The Struggle for Zero-Shot Customization in Generative AI

Originally published at https://www.unite.ai/the-struggle-for-zero-shot-customization-in-generative-ai/

If you want to place yourself into a popular image or video generation tool – but you're not already famous enough for the foundation model to recognize you – you'll need to train a low-rank adaptation (LoRA) model using a collection of your own photos. Once created, this personalized LoRA model allows the generative model to include your identity in future outputs.

This is commonly called *customization* in the image and video synthesis research sector. It first emerged a few months after the advent of Stable Diffusion in the summer of 2022, with Google Research's DreamBooth project offering high-gigabyte customization models, in a closed-source schema that was soon adapted by enthusiasts and released to the community.

LoRA models quickly followed, and offered easier training and far lighter file-sizes, at minimal or no cost in quality, quickly dominating the customization scene for Stable Diffusion and its successors, later models such as Flux, and now new generative video models like Hunyuan Video and Wan 2.1.

Rinse and Repeat

The problem is, as we've noted before, that every time a new model comes out, it needs a new generation of LoRAs to be trained, which represents considerable friction on LoRA-producers, who may train a range of custom models only to find that a model update or popular newer model means they need to start all over again.

Therefore zero-shot customization approaches have become a strong strand in the literature lately. In this scenario, instead of needing to curate a dataset and train your own sub-model, you simply supply one or more photos of the subject to be injected into the generation, and the system interprets these input sources into a blended output.

Below we see that besides face-swapping, a system of this type (here using PuLID) can also incorporate ID values into style transfer:

Examples of facial ID transference using the PuLID system. Source: https://github.com/ToTheBeginning/PuLID?tab=readme-ov-file

While replacing a labor-intensive and fragile system like LoRA with a generic adapter is a great (and popular) idea, it's challenging too; the extreme attention to detail and coverage obtained in the LoRA training process is very difficult to imitate in a one-shot IP-Adapter-style model, which has to match LoRA's level of detail and flexibility without the prior advantage of analyzing a comprehensive set of identity images.

HyperLoRA

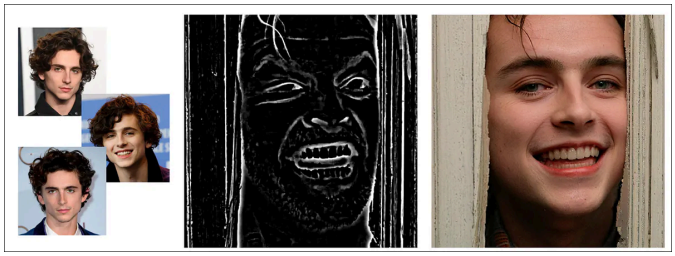
With this in mind, there's an interesting new paper from ByteDance proposing a system that generates actual LoRA code *on-the-fly*, which is currently unique among zero-shot solutions:

On the left, input images. Right of that, a flexible range of output based on the source images, effectively producing deepfakes of actors Anthony Hopkins and Ann Hathaway. Source: https://arxiv.org/pdf/2503.16944

The paper states:

> *'Adapter based techniques such as IP-Adapter freeze the foundational model parameters and employ a plug-in architecture to enable zero-shot inference, but they often exhibit a lack of naturalness and authenticity, which are not to be overlooked in portrait synthesis tasks.*
>
> *'[We] introduce a parameter-efficient adaptive generation method namely HyperLoRA, that uses an adaptive plug-in network to generate LoRA weights, merging the superior performance of LoRA with the zero-shot capability of adapter scheme.*
>
> *'Through our carefully designed network structure and training strategy, we achieve zero-shot personalized portrait generation (supporting both single and multiple image inputs) with high photorealism, fidelity, and editability.'*

Most usefully, the system as trained can be used with existing ControlNet, enabling a high level of specificity in generation:



Timothy Chalomet makes an unexpectedly cheerful appearance in 'The Shining' (1980), based on three input photos in HyperLoRA, with a ControlNet mask defining the output (in concert with a text prompt).

As to whether the new system will ever be made available to end-users, ByteDance has a reasonable record in this regard, having released the very powerful LatentSync lip-syncing framework, and having only just released also the InfiniteYou framework.

