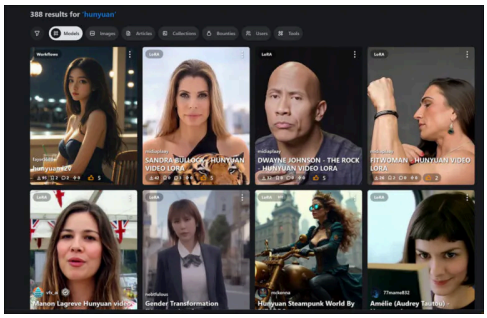
Towards LoRAs That Can Survive Model Version Upgrades

Originally published at https://www.unite.ai/towards-loras-that-can-survive-model-version-upgrades/

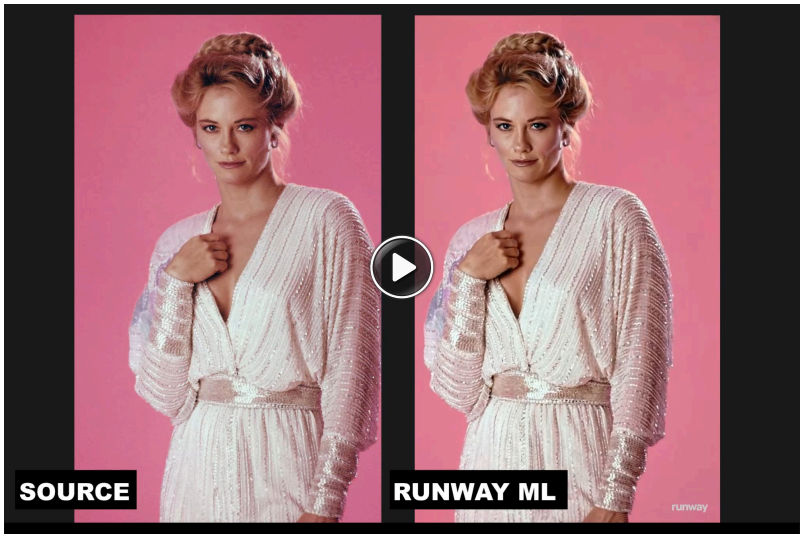
Since my recent coverage of the growth in hobbyist Hunyuan Video LoRAs (small, trained files that can inject custom personalities into multi-billion parameter text-to-video and image-to-video foundation models), the number of related LoRAs available at the Civit community has risen by 185%.



Despite the fact that there are no particularly easy or low-effort ways to make a Hunyuan Video LoRA, the catalog of celebrity and themed LoRAs at Civit is growing daily. Source: https://civitai.com/

The same community that is scrambling to learn how to produce these 'add-on personalities' for Hunyuan Video (HV) is also ulcerating for the promised release of an image-to-video (I2V) functionality in Hunyuan Video.

With regard to open source human image synthesis, this is a big deal; combined with the growth of Hunyuan LoRAs, it could enable users to transform photos of people into videos in a way that doesn't erode their identity as the video develops – which is currently the case in all state-of-the-art image-to-video generators, including Kling, Kaiber, and the much-celebrated RunwayML:



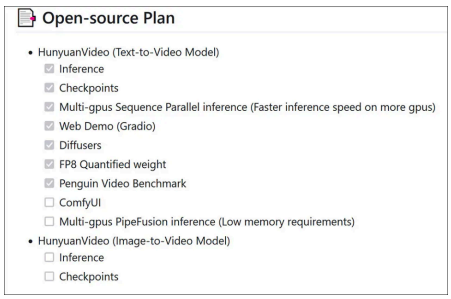
CLICK TO PLAY VIDEO ON SITE

Click to play. *An image-to-video generation from RunwayML's state-of-the-art Gen 3 Turbo model. However, in common with all similar and lesser rival models, it cannot maintain consistent identity when the subject turns away from the camera, and the distinct features of the starting image become a 'generic diffusion woman'.* Source: https://app.runwayml.com/

By developing a custom LoRA for the personality in question, one could, in a HV I2V workflow, use a real photo of them as a starting point. This is a far better 'seed' than sending a random number into the model's latent space and settling for whatever semantic scenario results. One could then use the LoRA, or multiple LoRAs, to maintain consistency of identity, hairstyles, clothing and other pivotal aspects of a generation.

