

LoFA: Learning to Predict Personalized Priors for Fast Adaptation of Visual Generative Models

Supplementary Material

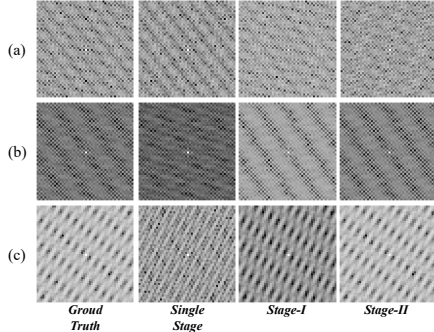


Figure 1. **Qualitative comparison of LoRA response maps.** Each row corresponds to a distinct task-specific LoRA, while columns represent the response map produced by different models or ground truth.

A. More Results

Fig. 4, Fig. 5, Fig. 6 and Fig. 7 present more qualitative results on text/pose-conditioned Personalized Human Action Video Generation, Text-to-Video Stylization and Identity-Personalized Image Generation tasks, where our method consistently outperforms other baselines.

We further provide a *demo* video in the supplementary file, named *demo.mp4*. In this video, LoRA [?] and its problems are introduced from start to 00:43. Next, the advantage of our LoFA is briefly illustrated. Finally, the qualitative comparisons across all downstream tasks, including the generated videos and images, are shown from 01:03 to the end.

B. User Study

To complement our evaluation, we conducted a subjective user study across all personalization tasks. We recruited 50 participants with no expertise in visual generation via an online questionnaire. For each task, we randomly select 5 prompts from the validation set and send them to each method to generate corresponding results, which are displayed to participants side-by-side in a random order alongside the original prompt. During the evaluation, we instruct subjects to select the best result based on prompt alignment and visual quality. As shown in Tab. 1, our method achieved a higher win rate across all tasks, demonstrating its clear superiority.

C. More Baselines on Video Stylization

As discussed in ?? and ??, we include a simplified HyperDreambooth[‡] to satisfy memory limitations. We additionally trained RPG [47] and ICM-LoRA [60], which are not for generation, using official scripts. Tab. 2 shows that our method consistently outperforms these baselines.

D. More Ablations and Analysis

Qualitative comparison of LoRA response maps. Fig. 1 visualizes the response maps derived from the ground truth (GT) LoRA [?], the single-stage prediction (*i.e.*, directly predicting LoRA weights without the guidance from response maps), and both stages of our proposed method. Compared to the GT LoRA, the single-stage prediction’s response map shows poor alignment. In contrast, both our Stage-I prediction and the final Stage-II prediction align closely with the GT. This demonstrates that our two-stage learning effectively guides the hypernetwork to identify and emphasize the regions requiring greater attention.

Threshold to get the LoRA response map. Tab. 3 ablates the threshold to get the binary-masked LoRA response map. A low threshold (*e.g.*, 0.01) masks a negligible area, revealing no distinct pattern, while a high threshold (*e.g.*, 0.03) masks an excessive area. We find that a threshold between 0.015 and 0.025 provides an optimal balance, yielding stable and high-quality results.

Are the masked parameters truly negligible? We investigate whether the parameters masked by our threshold are functionally useless. To test this, we perturb these parameters by adding Gaussian noise or setting them to zero. As shown in Fig. 2, these perturbations cause *no* essential performance difference, confirming that the masked parameters are indeed negligible.

Attending to the feature of the Stage-I model. In our Stage-II model, an additional cross-attention layer is introduced to attend to the final layer feature representation of the Stage-I model. We further conduct an ablation study on *how many layers* and *which layers* should be attended. First, Tab. 4 shows the results of attending the Stage-I feature at the *first N-th* layers in Stage-II, indicating that attending at the first 2-th layers yields the best result. Furthermore, Tab. 5 shows that injecting the features into the 4th and 8th blocks yields the best overall performance.

	Text-Cond Human Action Video			Pose-Cond Human Action Video		Text-to-Video Stylization		Identity-Personalized Image Generation		
	LoRA [?]]	Text-to-LoRA [?]]	Ours	LoRA [?]]	Ours	LoRA [?]]	Ours	DreamBooth [?]]	HyperDreamBooth [?]]	Ours
Win Rate (%)	40.8	2.4	56.8	48.8	51.2	47.6	52.4	15.2	36.8	48.0

Table 1. User studies across all personalization tasks.

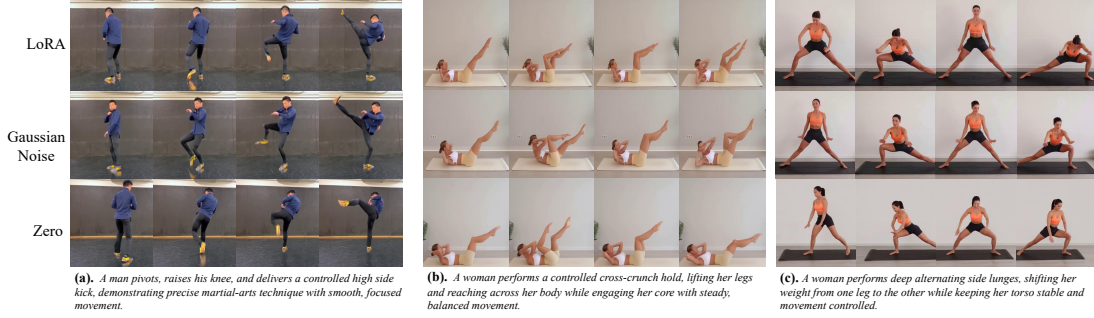


Figure 2. We investigate whether the parameters masked by our threshold are functionally useless. We perturb these parameters by adding Gaussian noise (**middle**) or setting them to zero (**bottom**). Compared with the original LoRA (**Top**), these perturbations cause *no* essential performance difference, confirming that the masked parameters are indeed negligible.

Injected Layers	FVD ↓	Dynamic Degree ↑
1st & 2nd	660.1	0.2067
2nd & 4th	644.3	0.2104
4th & 6th	612.2	0.2236
6th & 8th	633.9	0.2106
4th & 8th (ours)	589.8	0.2283

Table 5. Quantitative ablations on different cross-attention layers for injecting the Stage-I feature, on text-conditioned Personalized Human Action Video Generation.

Method	CSD-Score ↑	CLIP-T ↑	D.D. ↑
HyperDreamBooth [†]	0.374	0.2841	2.185
RPG [?]]	0.382	0.2857	2.161
ICM-LoRA [?]]	0.2927	0.3610	2.239
Ours	0.427	0.2943	2.394

Table 2. More baseline comparisons on video stylization.

Threshold	FVD ↓	Clip-T ↑	D.D ↑
0.01	634.7	0.3687	0.2219
0.015	602.9	0.3691	0.2245
0.02 (ours)	589.8	0.3719	0.2283
0.025	597.1	0.3712	0.2253
0.03	622.5	0.3620	0.2197

Table 3. Quantitative ablations on the threshold to get the response map, on text-conditioned Personalized Human Action Video Generation.

N Layers	2	4	6	8
FVD ↓	660.1	668.3	678.2	680.9
Dynamic Degree ↑	0.2067	0.2051	0.1913	0.1845

Table 4. Quantitative ablations on attending the Stage-I feature on the *first* N -th layers in Stage-II, on text-conditioned Personalized Human Action Video Generation.

E. Implementation Details

Training strategy. All experiments follow the same training protocol to train our framework: the first stage is trained for 4,000 steps with a learning rate of $1e-4$, and the second stage for 7,000 steps with a learning rate of $4e-5$. We set loss weights $\lambda_{\text{recon}} = 5$ and $\lambda_{\text{diff}} = 1$, use the default flow-matching objective used in the base model [?] , and apply the AdamW [?] optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.01, $\epsilon = 1e-8$) with 500 warm-up steps and a batch size of 4. We use the LoRA parameters from linear projections in attention layers as the training supervision. During the data preparation, all LoRAs [?] are optimized for 1,000 steps with a constant learning rate of $1e-4$ and a batch size of 4. We select representative base models, including a DiT-based model (WAN, rank 32) and a UNet-based model (SDXL, rank 16), using the default training scripts and recommended settings provided by the official implementations.

Architecture Details We adopt a standard ViT architecture for both stages with RMSNorm and qk_norm, consisting of 8 transformer blocks with 768 hidden dimensions and a [CLS] token. Cross-attention is applied as a residual module with a dropout rate of 0.1, where Q is derived

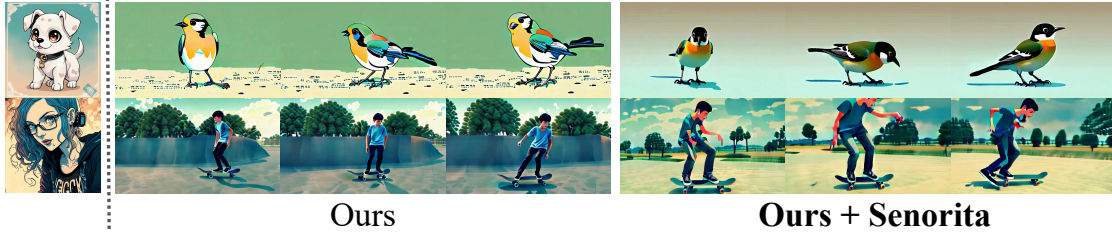


Figure 3. Qualitative comparison on video stylization. **Ours + Seniorita** preserves more sense of depth (1st row) and shifts further toward anime-style (2nd row).

Input Text
Prompt (a): A woman is performing a workout routine on a black yoga mat placed on a gray floor. The woman starts by lying on her back with her legs extended straight out. She then lifts one leg at a time, keeping it straight and tapping her toes. She repeats this movement several times, alternating between lifting each leg. The video captures her form and technique as she performs the exercise.
Prompt (b): In this video, a man is seen practicing with a katana in an indoor studio setting. His posture is dynamic and focused, indicating that he is engaged in a serious practice session. He begins by standing with his back to the camera, holding the katana in a ready position. As the video progresses, he transitions into various stances and movements, demonstrating different techniques and maneuvers with the sword. His movements are fluid and precise, showcasing his skill and control over the weapon. Throughout the video, the man maintains a strong and confident stance, occasionally shifting his weight from one foot to another while maintaining a firm grip on the katana. His facial expression remains neutral, suggesting concentration and dedication to his practice.
Prompt (c): In this video, a man is performing a series of dynamic movements in an indoor studio setting. He begins by standing with his feet apart, arms outstretched to the sides. He then transitions into a series of fluid, athletic movements. His right leg kicks forward while he maintains a balanced stance, showcasing agility and control. After the kick, he smoothly transitions back to a standing position, repeating the sequence with his left leg. Throughout the performance, the man's movements are precise and deliberate, emphasizing the fluidity and grace of the kick walking technique. His body language conveys strength and precision, highlighting the skill involved in executing such a complex movement.

Table 6. Samples of the detailed input prompts.

from backbone features and K/V from conditional embeddings. Stage-II further introduces an additional cross-attention module for response injection.

Training cost. Training the LoRAs requires approximately 789, 563, and 620 GPU hours for action video, stylization, and identity tasks, respectively. All training is conducted on 4 NVIDIA A100 80GB GPUs. Stage-I and Stage-II contain 199.3M and 213.4M parameters, respectively, requiring approximately 130 and 270 GPU hours for training, with peak memory usage of 56GB and 77GB.

Inference latency. The inference latency of LoFA is 1.7 s and 2.0 s per LoRA for Stage-I and Stage-II, respec-

tively, with memory usage of 30 GB and 32 GB, achieving throughputs of 8.43 and 9.39 TFLOPS. The reported latency corresponds to the time required to obtain the LoRAs. The subsequent LoRA injection and generation time are reported, as they are identical for ours and baselines, ensuring fairness.

Text prompts. Tab. 6 lists some samples of the text to prompt text-to-video tasks. It shows the *fine-grained* property of the prompts used to predict LoRA weights.