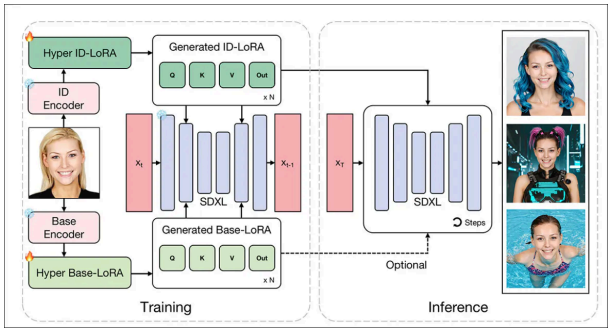
Negatively, the paper gives no indication of an intent to release, and the training resources needed to recreate the work are so exorbitant that it would be challenging for the enthusiast community to recreate (as it did with DreamBooth).

The new paper is titled *HyperLoRA: Parameter-Efficient Adaptive Generation for Portrait Synthesis*, and comes from seven researchers across ByteDance and ByteDance's dedicated Intelligent Creation department.

Method

The new method utilizes the Stable Diffusion latent diffusion model (LDM) SDXL as the foundation model, though the principles seem applicable to diffusion models in general (however, the training demands – see below – might make it difficult to apply to generative video models).

The training process for HyperLoRA is split into three stages, each designed to isolate and preserve specific information in the learned weights. The aim of this ring-fenced procedure is to prevent identity-relevant features from being polluted by irrelevant elements such as clothing or background, at the same time as achieving fast and stable convergence.



Conceptual schema for HyperLoRA. The model is split into 'Hyper ID-LoRA' for identity features and 'Hyper Base-LoRA' for background and clothing. This separation reduces feature leakage. During training, the SDXL base and encoders are frozen, and only HyperLoRA modules are updated. At inference, only ID-LoRA is required to generate personalized images.

The first stage focuses entirely on learning a *'Base-LoRA'* (lower-left in schema image above), which captures identity-irrelevant details.

To enforce this separation, the researchers deliberately blurred the face in the training images, allowing the model to latch onto things such as background, lighting, and pose – but not identity. This 'warm-up' stage acts as a filter, removing low-level distractions before identity-specific learning begins.

In the second stage, an *'ID-LoRA'* (upper-left in schema image above) is introduced. Here, facial identity is encoded using two parallel pathways: a CLIP Vision Transformer (CLIP ViT) for structural features and the InsightFace AntelopeV2 encoder for more abstract identity representations.

Transitional Approach

CLIP features help the model converge quickly, but risk overfitting, whereas Antelope embeddings are more stable but slower to train. Therefore the system begins by relying more heavily on CLIP, and gradually phases in Antelope, to avoid instability.

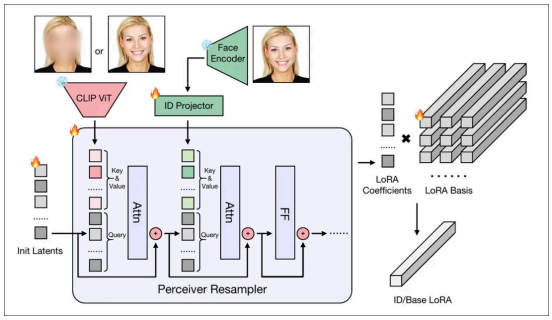
In the final stage, the CLIP-guided attention layers are frozen entirely. Only the AntelopeV2-linked attention modules continue training, allowing the model to refine identity preservation without degrading the fidelity or generality of previously learned components.

This phased structure is essentially an attempt at disentanglement. Identity and non-identity features are first separated, then refined independently. It's a methodical response to the usual failure modes of personalization: identity drift, low editability, and overfitting to incidental features.

While You Weight

After CLIP ViT and AntelopeV2 have extracted both structural and identity-specific features from a given portrait, the obtained features are then passed through a *perceiver resampler* (derived from the aforementioned IP-Adapter project) – a transformer-based module that maps the features to a compact set of coefficients.

