A1. Experimental Details

A1.1. Sampling Algorithm

Algorithm 1: Spatiotemporal Skip Guidance (STG)

Input: ϵ_θ, $\epsilon_\theta^{s,t}$: Main model and spatiotemporally perturbed model respectively.
w: Spatiotemporal guidance scale.
Σ_t: Variance at step t.
Output: Generated video V_{out}.

$x_T \sim \mathcal{N}(0, I)$
for $t \leftarrow T, T-1, \ldots, 1$ **do**
  $\epsilon_t \leftarrow \epsilon_\theta(x_t)$ $\epsilon_t^{s,t} \leftarrow \epsilon_\theta^{s,t}(x_t)$ $\tilde{\epsilon}_t \leftarrow \epsilon_t + w(\epsilon_t - \epsilon_t^{s,t})$ $x_{t-1} \sim \mathcal{N}\left(\frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\tilde{\epsilon}_t\right), \Sigma_t\right)$
return V_{out}

Algorithm 2: Spatiotemporal Skip Guidance (STG) for factorized attention

Input: ϵ_θ, ϵ_θ^s, ϵ_θ^t: Main model, spatially perturbed, and temporally perturbed models respectively.
w_1, w_2: Guidance scales.
Σ_t: Variance at step t.
Output: Generated video V_{out}.

$x_T \sim \mathcal{N}(0, I)$
for $t \leftarrow T, T-1, \ldots, 1$ **do**
  $\epsilon_t \leftarrow \epsilon_\theta(x_t)$ $\epsilon_t^s \leftarrow \epsilon_\theta^s(x_t)$ $\epsilon_t^t \leftarrow \epsilon_\theta^t(x_t)$ $\tilde{\epsilon}_t \leftarrow \epsilon_t + w_1(\epsilon_t - \epsilon_t^s) + w_2(\epsilon_t - \epsilon_t^t)$
  $x_{t-1} \sim \mathcal{N}\left(\frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\tilde{\epsilon}_t\right), \Sigma_t\right)$
return V_{out}

A1.2. Computational Resources

For evaluation, we utilized an NVIDIA A100 40GB GPU for SVD [4], while Open-Sora [38] and Mochi [33] were evaluated using NVIDIA H100 or A100 80GB GPUs.

A1.3. Implementation Details

The default model scales for CFG are as follows: SVD [4] uses a scale of 3.0, Open-Sora [38] uses 7.0, and Mochi [33] uses 4.5. For STG, the configurations vary, with STG-A using a scale of 2.0 and STG-R using 1.0. STG is applied to the 8^{th} layer of SVD, which has a total of 16 layers, and the 12^{th} layer of Open-Sora, which has a total of 28 layers. For Mochi, which has 48 layers in total, STG is applied at the 35^{th} layer.

A1.4. Metrics

To evaluate model performance across different datasets, several methodologies were employed. FVD was assessed using the VideoMAE [8] model. IS was evaluated with the C3D model [16, 17, 34], following the setup of TGAN [7]. For VBench - Imaging Quality, the MUSIQ image quality predictor, trained on the SPAQ dataset, was used. VBench - Aesthetic Quality was measured using the LAION aesthetic predictor, applied to individual video frames. VBench - Dynamic Degree was evaluated with the RAFT flow estimator to quantify the degree of dynamics. VBench - Motion Smoothness was calculated as the mean absolute error (MAE) between dropped and reconstructed frames using a video frame interpolation model. Finally, VBench - Temporal Flickering was assessed by generating static frames and computing the mean absolute difference between consecutive frames.