Potentially, the availability of such a combination could represent one of the most epochal shifts in generative AI since the launch of Stable Diffusion, with formidable generative power handed over to open source enthusiasts, without the regulation (or 'gatekeeping', if you prefer) provided by the content censors in the current crop of popular gen vid systems.

As I write, Hunyuan image-to-video is an unticked 'to do' in the Hunyuan Video GitHub repo, with the hobbyist community reporting (anecdotally) a Discord comment from a Hunyuan developer, who apparently stated that the release of this functionality has been pushed back to some time later in Q1 due to the model being 'too uncensored'.



The official feature release checklist for Hunyuan Video. Source: https://github.com/Tencent/HunyuanVideo?tab=readme-ov-file#-open-source-plan

Accurate or not, the repo developers have substantially delivered on the rest of the Hunyuan checklist, and therefore Hunyuan I2V seems set to arrive eventually, whether censored, uncensored or in some way 'unlockable'.

But as we can see in the list above, the I2V release is apparently a separate model entirely – which makes it pretty unlikely that any of the current burgeoning crop of HV LoRAs at Civit and elsewhere will function with it.

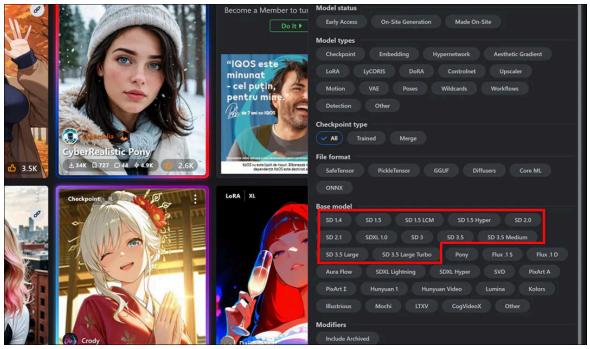
In this (by now) predictable scenario, LoRA training frameworks such as Musubi Tuner and OneTrainer will either be set back or reset in regard to supporting the new model. Meantime, one or two of the most tech-savvy (and entrepreneurial) YouTube AI luminaries will ransom their solutions via Patreon until the scene catches up.

Upgrade Fatigue

Almost no-one experiences upgrade fatigue as much as a LoRA or fine-tuning enthusiast, because the rapid and competitive pace of change in generative AI encourages model foundries such as Stability.ai, Tencent and Black Forest Labs to produce bigger and (sometimes) better models at the maximum viable frequency.

Since these new-and-improved models will at the very least have different biases and weights, and more commonly will have a different scale and/or architecture, this means that the fine-tuning community has to get their datasets out again and repeat the grueling training process for the new version.

For this reason, a multiplicity of Stable Diffusion LoRA version types are available at Civit:

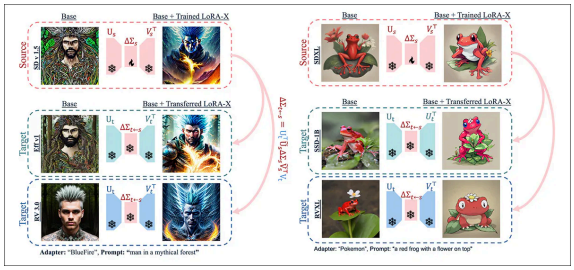


The upgrade trail, visualized in search filter options at civit.ai

Since none of these lightweight LoRA models are interoperable with higher or lower model versions, and since many of them have dependencies on popular large-scale merges and fine-tunes that adhere to an older model, a significant portion of the community tends to stick with a 'legacy' release, in much the same way as customer loyalty to Windows XP persisted years after official past support ended.

Adapting to Change

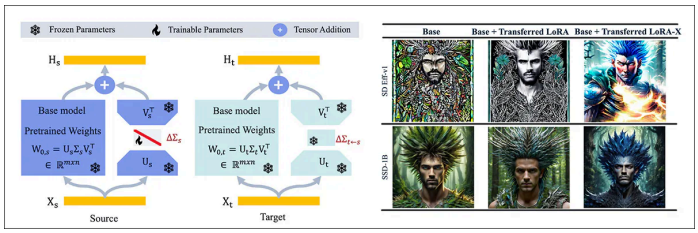
This subject comes to mind because of a new paper from Qualcomm QCOM 0.65% AI Research that claims to have developed a method whereby existing LoRAs can be 'upgraded' to a newly-released model version.



Example conversion of LoRAs across model versions. Source: https://arxiv.org/pdf/2501.16559

This does not mean that the new approach, titled *LoRA-X*, can translate freely between all models of the same type (i.e., text to image models, or Large Language Models [LLMs]); but the authors have demonstrated an effective transliteration of a LoRA from Stable Diffusion v1.5 > SDXL, and a conversion of a LoRA for the text-based TinyLlama 3T model to TinyLlama 2.5T.

LoRA-X transfers LoRA parameters across different base models by preserving the adapter within the source model's subspace; but only in parts of the model that are adequately similar across model versions.



On the left, a schema for the way that the LoRA-X source model fine-tunes an adapter, which is then adjusted to fit the target model. On the right, images generate target models SD Eff-v1.0 and SSD-1B, after applying adapters transferred from SD-v1.5 and SDXL without additional training.

While this offers a practical solution for scenarios where retraining is undesirable or impossible (such as a change of license on the original training data), the method is restricted to similar model architectures, among other limitations.

Though this is a rare foray into an understudied field, we won't examine this paper in depth because of LoRA-X's numerous shortcomings, as evidenced by comments from its critics and advisors at Open Review.

The method's reliance on subspace similarity restricts its application to closely related models, and the authors have conceded in the review forum that LoRA-X cannot be easily transferred across significantly different architectures

Other PEFT Approaches

The possibility of making LoRAs more portable across versions is a small but interesting strand of study in the literature, and the main contribution that LoRA-X makes to this pursuit is its contention that it requires no training. This is not strictly true, if one reads the paper, but it does require the least training of all the prior methods.

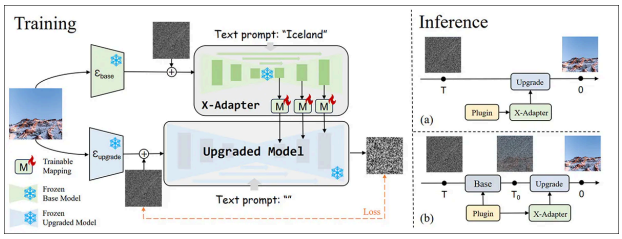
LoRA-X is another entry in the canon of Parameter-Efficient Fine-Tuning (PEFT) methods, which address the challenge of adapting large pre-trained models to specific tasks without extensive retraining. This conceptual approach aims to modify a minimal number of parameters while maintaining performance.

Notable among these are:

X-Adapter

The X-Adapter framework transfers fine-tuned adapters across models with a certain amount of retraining. The system aims to enable pre-trained plug-and-play modules (such as ControlNet and LoRA) from a base diffusion model (i.e., Stable Diffusion v1.5) to work directly with an upgraded diffusion model such as SDXL without retraining – effectively acting as a 'universal upgrader' for plugins.

The system achieves this by training an additional network that controls the upgraded model, using a frozen copy of the base model to preserve plugin connectors:

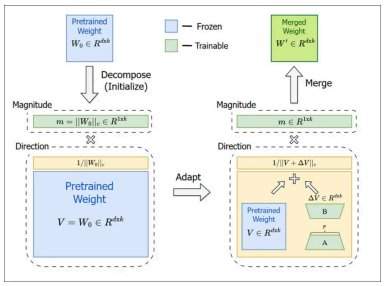


Schema for X-Adapter. Source: https://arxiv.org/pdf/2312.02238

X-Adapter was originally developed and tested to transfer adapters from SD1.5 to SDXL, while LoRA-X offers a wider variety of transliterations.