Two separate resamplers are used: one for generating Base-LoRA weights (which encode background and non-identity elements) and another for ID-LoRA weights (which focus on facial identity).



Schema for the HyperLoRA network.

The output coefficients are then linearly combined with a set of learned LoRA basis matrices, producing full LoRA weights without the need to fine-tune the base model.

This approach allows the system to generate personalized weights *entirely on the fly*, using only image encoders and lightweight projection, while still leveraging LoRA's ability to modify the base model's behavior directly.

Data and Tests

To train HyperLoRA, the researchers used a subset of 4.4 million face images from the LAION-2B dataset (now best known as the data source for the original 2022 Stable Diffusion models).

InsightFace was used to filter out non-portrait faces and multiple images. The images were then annotated with the BLIP-2 captioning system.

In terms of data augmentation, the images were randomly cropped around the face, but always focused on the face region.

The respective LoRA ranks had to accommodate themselves to the available memory in the training setup. Therefore the LoRA rank for ID-LoRA was set to 8, and the rank for Base-LoRA to 4, while eight-step gradient accumulation was used to simulate a larger batch size than was actually possible on the hardware.

The researchers trained the Base-LoRA, ID-LoRA (CLIP), and ID-LoRA (identity embedding) modules sequentially for 20K, 15K, and 55K iterations, respectively. During ID-LoRA training, they sampled from three conditioning scenarios with probabilities of 0.9, 0.05, and 0.05.

The system was implemented using PyTorch and Diffusers, and the full training process ran for approximately ten days on 16 NVIDIA A100 GPUs*.

ComfyUI Tests

The authors built workflows in the ComfyUI synthesis platform to compare HyperLoRA to three rival methods: InstantID; the aforementioned IP-Adapter, in the form of the IP-Adapter-FaceID-Portrait framework; and the above-cited PuLID. Consistent seeds, prompts and sampling methods were used across all frameworks.

The authors note that Adapter-based (rather than LoRA-based) methods generally require lower Classifier-Free Guidance (CFG) scales, whereas LoRA (including HyperLoRA) is more permissive in this regard.

So for a fair comparison, the researchers used the open-source SDXL fine-tuned checkpoint variant *LEOSAM's Hello World* across the tests. For quantitative tests, the Unsplash-50 image dataset was used.

Metrics

For a fidelity benchmark, the authors measured facial similarity using cosine distances between CLIP image embeddings (CLIP-I) and separate identity embeddings (ID Sim) extracted via CurricularFace, a model not used during training.

Each method generated four high-resolution headshots per identity in the test set, with results then averaged.

Editability was assessed in both by comparing CLIP-I scores between outputs with and without the identity modules (to see how much the identity constraints altered the image); and by measuring CLIP image-text alignment (CLIP-T) across ten prompt variations covering *hairstyles*, *accessories*, *clothing*, and *backgrounds*.

The authors included the Arc2Face foundation model in the comparisons – a baseline trained on fixed captions and cropped facial regions.

For HyperLoRA, two variants were tested: one using only the ID-LoRA module, and another using both ID- and Base-LoRA, with the latter weighted at 0.4. While the Base-LoRA improved fidelity, it slightly constrained editability.

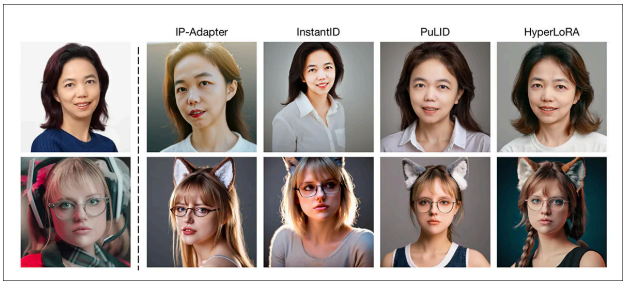
	Fidelity		Editability	
	CLIP-I↑	ID Sim.↑	CLIP-I↑	CLIP-T↑
IP-Adapter	0.764	0.566	0.725	0.244
InstantID	0.734	**0.681**	0.688	0.237
PuLID	0.771	0.613	**0.805**	**0.259**
Arc2Face	0.786	0.643	-	-
Ours (Full)	**0.853**	0.678	0.710	0.243
Ours (ID)	0.831	0.625	0.748	0.252

Results for the initial quantitative comparison.