A1.5. Prompts Used

EvalCrafter prompt

1. *2 Dog and a whale, ocean adventure*
2. *Teddy bear and 3 real bear*
3. *Goldfish in glass*
4. *A small bird sits atop a blooming flower stem.*
5. *A fluffy teddy bear sits on a bed of soft pillows surrounded by children's toys.*
6. *A peaceful cow grazing in a green field under the clear blue sky.*
7. *Unicorn sliding on a rainbow*
8. *Four godzillas*
9. *A fluffy grey and white cat is lazily stretched out on a sunny window sill, enjoying a nap after a long day of lounging.*
10. *A curious cat peers from the window, watching the world outside.*
11. *A horse*
12. *A pig*
13. *A squirrel*
14. *A bird*
15. *A zebra*
16. *Two elephants are playing on the beach and enjoying a delicious beef stroganoff meal.*
17. *Two fish eating spaghetti on a subway*
18. *A pod of dolphins gracefully swim and jump in the ocean.*
19. *A peaceful cow grazing in a green field under the clear blue sky.*
20. *A cute and chubby giant panda is enjoying a bamboo meal in a lush forest. The panda is relaxed and content as it eats, and occasionally stops to scratch its ear with its paw.*
21. *Dragon flying over the city at night*
22. *Pikachu snowboarding*
23. *A cat drinking beer*
24. *A dog wearing VR goggles on a boat*
25. *A giraffe eating an apple*
26. *Five camels walking in the desert*
27. *Mickey Mouse is dancing on white background*
28. *A happy pig rolling in the mud on a sunny day.*
29. *In an African savanna, a majestic lion is prancing behind a small timid rabbit. The rabbit tried to run away, but the lion catches up easily.*
30. *3 sheep enjoying spaghetti together*
31. *A photo of a Corgi dog riding a bike in Times Square. It is wearing sunglasses and a beach hat.*
32. *A pod of dolphins gracefully swim and jump in the ocean.*
33. *In the lush forest, a tiger is wandering around with a vigilant gaze while the birds chirp and monkeys play.*
34. *The teddy bear and rabbit were snuggled up together. The teddy bear was hugging the rabbit, and the rabbit was nuzzled up to the teddy bear's soft fur.*
35. *A slithering snake moves through the lush green grass.*
36. *A pair of bright green tree frogs cling to a branch in a vibrant tropical rainforest.*
37. *Four fluffy white Persian kittens snuggle together in a cozy basket by the fireplace.*
38. *Eight fluffy yellow ducklings waddle behind their mother, exploring the edge of a pond.*
39. *A family of four fluffy, blue penguins waddled along the icy shore.*
40. *Two white swans gracefully swam in the serene lake.*
41. *In a small forest, a colorful bird was flying around gracefully. Its shiny feathers reflected the sun rays, creating a beautiful sight.*
42. *A spider spins a web, weaving intricate patterns with its silk.*
43. *...*

Curated prompt for demos

1. *Sloth with pink sunglasses lays on a donut float in a pool. The sloth is holding a tropical drink. The world is tropical. The sunlight casts a shadow.*
2. *A vibrant, top-down video of a kayak gliding through multicolored waters, showcasing shifting hues from blue to red to illustrate varying flow speeds or temperatures. The paddle interacts with the water, creating dynamic ripples and currents.*
3. *A slow-motion capture of a beautiful woman in a flowing dress spinning in a field of sunflowers, with petals swirling around her.*
4. *An exotic video of rabbits on the moon making rice cakes under a star-filled sky, with Earth visible in the background.*
5. *A handsome man walking confidently through a bustling city street at night, illuminated by neon lights and reflections in puddles.*
6. *A close-up shot of a butterfly landing on the nose of a woman, highlighting her smile and the intricate details of the butterfly's wings.*
7. *A top-down video of a table filled with colorful dishes from different cuisines, with hands reaching in to serve food and clinking glasses.*
8. *A majestic bird's-eye view of a couple holding hands while walking along the shore of a beach with sparkling turquoise waves.*
9. *A surreal scene of a forest where the leaves glow in neon colors, with a person walking down a path as fireflies dance around them.*
10. *A drone shot of a desert at sunset, where shadows stretch and shift, capturing a lone traveler moving gracefully through the sand dunes.*
11. *A close-up of a beautiful woman's face with colored powder exploding around her, creating an abstract splash of vibrant hues.*
12. *A panoramic view of a tropical waterfall surrounded by lush greenery, with a rainbow forming in the mist.*
13. *A whimsical video of floating lanterns being released into the sky over a calm lake, with reflections on the water creating a mirror effect.*
14. *A time-lapse video of an artist painting a mural on a city wall, where each frame shows a burst of color and detail.*
15. *An overhead video of koi fish swimming in a pond with rippling water, with their scales reflecting shades of gold, orange, and white.*
16. *A slow-motion clip of a handsome person diving into a crystal-clear ocean, with water splashing and bubbles forming intricate patterns.*
17. *A fantastical scene of a meadow where flowers bloom and change colors in sync with the music, with a person dancing among them.*
18. *A top-down video of a hot air balloon festival, showing multicolored balloons lifting off and dotting the sky.*
19. *A beautiful woman sitting by a window as rain drizzles down, creating streaks and patterns on the glass.*
20. *A cinematic shot of a person walking through a field of lavender during golden hour, with the wind gently swaying the purple blossoms.*
21. *An exotic video of floating jellyfish in the ocean, their translucent bodies glowing with bioluminescence in shades of blue and purple.*
22. *A playful video of puppies running across a vibrant, flower-filled meadow, filmed in slow motion to capture their joyful expressions.*
23. *A captivating aerial view of a cityscape at sunrise, with skyscrapers casting long shadows and golden light reflecting on windows.*
24. *A stunning slow-motion shot of a bird taking flight over a reflective lake, with water droplets glistening as they scatter.*
25. *A romantic scene of a couple dancing under string lights in a backyard, with warm, golden tones highlighting their laughter.*
26. *...*

Prompt for figures in the main paper

Fig2:

- *A macro cinematography animation showing a butterfly emerging from its chrysalis, filmed with side-lit lighting to accentuate the texture of its wings.*

Fig4:

- *A 50mm lens shot of a couple embracing under string lights as the camera slowly tracks them, capturing their shared laughter in a soft, cinematic glow.*
- *An animation showing a floating castle drifting above the clouds, with birds flying around it and sunlight casting golden rays, evoking the feeling of wonder seen in classic animations.*
- *A realistic documentary-style video of artisans crafting pottery, with the scene unfolding and transforming as hands shape clay under diffused lighting.*
- *A ghost in a white bedsheet faces a mirror. The ghost's reflection can be seen in the mirror. The ghost is in a dusty attic, filled with old beams, cloth-covered furniture. The attic is reflected in the mirror. The light is cool and natural. The ghost dances in front of the mirror.*

A2. Theoretical justification for layer skipping

In this section, we provide additional theoretical justification for layer skipping. We adopt an energy-based perspective using Hopfield energy. The modern Hopfield network [28] retrieves relevant information given a query, with its energy function defined as:

$$E(\boldsymbol{\xi}) = -\text{lse}(\beta, \boldsymbol{X}^T\boldsymbol{\xi}) + \frac{1}{2}\boldsymbol{\xi}^T\boldsymbol{\xi} + C, \tag{22}$$

where $\boldsymbol{\xi} \in \mathbb{R}^d$ is the query, $\boldsymbol{X} \in \mathbb{R}^{N\times d}$ represents stored patterns to be retrieved (keys), $\beta > 0$ and lse is the *log-sum-exp function*. Lower energy indicates better query states for retrieval. The update rule $\boldsymbol{\xi}^{\text{t+1}} = f(\boldsymbol{\xi}^t) = \boldsymbol{X}^T\text{softmax}(\beta\boldsymbol{X}\boldsymbol{\xi}^t)$ monotonically decreases energy and is proven to converge globally [28]. Notably, the attention mechanism is equivalent to a single Hopfield update step [28] when $\boldsymbol{X} = \boldsymbol{K} = \boldsymbol{W}_K\boldsymbol{Y}$, $\boldsymbol{\xi} = \boldsymbol{q} = \boldsymbol{W_Q}\boldsymbol{Y_i}$, and $\beta = 1/\sqrt{d_k}$. Even with a single update step per attention layer, the retrieval quality improves exponentially [28]. Therefore, the forward pass of the transformers can be interpreted as iterative energy minimization, where each attention layer refines the retrieved patterns (e.g., DiT tokens). Now we formally state that skipping layers disrupts the energy minimization:

Lemma 1 (Skipping Layers Prevents Energy Reduction)**.** *Let E^{l^-} and E^l be the Hopfield energy before and after layer l. The total energy for a model with L layers is: $E_{total} = \sum_{i=1}^{L} E^i$. If layer l is skipped, the energy becomes: $E_{total}^{skip} = \sum_{i\neq l}^{L} E^i + E^{l^-}$. Then, $E_{total} < E_{total}^{skip}$.*

Proof. It suffices to show that $E^l < E^{l^-}$ for any l. Since attention acts as a Hopfield update, which is the Concave-Convex Procedure (CCCP) [28], it monotonically decreases energy. Hence, $E^l < E^{l^-}, \forall l$.

□

Now recall the Boltzmann distribution in probabilistic models:

$$P(\boldsymbol{\xi}) = \frac{1}{Z}\exp(-\beta E(\boldsymbol{\xi})), \tag{23}$$

where the higher energy states correspond to a lower probability (Z is the partition function ensuring normalization). Since skipping a layer increases energy (Lemma 1), it results in retrieved feature representations with lower likelihood, leading to lower-likelihood samples (**Conjecture**: Low-likelihood features produce low-likelihood samples).

Additionally, layer skipping helps preserve in-distribution features, as consecutive DiT layers exhibit high feature similarity [24] due to residual connections. By treating in-distribution but low-likelihood samples as negative predictions, *STG effectively guides sampling away from them, enhancing generation quality.*

A3. Layer selection

Layer selection is straightforward and efficient: we directly assess the VBench score across different layers (Fig. 6), requiring only a few evaluations to find optimal layers. The selected layers perform consistently better across diverse samples. STG on later layers tends to yield better results, since these layers are primarily related to texture details.

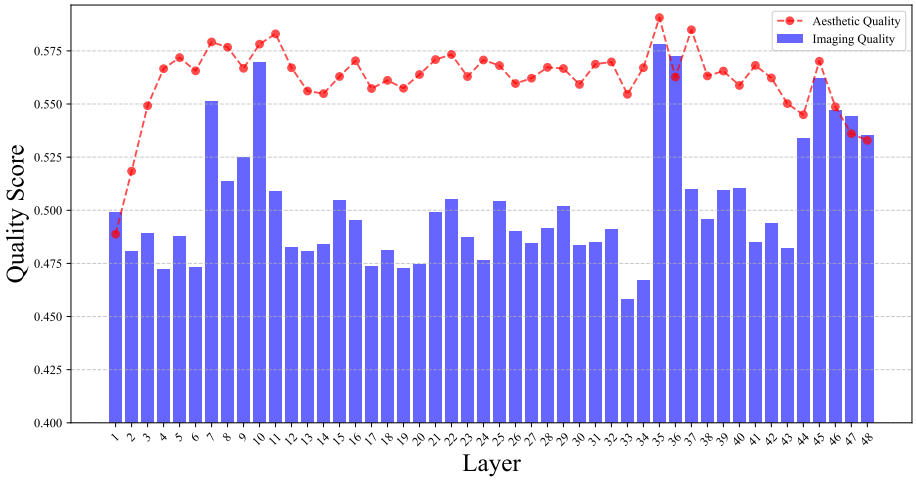


Figure 6. Ablation study on layer selection - layer 35 is selected.

A4. Human Evaluation

We conducted user studies following EvalCrafter [22] to evaluate subjective opinions across five key aspects: (1) Video Quality, reflecting the clarity of the generated video, with higher scores indicating reduced blur, noise, or visual artifacts; (2) Text-Video Alignment, assessing the correspondence between the input text prompt and the generated video, particularly focusing on the accuracy of generated motions; (3) Motion Quality, evaluating the correctness and realism of the motions depicted in the video; (4) Temporal Consistency, measuring the frame-to-frame coherence, distinct from Motion Quality as it requires users to assess the smoothness of movement; and (5) Subjective Likeness, akin to an aesthetic score, where higher values signify better alignment with human preferences.

For each metric, feedback was collected from seven users, who rated videos on a scale from 1 to 5, with higher scores representing better alignment. To ensure fairness, the video sequences were randomly shuffled before being presented to users.

We used 700 prompts from EvalCrafter for text-to-video (T2V) generation with Mochi [33]. Additionally, we employed FLUX.1 [dev] [20] to generate images from these prompts, which served as input to the image-to-video (I2V) model (SVD [4]). The results, shown in Fig. 7, demonstrate that incorporating STG leads to improved quality across all evaluated aspects.

A5. Ablation Study

A5.1. Manifold Constrained Guidance

As discussed in the main paper, sampling guidance techniques, including STG, utilize scale guidance, which can sometimes cause the sampling trajectory to deviate from the data manifold. This deviation is particularly noticeable when STG is applied with large scales or to videos that are already bright, often resulting in broken videos or over-saturation due to manifold overshooting. To mitigate these issues, we propose a set of optional techniques that can serve as effective remedies.

First, we leverage the error contraction property of stochastic processes [36] by incorporating stochastic forward processes into the sampling guidance framework. This technique, referred to as **STG with Restart**, is detailed in Algorithm 10. While this method moderately enhances the quality of the final samples and resolves issues such as broken videos (as illustrated in Fig. 8), it introduces additional computational overhead.

Additionally, increased variance in the latent code [21] has been observed in over-saturated results. Consequently, over-saturation can be effectively mitigated using a rescaling technique [21], which constrains the variance of the latent code. This method, referred to as **STG with Rescaling**, is detailed in Algorithm 6. As shown in Fig. 9, videos generated with larger variance (second row) often display saturated colors, which are successfully resolved by applying variance rescaling (third row). Unlike the Restart method, Rescaling introduces negligible computational overhead, making it the preferred approach for addressing over-saturation.

Algorithm 3: Spatiotemporal Skip Guidance with Restart

Input: ϵ_θ, $\epsilon_\theta^{s,t}$: Main model and spatiotemporal perturbed model respectively.
w: Spatiotemporal guidance scale.
Σ_t: Variance at step t.
$t_{\text{min}}, t_{\text{max}}$: Restart interval.
K: Number of Restart iterations.
Output: Generated video V_{out}.

$x_T \sim \mathcal{N}(0, I)$
for $t \leftarrow T, T-1, \ldots, 1$ **do**
  $\epsilon_t \leftarrow \epsilon_\theta(x_t)\ \epsilon_t^{s,t} \leftarrow \epsilon_\theta^{s,t}(x_t)\ \tilde{\epsilon}_t \leftarrow \epsilon_t + w(\epsilon_t - \epsilon_t^{s,t})\ x_{t-1} \sim \mathcal{N}\left(\frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\tilde{\epsilon}_t\right), \Sigma_t\right)$
  if $t = t_{min}$ **then**
    $x_{t_{\text{min}}}^0 \leftarrow x_{t-1}$
    for $k \leftarrow 0, \ldots, K-1$ **do**
      $\epsilon_{\text{restart}} \sim \mathcal{N}(0, \Sigma_{\text{restart}})\ x_{t_{\text{max}}}^{k+1} \leftarrow x_{t_{\text{min}}}^k + \epsilon_{\text{restart}}$
      for $t' \leftarrow t_{max}, t_{max}-1, \ldots, t_{min}$ **do**
        $\epsilon_{t'} \leftarrow \epsilon_\theta(x_{t'}^{k+1})\ \epsilon_{t'}^{s,t} \leftarrow \epsilon_\theta^{s,t}(x_{t'}^{k+1})\ \tilde{\epsilon}_{t'} \leftarrow \epsilon_{t'} + w(\epsilon_{t'} - \epsilon_{t'}^{s,t})$
        $x_{t'-1}^{k+1} \sim \mathcal{N}\left(\frac{1}{\sqrt{\alpha_{t'}}}\left(x_{t'}^{k+1} - \frac{1-\alpha_{t'}}{\sqrt{1-\bar{\alpha}_{t'}}}\tilde{\epsilon}_{t'}\right), \Sigma_{t'}\right)$
return V_{out}

A5.2. STG with Orthogonalization

As discussed in the main paper, for SVD and Open-Sora, which utilize factorized spatial and temporal attention, it is possible to orthogonalize spatial and temporal guidance. The detailed algorithm for this approach is provided in Algorithm 8. However, we do not implement orthogonalization in practice, as it does not demonstrate any performance improvement, as shown in Table 5.

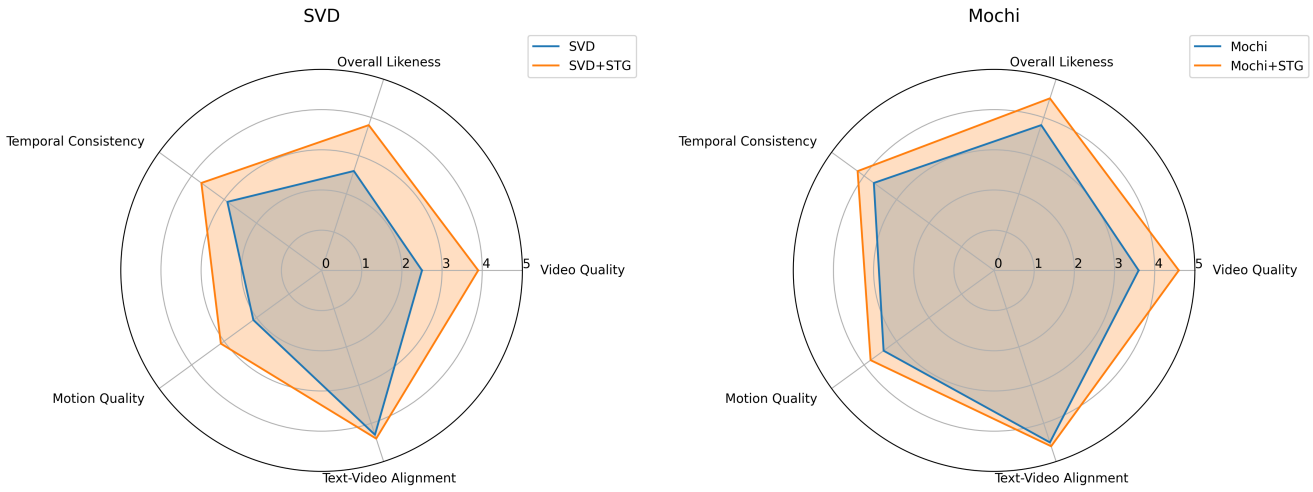


Figure 7. User study results for STG on SVD and Mochi, using 700 prompts from EvalCrafter [22]. For I2V generation of SVD, we employed FLUX.1 [dev] [20] to generate images from these prompts, which served as input to the model. The results demonstrate that incorporating STG leads to improved quality across all evaluated aspects.

Algorithm 4: Spatiotemporal Skip Guidance (STG) with Rescaling

Input: ϵ_θ, $\epsilon_\theta^{s,t}$: Main model and spatiotemporal perturbed model respectively.
w: Spatiotemporal guidance scale.
$rescale$: Rescaling factor.
Σ_t: Variance at step t.
Output: Generated video V_{out}.

$x_T \sim \mathcal{N}(0, I)$
for $t \leftarrow T, T-1, \ldots, 1$ **do**
  $\epsilon_t \leftarrow \epsilon_\theta(x_t)$ $\epsilon_t^{s,t} \leftarrow \epsilon_\theta^{s,t}(x_t)$ $\tilde{\epsilon}_t \leftarrow \epsilon_t + w(\epsilon_t - \epsilon_t^{s,t})$
  $\text{std}_\epsilon \leftarrow \text{std}(\epsilon_t)$ $\text{std}_{\tilde{\epsilon}} \leftarrow \text{std}(\tilde{\epsilon}_t)$ $\text{factor} \leftarrow \frac{\text{std}_\epsilon}{\text{std}_{\tilde{\epsilon}}}$ $\text{factor} \leftarrow rescale \cdot \text{factor} + (1 - rescale)$ $\tilde{\epsilon}_t \leftarrow \tilde{\epsilon}_t \cdot \text{factor}$
  $x_{t-1} \sim \mathcal{N}\left(\frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\tilde{\epsilon}_t\right), \Sigma_t\right)$
return V_{out}

Models	FVD (↓)	IS	Imaging Quality	Aesthetic Quality	Motion Smoothness	Dynamic Degree
SVD (STG)	**128.7**	**38.5**	**0.694**	**0.639**	**0.968**	**0.694**
SVD (STG-ORTH)	130.4	38.4	0.691	0.637	0.967	0.692

Table 5. Ablation results of STG on SVD [4], evaluating the impact of orthogonalizing spatial and temporal guidance (STG-ORTH). Our findings show no performance gain from applying orthogonalization; therefore, we do not adopt it.

Algorithm 5: Spatiotemporal Skip Guidance (STG) with Orthogonalization

Input: ϵ_θ, ϵ_θ^s, ϵ_θ^t: Main model, spatially perturbed, and temporally perturbed models respectively.
w_1, w_2: Guidance scales.
Σ_t: Variance at step t.
Output: Generated video V_{out}.

$x_T \sim \mathcal{N}(0, I)$
for $t \leftarrow T, T-1, \ldots, 1$ **do**
  $\epsilon_t \leftarrow \epsilon_\theta(x_t)$ $\epsilon_t^s \leftarrow \epsilon_\theta^s(x_t)$ $\epsilon_t^t \leftarrow \epsilon_\theta^t(x_t)$
  $\Delta_s \leftarrow \epsilon_t - \epsilon_t^s$ $\Delta_t \leftarrow \epsilon_t - \epsilon_t^t$
  $\Delta_t^\perp \leftarrow \Delta_t - \frac{\langle \Delta_s, \Delta_t \rangle}{\|\Delta_s\|^2} \cdot \Delta_s$
  $\tilde{\epsilon}_t \leftarrow \epsilon_t + w_1\Delta_s + w_2\Delta_t^\perp$
  $x_{t-1} \sim \mathcal{N}\left(\frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\tilde{\epsilon}_t\right), \Sigma_t\right)$
return V_{out}

A5.3. Layer Ablation

STG can be applied to different layers, and we conduct an ablation study to evaluate the impact of skipping various layers for STG on Mochi [33]. The results are presented in Fig. 10. Mochi consists of 48 layers in total, and we experimented with layer skipping at layers 30, 32, and 35. Our findings show that skipping later layers has a more significant effect on quality improvements, as these layers are primarily responsible for refining texture details. Throughout all experiments in this paper, we consistently skip layer 35.

A5.4. Effect of Spatial and Temporal Guidance

For models with factorized attention layers, guidance can be applied separately to spatial and temporal layers. When using STG-A, it functions similarly to applying PAG [1] to the spatial attention layers, and we refer to this method as Spatial PAG (SPAG). When spatial guidance is applied alone, as shown for SVD in Fig. 11 and for Open-Sora in Fig. 12, the results struggle to maintain clarity during motion and exhibit poor temporal consistency. For instance, significant artifacts appear near the wings in the second row of Fig. 11, and around the legs of the chicken in Fig. 12.

