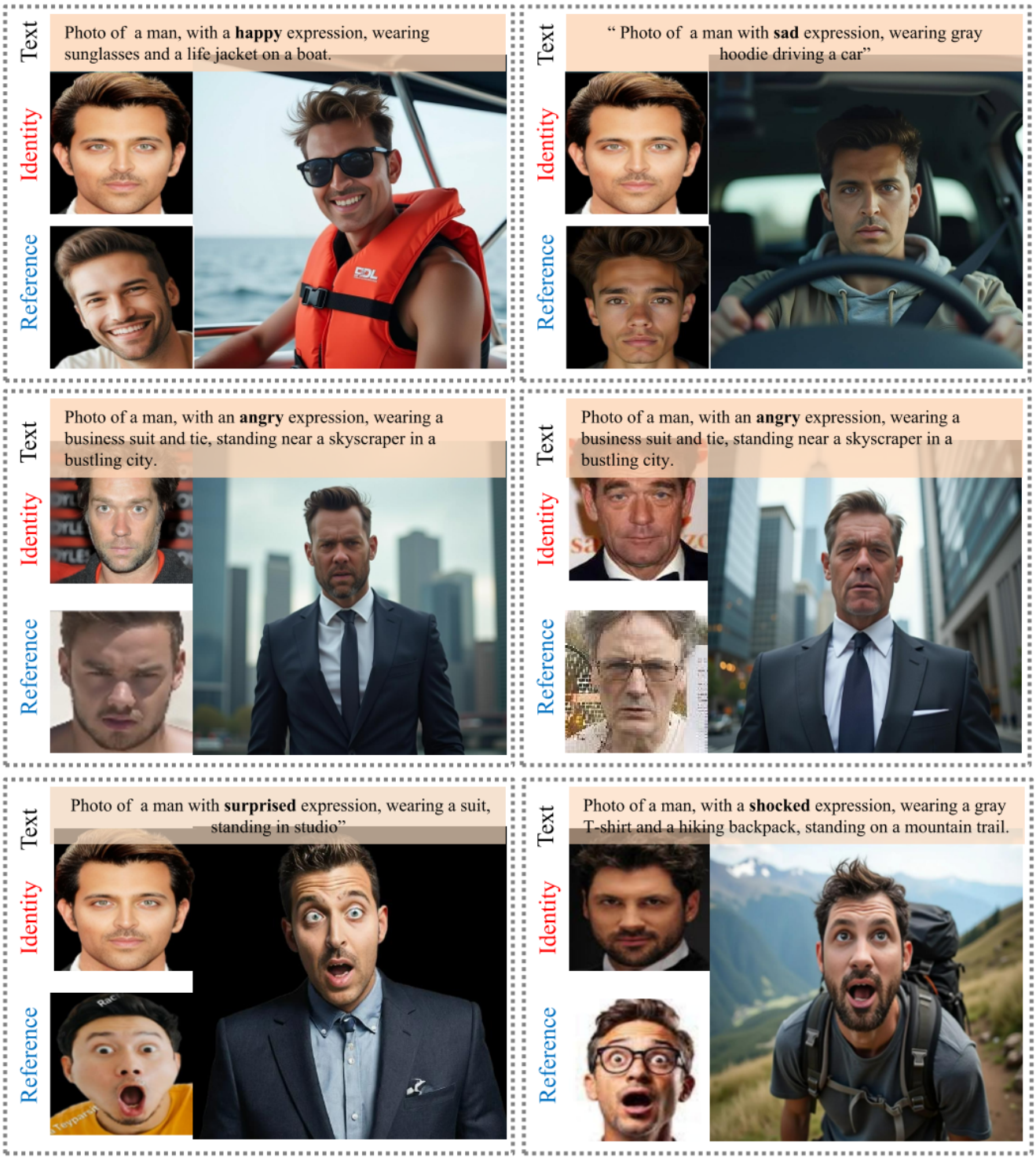
Table 3: Quantitative comparison against personalized and expression generation baselines. **Bold** indicates the best, and underline the second-best performance for each metric.

	Method	ID.↓	SEMCHG.↑	SIM.↓	EXP.(%)↑	POSE↑	FID↓	TIME↓
Personalized	TI [9]	0.48	0.35	-	30	-	128.3	10s
	DreamBooth [10]	**0.25**	0.20	-	-	-	119.5	14s
	DisenBooth [11]	0.47	0.42	-	35	-	121.0	14s
	Celeb Basis [12]	0.56	0.40	-	38	-	122.5	20s
	CIP-TIG [13]	0.85	0.70	-	80	-	124.3	22s
	MW [14]	0.53	0.66	-	75	-	124.9	8s
	DeepExPo	0.40	**0.80**	**0.29**	**90**	**2.50**	**119.21**	**4s**
Express	EmojiDiff [15]	0.57	0.68	0.45	85	4.59	124.3	6s
	DiffSFSR [16]	0.50	0.77	0.38	87	3.00	122.8	5s
	DeepExPo	**0.40**	**0.80**	**0.29**	**90**	**2.50**	**119.21**	**4s**

Description: Quantitative evaluation highlights **DeepExPo**'s superior performance in identity preservation and expression accuracy. Qualitative results emphasize expression realism, head pose accuracy, and facial fidelity.

🎯 Additional Results

🧔 Male Subjects



DeepExPo successfully generates realistic male facial expressions while preserving identity.

👩 Female Subjects

Expression synthesis for female subjects across five expression types with high visual fidelity.

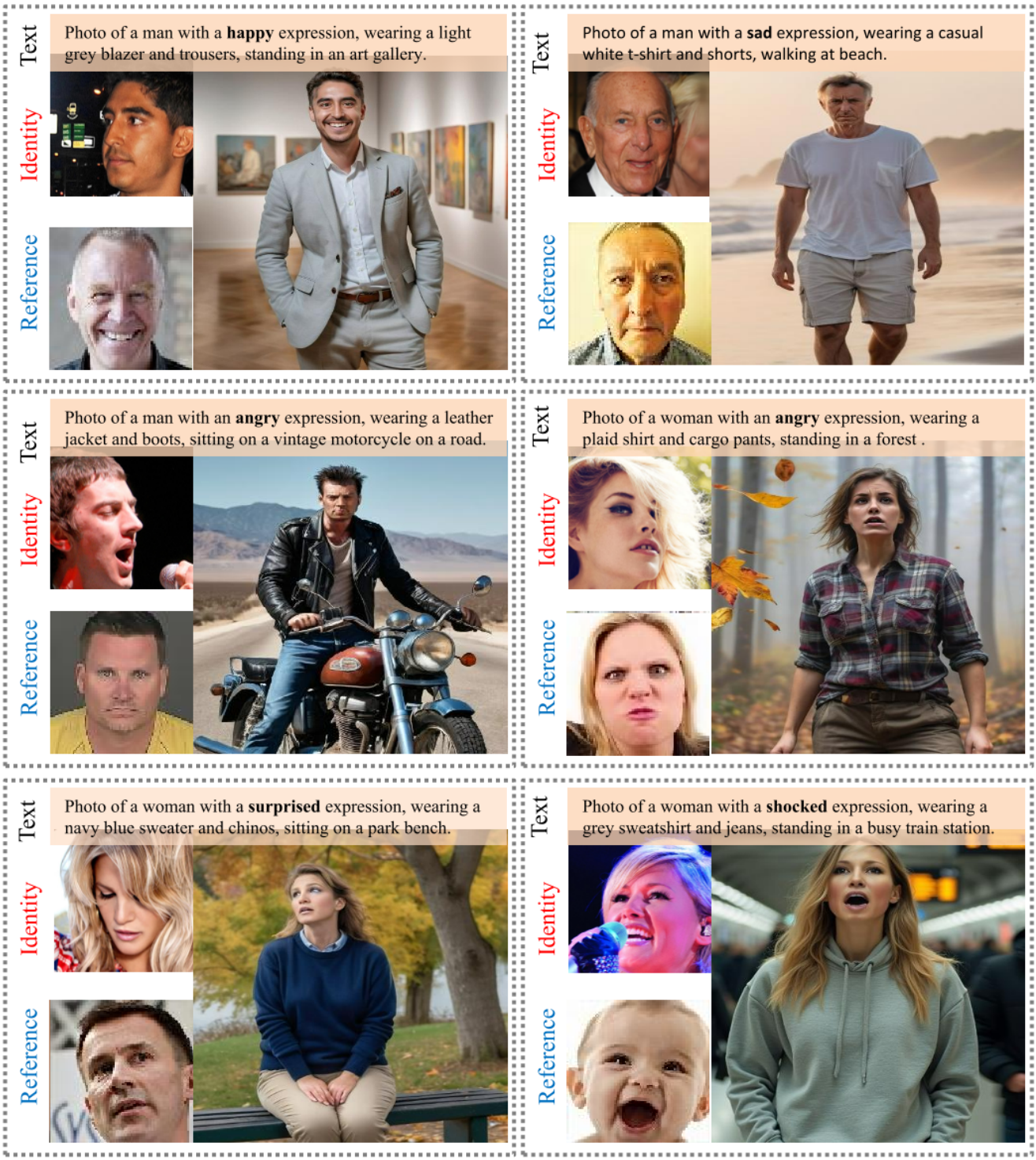
🧒 Children Across Ethnic Groups

Results demonstrate effective generation for young boys and girls of Indian, African, and European descent.

🌍 Ethnic Diversity

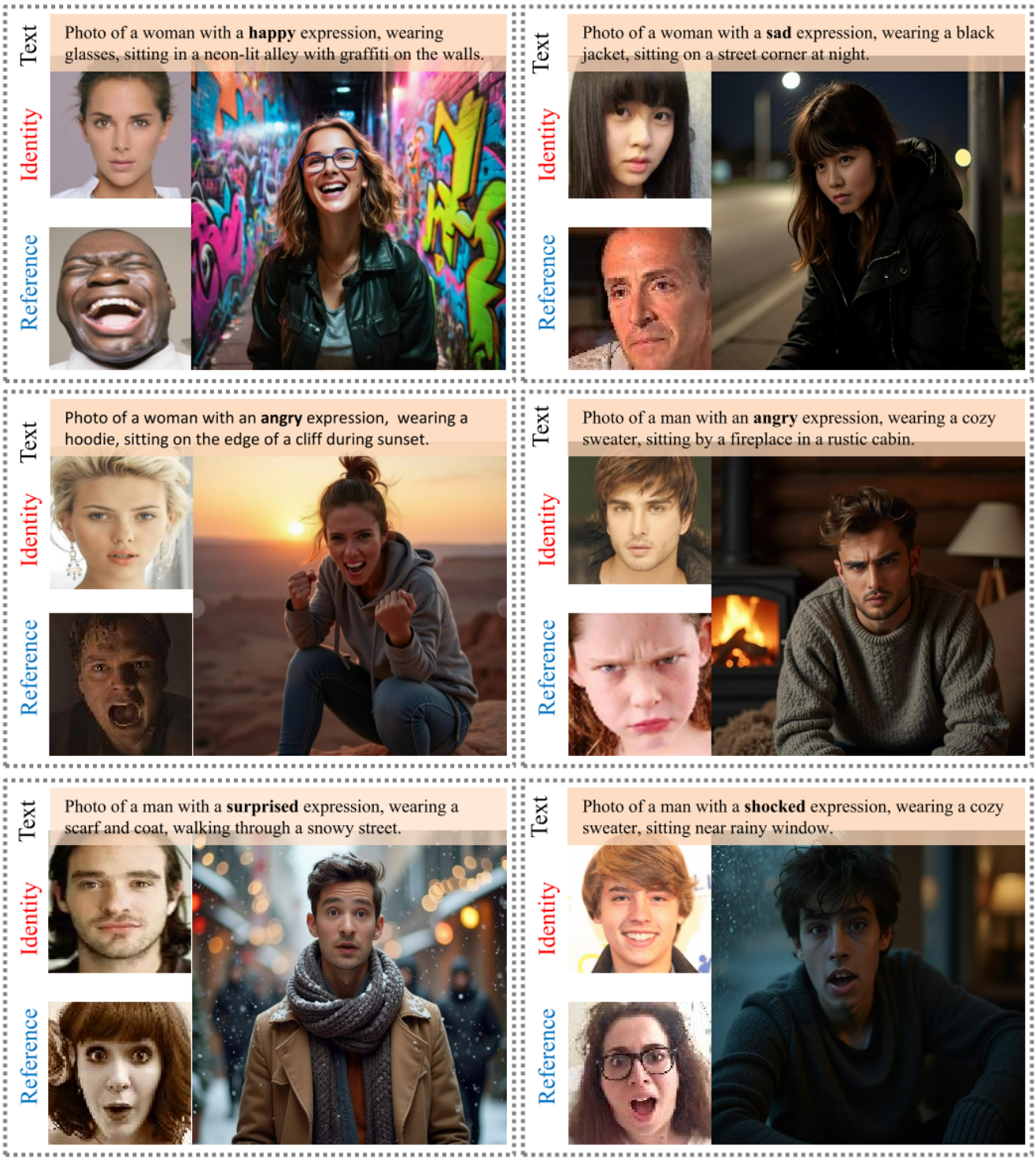
The model maintains identity and expression accuracy across various ethnic groups.

🔄 Cross-Identity/Reference Combinations



Demonstrates flexibility in handling cross-gender and cross-ethnicity transformations while preserving identity.

🧑‍🦰 Extreme Orientations



Results show model limitations when both identity and reference inputs have extreme head poses, reflecting the boundaries of identity fidelity under such conditions.

✅ Conclusion

DeepExPo demonstrates robust performance in identity-preserving facial expression synthesis across diverse subjects, conditions, and contexts. Its ability to handle complex semantic cues and maintain realism positions it as a strong foundation for personalized human image generation in real-world applications.

License

Acknowledgments

- **Diffusion Models**: We used diffusion-based image generation techniques for creating high-quality images.
- **Open Source Tools**: The project leverages various open-source libraries such as PyTorch, Hugging Face Transformers, and others.

📬 Contact

For questions, collaboration, or code access, please contact:

Muhammad Afgan
✉️ [msafgan@mail.ustc.edu.cn]