Of the quantitative tests, the authors comment:

> *'Base-LoRA helps to improve fidelity but limits editability. Although our design decouples the image features into different LoRAs, it's hard to avoid leaking mutually. Thus, we can adjust the weight of Base-LoRA to adapt to different application scenarios.*
>
> *'Our HyperLoRA (Full and ID) achieve the best and second-best face fidelity while InstantID shows superiority in face ID similarity but lower face fidelity.*
>
> *'Both these metrics should be considered together to evaluate fidelity, since the face ID similarity is more abstract and face fidelity reflects more details.'*

In qualitative tests, the various trade-offs involved in the essential proposition come to the fore (please note that we do not have space to reproduce all the images for qualitative results, and refer the reader to the source paper for more images at better resolution):



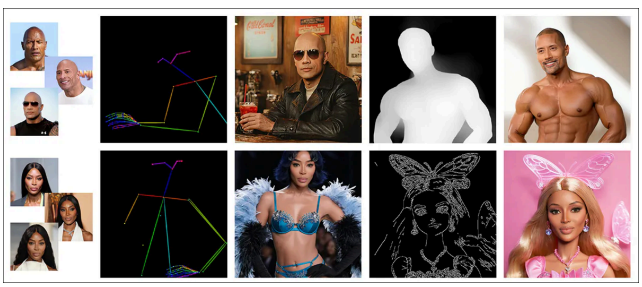
Qualitative comparison. From top to bottom, the prompts used were: 'white shirt' and 'wolf ears' (see paper for additional examples).

Here the authors comment:

> *'The skin of portraits generated by IP-Adapter and InstantID has apparent AI-generated texture, which is a little [oversaturated] and far from photorealism.*
>
> *'It is a common shortcoming of Adapter-based methods. PuLID improves this problem by weakening the intrusion to base model, outperforming IP-Adapter and InstantID but still suffering from blurring and lack of details.*
>
> *'In contrast, LoRA directly modifies the base model weights instead of introducing extra attention modules, usually generating highly detailed and photorealistic images.'*

The authors contend that because HyperLoRA modifies the base model weights directly instead of relying on external attention modules, it retains the nonlinear capacity of traditional LoRA-based methods, potentially offering an advantage in fidelity and allowing for improved capture of subtle details such as pupil color.

In qualitative comparisons, the paper asserts that HyperLoRA's layouts were more coherent and better aligned with prompts, and similar to those produced by PuLID, while notably stronger than InstantID or IP-Adapter (which occasionally failed to follow prompts or produced unnatural compositions).



Further examples of ControlNet generations with HyperLoRA.

Conclusion

The consistent stream of various one-shot customization systems over the last 18 months has, by now, taken on a quality of desperation. Very few of the offerings have made a notable advance on the state-of-the-art; and those that have advanced it a little tend to have exorbitant training demands and/or extremely complex or resource-intensive inference demands.

While HyperLoRA's own training regime is as gulp-inducing as many recent similar entries, at least one finishes up with a model that can handle *ad hoc* customization out of the box.

From the paper's supplementary material, we note that the inference speed of HyperLoRA is better than IP-Adapter, but worse than the two other former methods – and that these figures are based on a NVIDIA V100 GPU, which is not typical consumer hardware (though newer 'domestic' NVIDIA GPUs can match or exceed this the V100's maximum 32GB of VRAM).

Method	IP-Adapter	InstantID	PuLID	HyperLoRA
Preprocess	2996	758	236	1143
Inference	6148	8037	6616	4327

The inference speeds of competing methods, in milliseconds.

It's fair to say that zero-shot customization remains an unsolved problem from a practical standpoint, since HyperLoRA's significant hardware requisites are arguably at odds with its ability to produce a truly long-term single foundation model.

** Representing either 640GB or 1280GB of VRAM, depending on which model was used (this is not specified)*

First published Monday, March 24, 2025

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai