Supplemtary Material for PYoCo (Preserve Your Own Correlation): A Noise Prior for Video Diffusion Models

Songwei Ge
University of Maryland

Seungjun Nah
NVIDIA

Guilin Liu
NVIDIA

Tyler Poon
University of Chicago

Andrew Tao
NVIDIA

Bryan Catanzaro
NVIDIA

David Jacobs
University of Maryland

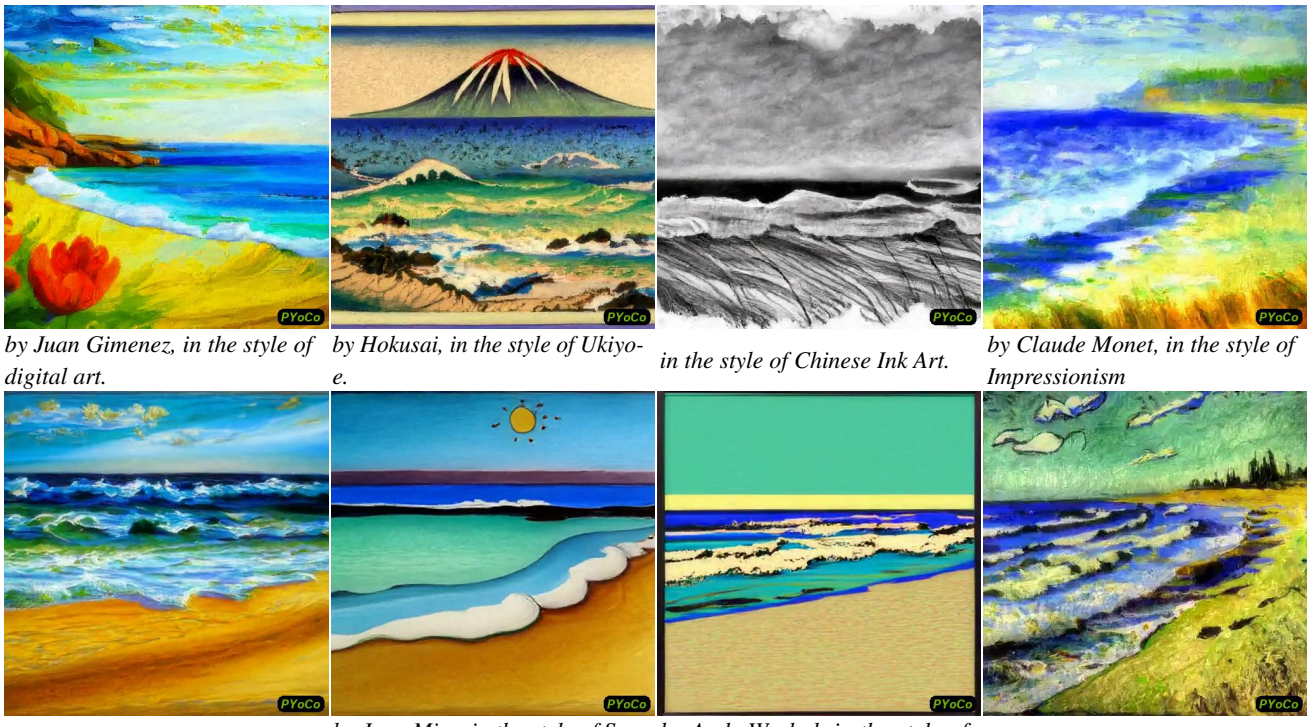
Jia-Bin Huang
University of Maryland

Ming-Yu Liu
NVIDIA

Yogesh Balaji
NVIDIA

In this supplementary material, we provide additional experimental results and details. In section A, we show the videos generated by our model with various styles, compositionality, and random seeds. We also provide the Inception Score plots for the ablation experiments on the UCF-101 dataset. In section B, we provide additional details on the implementation, training, inference, dataset, and evaluation.

A. Additional Results



by Juan Gimenez, in the style of digital art. | *by Hokusai, in the style of Ukiyo-e.* | *in the style of Chinese Ink Art.* | *by Claude Monet, in the style of Impressionism*

in the style of Baroque. | *by Joan Miro, in the style of Surrealism.* | *by Andy Warhol, in the style of Pop Art.* | *by Vincent van Gogh.*

Figure A: **Videos with different styles** generated by our approach for the caption *"a beautiful coastal beach in spring, waves lapping on sand X"*, where X is the style description provided under each video. *The figure is best viewed with Acrobat Reader. Click the images to play video clips.*

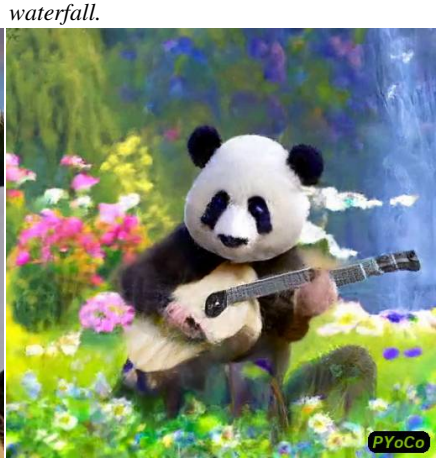
a cunning fluffy fox playing guitar in a boat on the ocean.

a cunning fluffy fox playing guitar nearby a campfire, snow mountain in the background.

a cunning fluffy fox playing guitar on the meadow full of flowers, in front of a graceful waterfall.

a happy fuzzy panda playing guitar in a boat on the ocean.

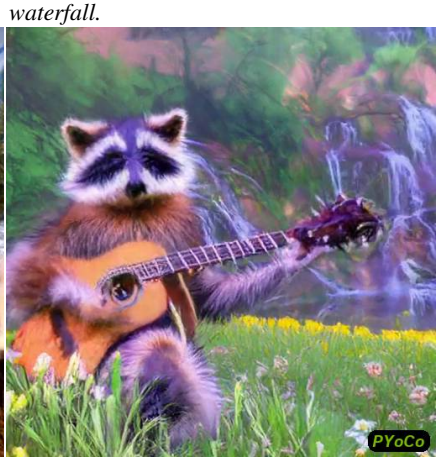
a happy fuzzy panda playing guitar nearby a campfire, snow mountain in the background.



a happy fuzzy panda playing guitar on the meadow full of flowers, in front of a graceful waterfall.

a cute raccoon playing guitar in a boat on the ocean.

a cute raccoon playing guitar nearby a campfire, snow mountain in the background.



a cute raccoon playing guitar on the meadow full of flowers, in front of a graceful waterfall.

Figure B: **Videos with different compositions** generated by our approach for the caption *"[PROTAGONIST] playing guitar [LOCATION]"*, where three different protagonists (fox, panda, raccoon) and three different locations (boat, snow mountain, meadow) are provided as inputs. Our model can faithfully generate videos consistent with the input text in each category. *The figure is best viewed with Acrobat Reader. Click the images to play video clips.*

Figure C: **Variations generated by our approach.** We generate four samples for the caption *"A cute tabby cat is doing homework on the desk"* using different random seeds. Our model can generate videos of cats in diverse backgrounds exhibiting different motions. *The figure is best viewed with Acrobat Reader. Click the images to play video clips.*

A.1. Additional ablation study

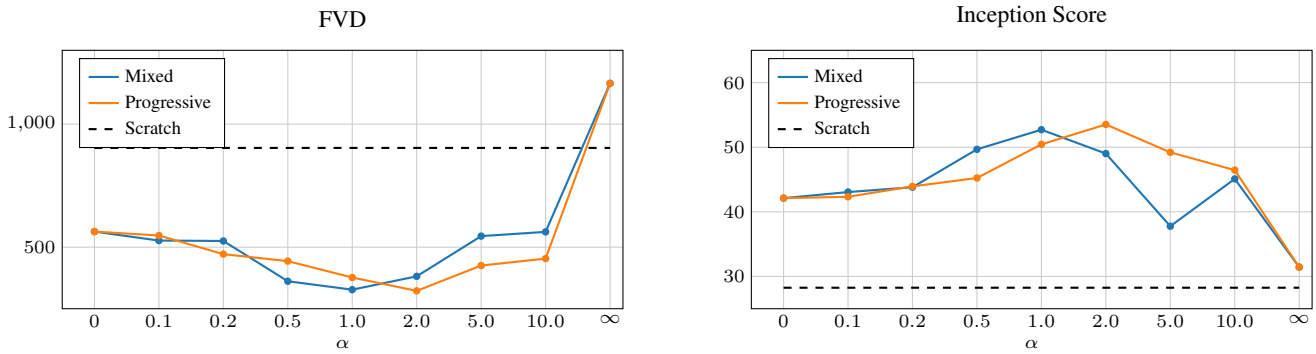


Figure D: **Ablation on hyperparameter α.** Finetuning with temporally correlated prior improves over training from scratch. Using too large or too small value for α leads to inferior results. $\alpha = 1$, $\alpha = 2$ each works the best for mixed and progressive noising, respectively.

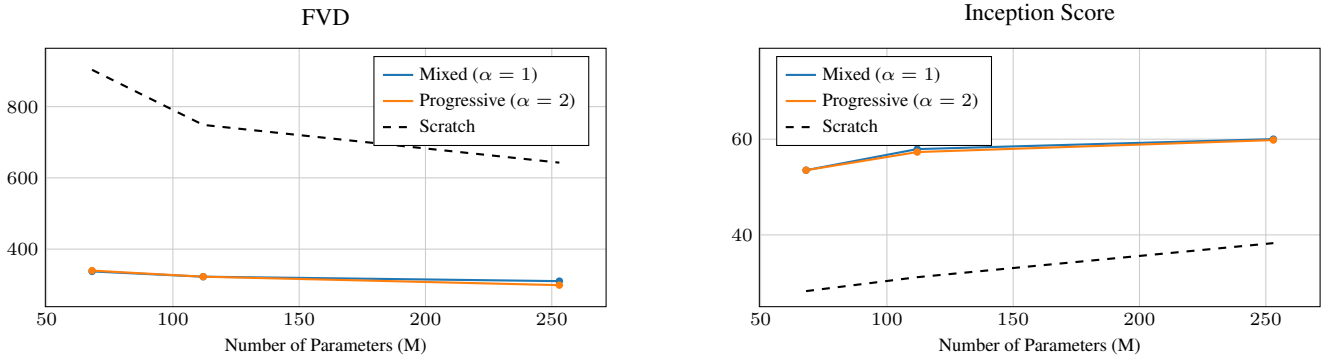


Figure E: **Ablation on model size.** Larger models consistently improve the performance of both finetuning and training from scratch. Finetuning consistently outperforms training from scratch.

In the main manuscript, we provided ablation study to better understand the behavior of our proposed noising schemes for video diffusion model under different circumstances. We show that the mixed and progressive noising schemes achieve improved FVDs in various configurations. In addition to the FVD metric, we also present that a consistent trend is observed in the Inception Score metric. In Figure D, we show the performance of our noising schemes by sweeping the hyperparameter α that controls the correlation among different frames in the noising schemes. In Figure E, we show the video quality in terms of both FVD and inception score with various model capacity compared with the baseline.

B. Experimental Setups

In this section, we provide additional details of our experiments in terms of implementation, dataset, evaluation, model, and training.

B.1. Implementation details

Similar to prior works [2, 7], we adapt the image-based U-Net model for the video synthesis task by making the following changes: (1) We transform the 2D convolution layers to 3D by adding a dimension of 1 to the temporal axis. For instance, we convert a 3×3 convolution layer to $1 \times 3 \times 3$ layer. (2) We replace the attention layers in the base and temporal interpolation models with a cascade of spatial and temporal attention layers. The spatial attention layers are reused from eDiff-I [1], while the temporal attention layers are initialized randomly with a projection layer at the end using zero-initialization. We apply temporal attention to the activation maps obtained by moving the spatial dimension of the feature tensor to the batch axis. (3) For the temporal interpolation model, we concatenate the input noise in the channel axis with 16 frames by infilling 4 real frames with zero frames. (4) We add a $3 \times 1 \times 1$ convolution layer at the end of each efficient block of the super-resolution model [6]. (5) For all the models, we apply spatial attention to the reshaped activation maps obtained by moving the temporal dimension of the feature tensor to the batch axis. We apply the same operation to the feature maps input the GroupNorm [8] to mimic better the statistics the image model learned. We use cross-attention layers (between text and videos) only in the spatial attention block, as adding it to the temporal attention resulted in significant memory overhead. (6) We utilize eDiff-I [1] to initialize our base and spatial super-resolution models. We use a similar model architecture as the base model for our temporal interpolation model, as they share the same function of hallucinating unseen frames. After finetuning the base model for some time, we use its checkpoint to initialize the temporal interpolation model. (7) Similar to Ho *et al*. [2], we jointly finetune the model on video and image datasets by concatenating videos and images in the temporal axis and applying our temporal modules only on the video part. (8) Similarly to eDiff-I, our model uses both T5 [5] text embeddings and CLIP text embeddings [4]. During training, we drop each of the embeddings independently at random, as in eDiff-I.

B.2. Dataset and evaluation details

Caption templates for categorical video datasets Given the name of the category [*class*] such as *kayaking* and *yoga*, we consider the following templates to create video captions:

- a man is [*class*].
- a woman is [*class*].
- a kid is [*class*].
- a group of people are [*class*].
- doing [*class*].
- a man is doing [*class*].
- a woman is doing [*class*].
- a kid is doing [*class*].
- a group of people are doing [*class*].
- [*class*].

Prompts used for UCF-101 evaluation In our initial explorations, we find that the original class labels in the UCF-101 dataset often cannot describe the video content correctly. For example, the class *jump rope* is more likely describing an object rather than a complete video. Therefore, we write one sentence for each class as the caption for video generation. We list these prompts for evaluating text-to-video generation models on the standard UCF-101 benchmark below.

applying eye makeup, applying lipstick, archery, baby crawling, gymnast performing on a balance beam, band marching, baseball pitcher throwing baseball, a basketball player shooting basketball, dunking basketball in a basketball match, bench press, biking, billiards, blow dry hair, blowing candles, body weight squats, a person bowling on bowling alley, boxing punching bag, boxing speed bag, swimmer doing breast stroke, brushing teeth, weightlifting with barbell, clean and jerk, cliff diving, bowling in cricket gameplay, batting in cricket gameplay, cutting in kitchen, diver diving into a swimming pool from a springboard, drumming, two fencers have fencing match indoors, field hockey match, gymnast performing on the floor, group of people playing frisbee on the playground, swimmer doing front crawl, golfer swings and strikes the ball, haircuting, a person hammering a nail, an athlete performing the hammer throw, an athlete doing handstand push up, an athlete doing handstand walking, massagist doing head massage to man, an athlete doing high jump, horse race, group of people racing horse, person riding a horse, a woman doing hula hoop, man and woman dancing on the ice, ice dancing, athlete practicing javelin throw, a person juggling with balls, a young person doing jumping jacks, a person skipping with jump rope, a person kayaking in rapid water, knitting, an athlete doing long jump, a person doing lunges with barbell, military parade, mixing in the kitchen, mopping floor, a person practicing nunchuck, gymnast performing on parallel bars, a person tossing pizza dough, a musician playing the cello in a room, a musician playing the daf, a musician playing the indian dhol, a musician playing the flute, a musician playing the guitar, a musician playing the piano, a musician playing the sitar, a musician playing the tabla, a musician playing the violin,

an athlete jumps over the bar, gymnast performing pommel horse exercise, a person doing pull ups on bar, boxing match, push ups, group of people rafting on fast moving river, rock climbing indoor, rope climbing, several people rowing a boat on the river, couple salsa dancing, young man shaving beard with razor, an athlete practicing shot put throw, a teenager skateboarding, skier skiing down, jet ski on the water, sky diving, soccer player juggling football, soccer player doing penalty kick in a soccer match, gymnast performing on still rings, sumo wrestling, surfing, kids swing at the park, a person playing table tennis, a person doing TaiChi, a person playing tennis, an athlete practicing discus throw, trampoline jumping, typing on computer keyboard, a gymnast performing on the uneven bars, people playing volleyball, walking with dog, a person standing, doing pushups on the wall, a person writing on the blackboard, a kid playing Yo-Yo

B.3. Training details

UCF-101 experiments. For image pretraining phase on the UCF-101 frames, we use an ADAM optimizer with a base learning rate of $2e-4$. For video finetuning phase, we adopt an ADAM optimizer with a base learning rate of $1e-4$. We use a linear warm up of $5,000$ steps for both phases. For sampling, we use stochastic DEIS sampler [9, 3] with 3kutta, order 6 and 25 steps.

Large-scale experiments. The hyper-parameters we use for the large-scale text-to-video experiments are provided in Table A.

Run time. To sample a video using the settings in Tables E - H, it takes 16.9, 3.0, 1.5, and 15.6 mins for each stage on a single NVIDIA A100 GPU.

Table A: Hyperparameters

Hyperparameters for large-scale experiments	
Optimizer	AdamW
Learning rate	0.0001
Weight decay	0.01
Betas	(0.9, 0.999)
EMA	0.9999
CLIP text embedding dropout rate	0.2
T5 text embedding dropout rate	0.25
Gradient checkpointing	Enabled
# iterations for base model	$150K$
# iterations for super-res model	$220K$
Sampler for base model	Stochastic DEIS [9, 3], 3kutta, Order 3, 60 steps
Sampler for super-res models	DEIS, 3kutta Order 3, 20 steps

B.4. Architecture details

The architectures used for the small-scale UCF experiments are provided in Tables B, C and D. For the large-scale experiment, the architectures used for base model, temporal interpolation model, and the two spatial super-resolution stacks are provided in tables E, F, G and H respectively.

Table B: Small (69M parameters) UCF-101 model architecture.

Small (69M parameters) UCF-101 model	
Channel multiplier	$[1, 2, 2, 3]$
Dropout	0.1
Number of channels	128
Number of residual blocks	2
Spatial self attention resolutions	$[32, 16, 8]$
Spatial cross attention resolutions	$[32, 16, 8]$
Temporal attention resolution	$[32, 16, 8]$
Number of channels in attention heads	64
Use scale shift norm	True

Table C: Medium (112M parameters) UCF-101 model architecture.

Medium (112M parameters) UCF-101 model	
Channel multiplier	$[1, 2, 3, 4]$
Dropout	0.1
Number of channels	128
Number of residual blocks	2
Spatial self attention resolutions	$[32, 16, 8]$
Spatial cross attention resolutions	$[32, 16, 8]$
Temporal attention resolution	$[32, 16, 8]$
Number of channels in attention heads	64
Use scale shift norm	True

Table D: Large (253M parameters) UCF-101 model architecture.

Large (253M parameters) UCF-101 model	
Channel multiplier	$[1, 2, 3, 4]$
Dropout	0.1
Number of channels	192
Number of residual blocks	2
Spatial self attention resolutions	$[32, 16, 8]$
Spatial cross attention resolutions	$[32, 16, 8]$
Temporal attention resolution	$[32, 16, 8]$
Number of channels in attention heads	64
Use scale shift norm	True

Table E: Architecture for the base model in text-to-video experiments.

Text-to-video base model (1.08B parameters)	
Channel multiplier	$[1, 2, 4, 4]$
Dropout	0
Number of channels	256
Number of residual blocks	3
Spatial self attention resolutions	$[32, 16, 8]$
Spatial cross attention resolutions	$[32, 16, 8]$
Temporal attention resolution	$[32, 16, 8]$
Number of channels in attention heads	64
Use scale shift norm	True

Table F: Architecture for the temporal interpolation model in text-to-video experiments.

Temporal interpolation model (1.08B parameters)	
Channel multiplier	$[1, 2, 4, 4]$
Dropout	0
Number of channels	256
Number of residual blocks	3
Spatial self attention resolutions	$[32, 16, 8]$
Spatial cross attention resolutions	$[32, 16, 8]$
Temporal attention resolution	$[32, 16, 8]$
Number of channels in attention heads	64
Use scale shift norm	True

Table G: Architecture for the spatial super-resolution model in text-to-video experiments.

Spatial super-resolution 256 (300M parameters)	
Channel multiplier	$[1, 2, 4, 8]$
Block multiplier	$[1, 2, 4, 4]$
Dropout	0
Number of channels	128
Number of residual blocks	2
Spatial self attention resolutions	$[32]$
Spatial cross attention resolutions	$[32]$
Number of channels in attention heads	64
Use scale shift norm	True

Table H: Architecture for the spatial super-resolution model in text-to-video experiments.

Spatial super-resolution 1024 (170M parameters)	
Patch size	256×256
Channel multiplier	$[1, 2, 4, 4]$
Block multiplier	$[1, 2, 4, 4]$
Number of channels	128
Number of residual blocks	2
Spatial cross attention resolutions	$[32]$
Use scale shift norm	True

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. eDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 4

[2] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 4

[3] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022. 5

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4

[5] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140):1–67, 2020. 4

[6] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 4

[7] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 4

[8] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, pages 3–19, 2018. 4

[9] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. In *ICLR*, 2023. 5