F. Generalization beyond expert models

In the core experiments, we follow prior work and train LoRAs using synthetic data generated by expert models. This setup may raise the concern that the hypernetwork merely imitates a specific expert. We therefore investigate whether LoFA can go beyond individual experts by leveraging data synthesized from multiple sources, improving generalization. We finetune our LoFA with 500 extra LoRAs trained from Seniorita [?] outputs, and validate using *real-world* user prompts from MidJourney. Fig. 3 shows our method further gains with additional data source.

G. Limitation and Future Work.

A key limitation of our method is that handling different domain-specific prompts—such as those for human actions, identities, or artistic styles—currently requires training separate networks. The ideal solution is a *single, unified* hypernetwork with strong zero-shot capabilities. Given that our architecture has demonstrated promising scalability, we are confident that scaling up the quantity and diversity of training data will enable the development of such a model in the future.

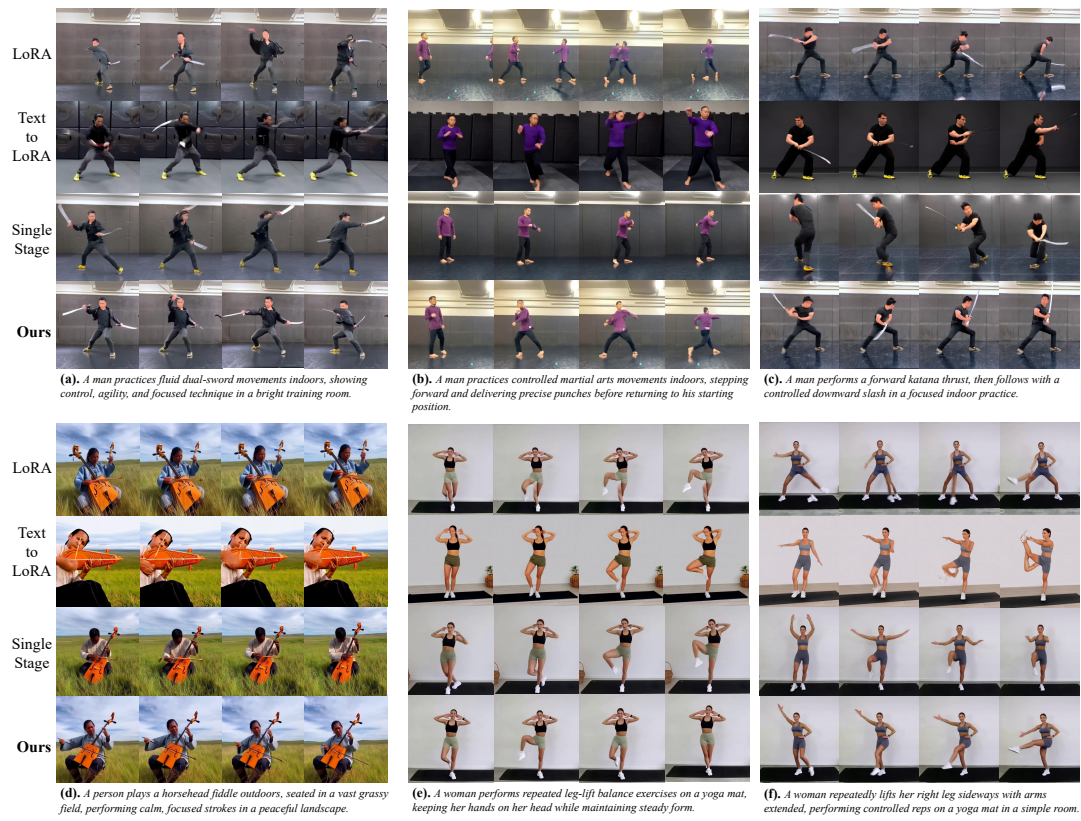


Figure 4. More qualitative results on **text-conditioned Personalized Human Action Video Generation**.

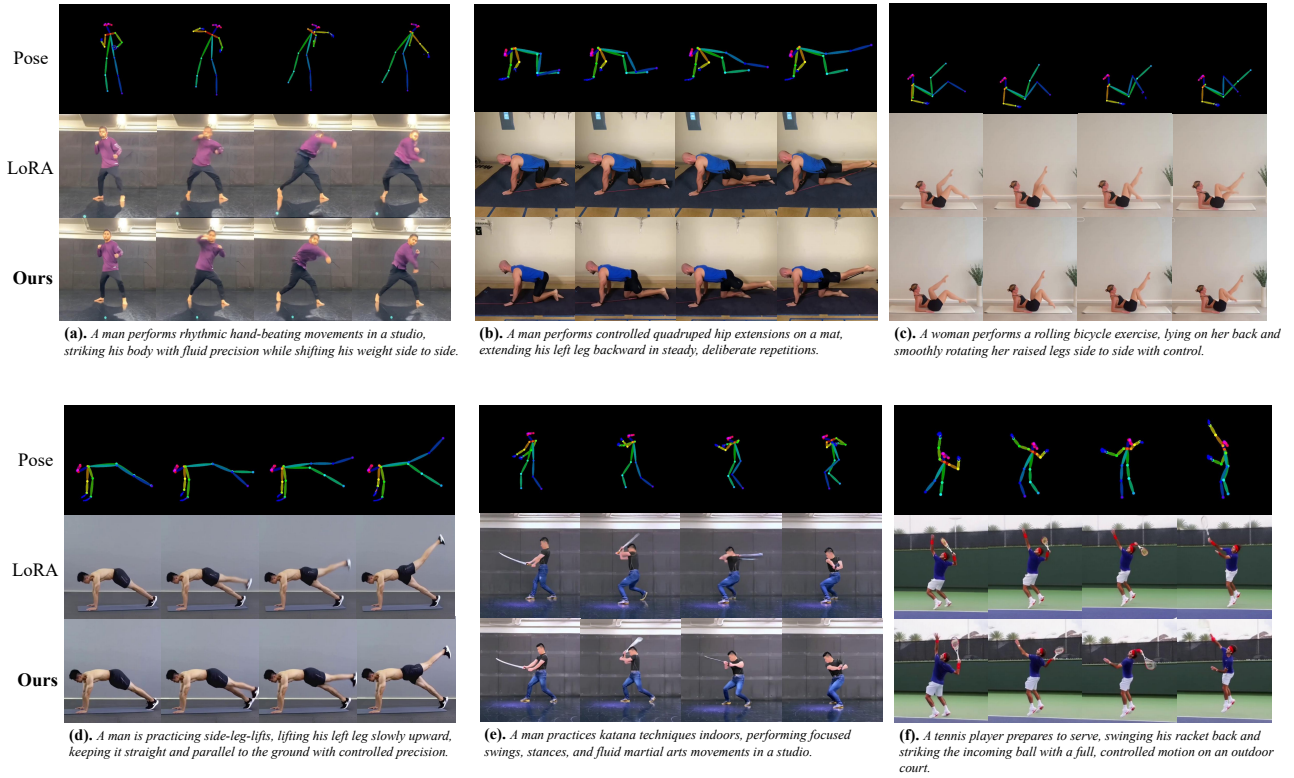


Figure 5. More qualitative results on **pose-conditioned Personalized Human Action Video Generation**.

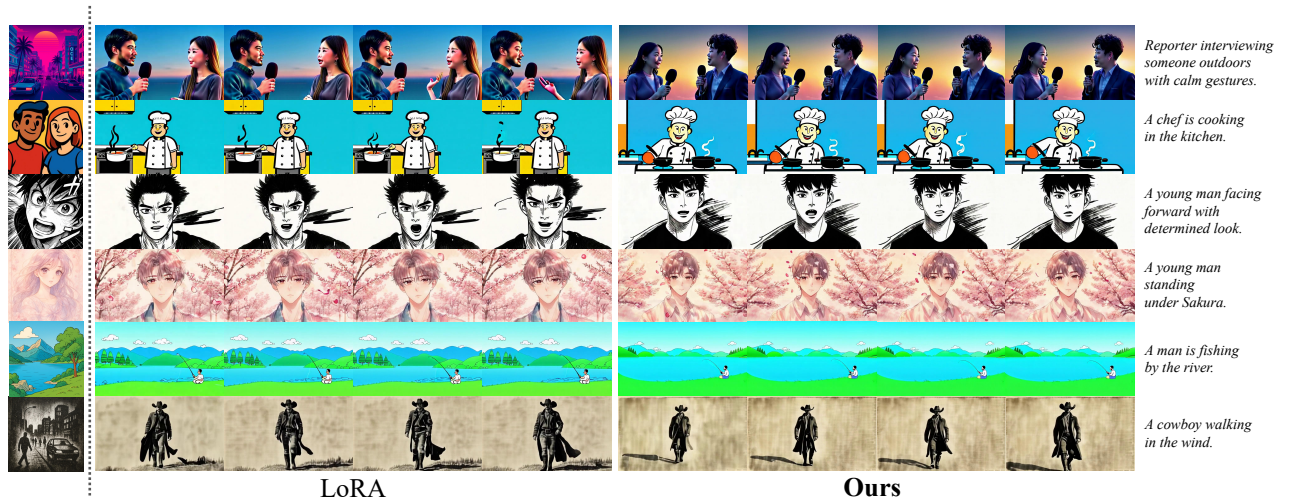


Figure 6. More qualitative results on **Text-to-Video Stylization**.

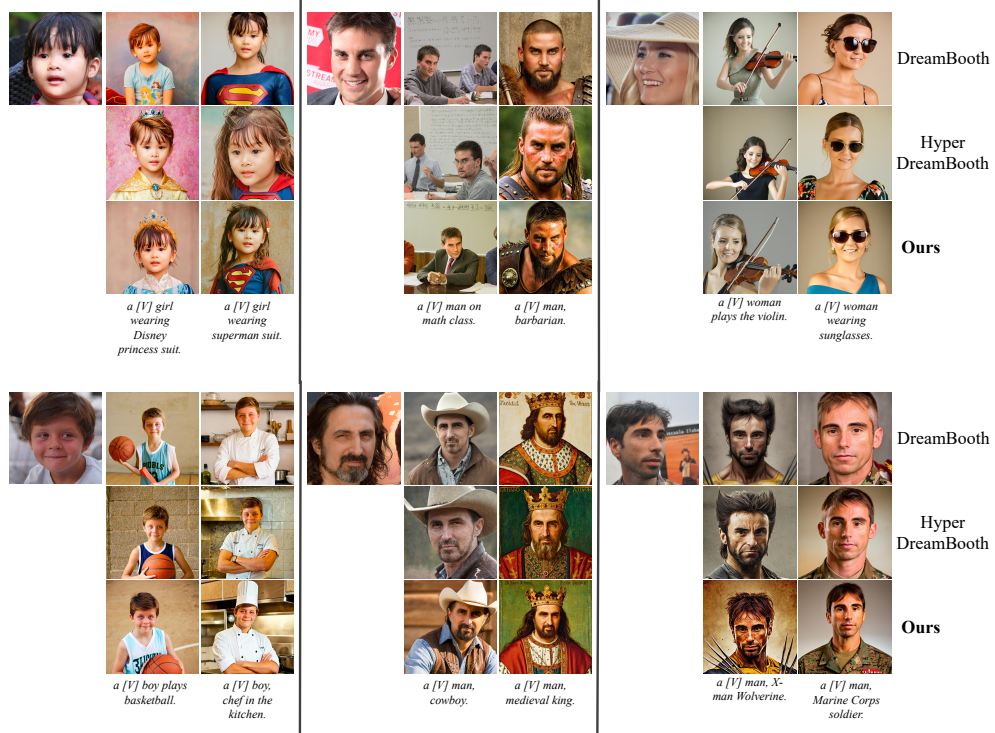


Figure 7. More qualitative results on **Identity-Personalized Image Generation**.