Table 6. Quantitative comparison of Stage III application and varying pruning ratios on WanX-TI2V. Applying Stage III, which performs coordinated distillation of sampling step and model size, results in slight declines across metrics like i2v subject, subject consistency, motion smoothness, and temporal flickering. Crucially, however, it achieves significant improvements in dynamic degree, aesthetic quality, and imaging quality. We also observe that larger pruning ratios correlate with poorer model performance overall.

Stage III	sampling steps	pruning ratio	i2v subject	subject consistency	motion smoothness	dynamic degree	aesthetic quality	imaging quality	temporal flickering	average
✗	28	30%	**0.961**	**0.970**	**0.991**	0.268	0.655	0.711	**0.984**	0.791
✓	4	30%	0.960	0.966	0.987	**0.378**	**0.658**	**0.712**	0.977	**0.805**
✗	28	50%	**0.958**	**0.964**	**0.988**	0.174	0.632	0.692	**0.989**	0.771
✓	4	50%	0.957	0.958	0.971	**0.273**	**0.649**	**0.701**	0.976	**0.784**
✗	28	70%	**0.959**	**0.959**	**0.965**	0.064	0.552	0.591	**0.983**	0.725
✓	4	70%	0.955	0.958	0.953	**0.132**	**0.579**	**0.619**	0.970	**0.738**

Table 7. Quantitative results for different pruning ratios on HunyuanVideo-ATI2V. As the pruning ratio increases, we observe a significant degradation in model performance, specifically in dynamic degree, aesthetic quality, and imaging quality.

Pruning Ratio	i2v subject	subject consistency	motion smoothness	dynamic degree	aesthetic quality	imaging quality	temporal flickering	average
0%	0.965	**0.984**	0.993	**0.696**	**0.603**	**0.722**	**0.982**	**0.849**
30%	**0.966**	0.972	**0.996**	0.622	0.571	0.693	**0.982**	0.829
50%	0.948	0.979	**0.996**	0.501	0.538	0.553	0.977	0.785
70%	0.949	0.967	0.993	0.249	0.465	0.519	0.952	0.728

A. Benchmarks and Hyperparameter Settings

Our Benchmark. Our benchmarking strategy is twofold: we use the generic VBench-I2V [14] to evaluate WanX-I2V and a customized VBench-I2V to assess HunyuanVideo-ATI2V. Regarding the generic VBench-I2V, the original VBench [14] was designed exclusively for T2V generators and could not assess I2V performance. Consequently, VBench-I2V [15] was developed as an extension, building upon the foundational VBench framework. This enhancement adds over 1118 new text-image pairs, facilitating a comprehensive evaluation of I2V generators across multiple dimensions—such as visual quality, motion dynamics, and subject consistency—with all testing conducted at a 16:9 aspect ratio. Our customized VBench-I2V is specifically designed for testing digital human scenarios. We collected 58 widescreen images featuring real people and anime characters and generated corresponding prompts using InternVL-26B. Similar to the generic VBench-I2V, we employed a 16:9 aspect ratio and repeated each sample five times to mitigate random errors.

Hyperparameter Settings for Training. All experiments were conducted on 16 NVIDIA H100 GPUs (2 nodes of 8) with a per-GPU batch size of 1, resulting in a global batch size of 16. We employed DeepSpeed ZeRO-3 and AdamW with CPU offloading to mitigate GPU memory overhead. For Stage II, we set the learning rate to $1e-5$ and trained for 4000 iterations. For Stage III, the learning rate was $5e-7$ and training ran for 1000 iterations. Specifically, Stage II consumes approximately 64 GPU days, while Stage III requires approximately 16 GPU days. The video memory footprint for both stages remains under 80GB. This shorter duration for Stage III was chosen empirically, as we observed that training beyond this point produced overly abrupt video motions and oversaturated colors. The hyperparameters $(\alpha, \beta_1, \beta_2)$ were set to $(1, 2.0, 0.25)$ for the HunyuanVideo-TI2V experiments and $(1, 3.5, 0.25)$ for the WanX-TI2V experiments.

Hyperparameter Settings for Evaluation. For the general-purpose VBench-I2V benchmark, we adopted the official evaluation prompts and the standard 16:9 resolution. On our customized VBench-I2V benchmark, identical prompts and a 16:9 resolution were used for all evaluations. To account for generative stochasticity, we generated five distinct videos for each prompt using different random seeds.

B. Additional Ablation Study

The Importance of Stage III. Stage III is critical, as it performs the truly coordinated distillation of both sampling steps and model size. As demonstrated in Table 6, without this stage, FastLightGen remains an underperforming pruned model; its

Table 8. Quantitative results comparison of β_1 and β_2 on HunyuanVideo-ATI2V. Fixing β_2 at 0.25 yields the best overall performance when β_1 is set β_1 to 2.0. With β_1 fixed at 2.0, we found that higher values for β_2 (e.g., 0.5,0.75,1.0) improved aesthetic and image quality but significantly degraded lipsyncing. Therefore, we select β_2=0.25 as it provides the best trade-off between these two objectives.