DoRA (Weight-Decomposed Low-Rank Adaptation)

DoRA is an enhanced fine-tuning method that improves upon LoRA by using a weight decomposition strategy that more closely resembles full fine-tuning:



DORA does not just attempt to copy over an adapter in a frozen environment, as LoRA-X does, but instead changes fundamental parameters of the weights, such as magnitude and direction. Source: https://arxiv.org/pdf/2402.09353

DoRA focuses on improving the fine-tuning process itself, by decomposing the model's weights into magnitude and direction (see image above). Instead, LoRA-X focuses on enabling the transfer of existing fine-tuned parameters between different base models

However, the LoRA-X approach adapts the *projection* techniques developed for DORA, and in tests against this older system claims an improved DINO score.

FouRA (Fourier Low Rank Adaptation)

Published in June of 2024, the FouRA method comes, like LoRA-X, from Qualcomm AI Research, and even shares some of its testing prompts and themes.

Examples of distribution collapse in LoRA, from the 2024 FouRA paper, using the Realistic Vision 3.0 model trained with LoRA and FouRA for 'Blue Fire' and 'Orig style adapters, across four seeds. LoRA images exhibit distribution collapse and reduced diversity, whereas FouRA generates more varied outputs. Source: https://arxiv.org/pdf/2406.08798

FouRA focuses on improving the diversity and quality of generated images by adapting LoRA in the frequency domain, using a Fourier transform approach.

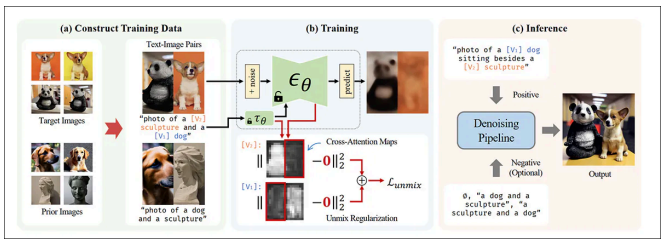
Here, again, LoRA-X was able to achieve better results than the Fourier-based approach of FouRA.

Though both frameworks fall within the PEFT category, they have very different use cases and approaches; in this case, FouRA is arguably 'making up the numbers' for a testing round with limited like-for-like rivals for the new paper's authors engage with.

SVDiff

SVDiff also has different goals to LoRA-X, but is strongly leveraged in the new paper. SVDiff is designed to improve the efficiency of the fine-tuning of diffusion models, and directly modifies values within the model's weight matrices, while keeping the singular vectors unchanged. SVDiff uses truncated SVD, modifying only the largest values, to adjust the model's weights.

This approach uses a data augmentation technique called *Cut-Mix-Unmix*:



Multi-subject generation operates as a concept-isolating system in SVDiff. Source: https://arxiv.org/pdf/2303.11305

Cut-Mix-Unmix is designed to help the diffusion model learn multiple distinct concepts without intermingling them. The central idea is to take images of different subjects and concatenate them into a single image. Then the model is trained with prompts that explicitly describe the separate elements in the image. This forces the model to recognize and preserve distinct concepts instead of blending them.

During training, an additional regularization term helps prevent cross-subject interference. The authors' theory contends that this facilitates improved multi-subject generation, where each element remains visually distinct, rather than being fused together.

SVDiff, excluded from the LoRA-X testing round, aims to create a compact parameter space. LoRA-X, instead, focuses on the transferability of LoRA parameters across different base models by operating within the subspace of the original model.

Conclusion

The methods discussed here are not the sole denizens of PEFT. Others include QLoRA and QA-LoRA; Prefix Tuning; Prompt-Tuning; and adapter-tuning, among others.

The 'upgradable LoRA' is, perhaps, an alchemical pursuit; certainly, there's nothing immediately on the horizon that will prevent LoRA modelers from having to drag out their old datasets again for the latest and greatest weights release. If there is some possible prototype standard for weights revision, capable of surviving changes in architecture and ballooning parameters between model versions, it hasn't emerged in the literature yet, and will need to keep being extracted from the data on a per-model basis.

First published Thursday, January 30, 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

Discover More