Inter CFG (β_1)	Intra CFG (β_2)	i2v subject	subject consistency	motion smoothness	dynamic degree	aesthetic quality	imaging quality	temporal flickering	average	lipsyncing
Fixed $\beta_1 = 2.0$:										
2.0	0.00	0.947	0.941	0.992	**0.667**	0.585	0.687	0.986	**0.829**	1.642
2.0	0.25	0.962	0.967	0.994	0.512	0.586	0.702	0.991	0.816	**1.967**
2.0	0.50	0.962	0.962	0.994	0.217	0.584	0.693	0.990	0.771	1.702
2.0	1.00	**0.966**	**0.970**	**0.996**	0.000	**0.587**	**0.704**	**0.995**	0.745	1.670
Fixed $\beta_2 = 0.25$:										
1.0	0.25	**0.997**	0.959	**0.995**	0.052	0.585	0.695	**0.991**	0.753	1.851
2.0	0.25	0.962	**0.967**	0.994	0.512	**0.586**	**0.702**	**0.991**	0.816	**1.967**
3.0	0.25	0.942	0.950	0.992	0.744	0.573	0.699	0.987	0.841	1.551
4.0	0.25	0.963	0.962	0.994	0.755	0.570	0.688	**0.991**	**0.846**	1.496
5.0	0.25	0.924	0.931	0.992	**0.837**	0.561	0.678	0.986	0.844	1.250

Table 9. Quantitative results on HunyuanVideo-ATI2V. This ablation investigates the choice of teacher model for the real DiT and the fake DiT in Stage III. We find that employing a pruned model as the teacher significantly outperforms using the unpruned version. This result suggests that the optimal teacher is not necessarily the most powerful, but rather one whose capacity is better matched to the student.

Teacher Setting	i2v subject	subject consistency	motion smoothness	dynamic degree	aesthetic quality	imaging quality	temporal flickering	average
Pruned	0.947	**0.941**	**0.992**	0.667	**0.585**	**0.687**	**0.986**	**0.829**
Unpruned	**0.956**	0.932	0.989	**0.670**	0.525	0.612	0.980	0.809

28-step sampling performance is markedly inferior to that of the final 4-step distilled model.

Pruning Ratio. As shown in Fig. 1, FastLightGen achieves an optimal balance between acceleration and performance at a 70% parameter retention rate with 4 distillation steps. The contour plot indicates that performance metrics decline sharply when retention falls below this 70% threshold, as shown by the denser contour lines. This finding is corroborated by the results in Tables 6 and 7. For instance, Table 6 demonstrates that pruning beyond this 30% limit (i.e., < 70% retention) causes a severe degradation in dynamic degree, aesthetic quality, and overall fidelity, leading to blurry and uncontrollable outputs.

Teacher Category. In Table 9, we provide further empirical evidence supporting the assertion that "stronger teachers do not necessarily benefit students more" On HunyuanVideo-ATI2V, we compare initializing both components with the unpruned model (i.e., a 'strong teacher') versus initializing both with the pruned model (i.e., a 'weak teacher'). The results show that the 'weak teacher' setup (using pruned models) outperforms the ·strong teacher' setup (using unpruned models) on the overall average score in our customized VBench-I2V evaluation. Therefore, our proposed well-guided teacher guidance is effective precisely because it constructs a teacher model that is optimally suited to the student.

C. Analysis of β_2

In Stage III, the hyperparameter β_2 controls the teacher model's interpolation between its unpruned and pruned states. Increasing β_2 biases the teacher toward the unpruned model, while decreasing β_2 biases it toward the pruned model. However, a 'stronger' teacher (i.e., one closer to the unpruned model) does not necessarily yield superior distillation results, as illustrated in Fig. 7. Specifically, setting $\beta_2 = 0$ (using the fully pruned model) results in overly static videos with poor motion quality. Conversely, setting $\beta_2 = 1$ (using the fully unpruned model) produces videos with drastic changes and uncontrollable content. Therefore, an appropriate intermediate value for β_2 is required to construct an optimal teacher for student model distillation.

D. Ethics Statement

We propose FastLightGen, a method for enabling coordinated step and size optimization in large-scale VDMs. The publicly available data used for this method is sourced from online animation clips and authorized public portrait videos. All UGC

Figure 7. In WanX-TI2V, the β_2 hyperparameter modulates the dynamic quality of the generated video. A value of $\beta_2 = 0$ yields overly static results, whereas $\beta_2 = 1$ introduces abrupt temporal discontinuities and excessive visual artifacts.

data is filtered using an automated model to ensure compliance with safety standards and ethical guidelines. We emphasize the responsible deployment and ethical regulation of this technology, aiming to maximize social benefits while proactively monitoring and mitigating potential risks.

E. Additional Visualization

Here, we present further visualizations of FastLightGen trained with optimal parameters, as shown in Figs. 8, 9 and 10.

Figure 8. Additional visualization of FastLightGen (i.e., 4-step generator that retains 70% of the parameters).

Figure 9. Additional visualization of FastLightGen (i.e., 4-step generator that retains 70% of the parameters).

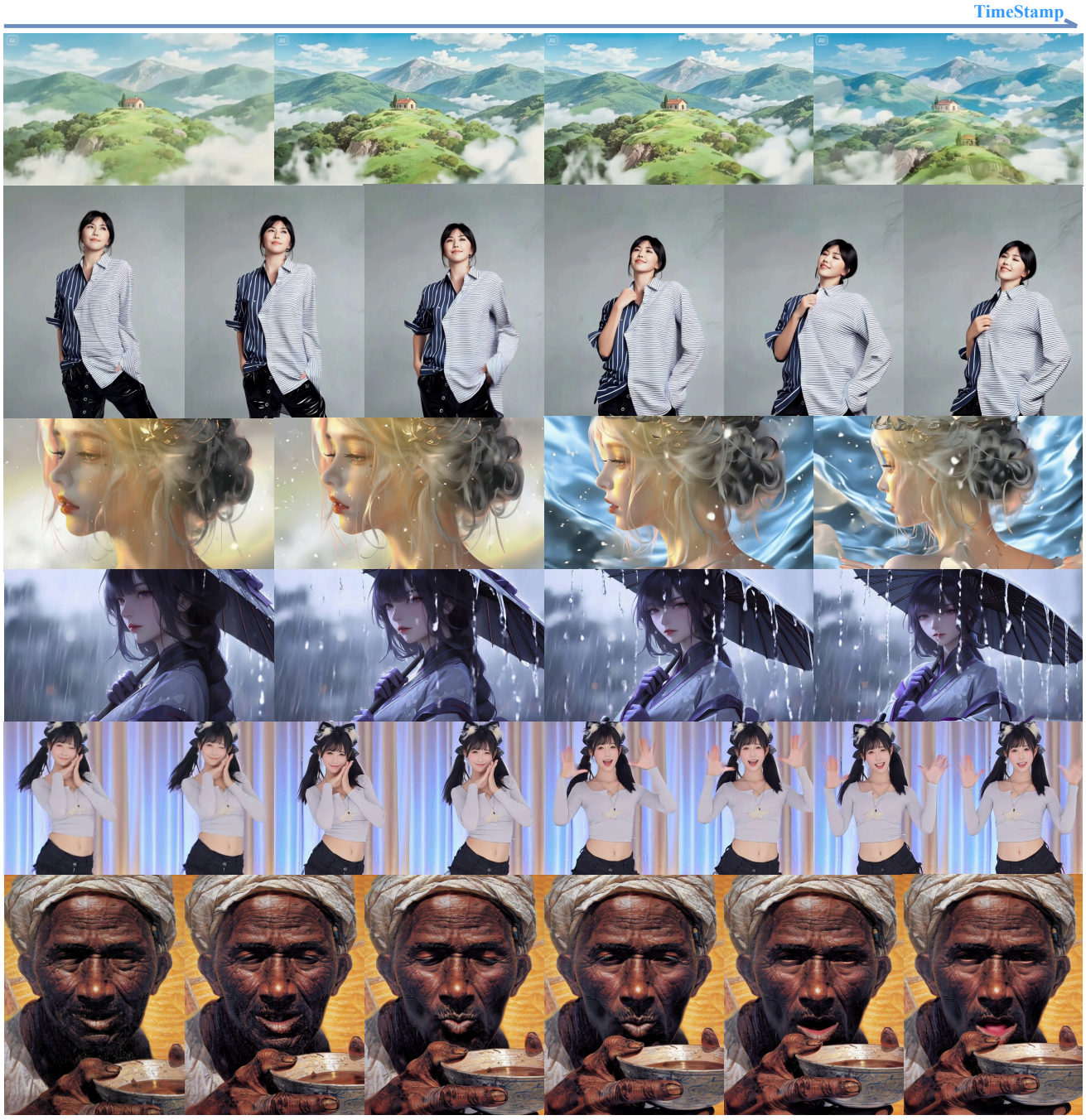

Figure 10. Additional visualization of FastLightGen (i.e., 4-step generator that retains 70% of the parameters).