We further investigate the individual contributions of spatial and temporal guidance. In Fig. 13, we compare results with and without Spatial Guidance (SPAG). The results show that while CFG fails to maintain clear object structures, resulting in blurry videos, SPAG significantly enhances object structure and improves clarity.

Similarly, in Fig. 14, we present results with and without Temporal Guidance (TPAG). The results reveal that CFG struggles to ensure frame-to-frame consistency, with the shape and color of the jelly varying noticeably across frames, leading to a disjointed video. In contrast, TPAG effectively preserves the jelly's appearance throughout the sequence, creating a more cohesive video and significantly improving Temporal Consistency.

A5.5. Attention Skip and Residual Skip

We compare the performance of attention skip (STG-A) and residual skip (STG-R) in Mochi [33] and Open-Sora [38]. The results for Mochi in Fig. 15 indicate that STG-R delivers greater qualitative improvements for Mochi. On the other hand, the results for Open-Sora in Fig. 16 and SVD in Fig. 17 demonstrate that STG-A delivers greater qualitative improvements for these models. Based on these findings, we use STG-A for Open-Sora and SVD, and use STG-R for Mochi in all experiments presented in the paper.

A5.6. Weak Model Visualization

We visualize the results of one-step prediction using different denoising methods (weak models, CFG, and STG) at timestep 30 in Fig. 19 and timestep 24 in Fig. 20, rows (a) to (e). Row (c) shows results denoised using the spatiotemporally perturbed model, $\epsilon_\theta^{s,t}(x_t)$, which generally produces blurrier outcomes compared to row (b), where the unconditional weak model $\epsilon_\theta(x_t|\phi)$ of CFG is applied. By moving away from the blurry weak model, STG achieves clear and well-defined structures with natural color tones. In contrast, CFG often produces unnatural color artifacts and broken structures. For example, the video predicted by CFG renders the girl's arm on the left unnaturally red, the man's arm on the right unnaturally dark, and the trees and leaves in the background blurry. By comparison, STG consistently generates videos with enhanced structure and natural, well-balanced color tones.

A5.7. Other Perturbation Methods

In addition to SPAG (spatial perturbation using PAG), we explore other perturbation techniques. One such approach is SEG [13], which applies Gaussian blurring to the attention map. A comparison of CFG, SEG, and STG is presented in Fig. 18. The results frequently show broken outputs in both CFG and SEG. In contrast, incorporating layer skipping alongside temporal perturbation, as in STG, consistently produces improved results.

A6. Qualitative Comparison

We provide additional qualitative comparisons using STG for SVD, Open-Sora, and Mochi. The results demonstrate that applying STG enhances the aesthetic appeal and fidelity of the videos, as shown in Fig. 21. In Open-Sora, we observe flickering artifacts frequently in the videos. By applying STG, these flickering artifacts are noticeably reduced, as illustrated in Fig. 22.

For I2V models such as SVD, as discussed in the main paper, STG not only enhances the structural quality of the generated videos but also increases their dynamic degree. This is because STG mitigates the effect of CFG, which tends to force generated videos to rigidly adhere to the conditioning image. This effect is visualized in Fig. 23.

We provide more video results in the zip file.

Figure 8. Quality Improvement with Restart STG. *Top*: Results for SVD [4]. *Bottom*: Results for Mochi [33]. The results demonstrate that while STG occasionally fails to generate videos correctly in certain cases, applying Restart resolves these issues, producing high-quality and accurate outputs.

Prompt: A young woman with glasses is jogging in the park wearing a pink headband.

Figure 9. Comparison of CFG, STG, and Rescaled STG on Mochi [33]. When STG is applied using large scales or to bright videos, it often suffers from over-saturation caused by manifold deviation. One potential cause of this issue is the increased variance in the latent code, which is effectively mitigated by the rescaling technique proposed in [21].

Prompt: A close-up shot of a butterfly landing on the nose of a woman, highlighting her smile and the details of the butterfly's wings.

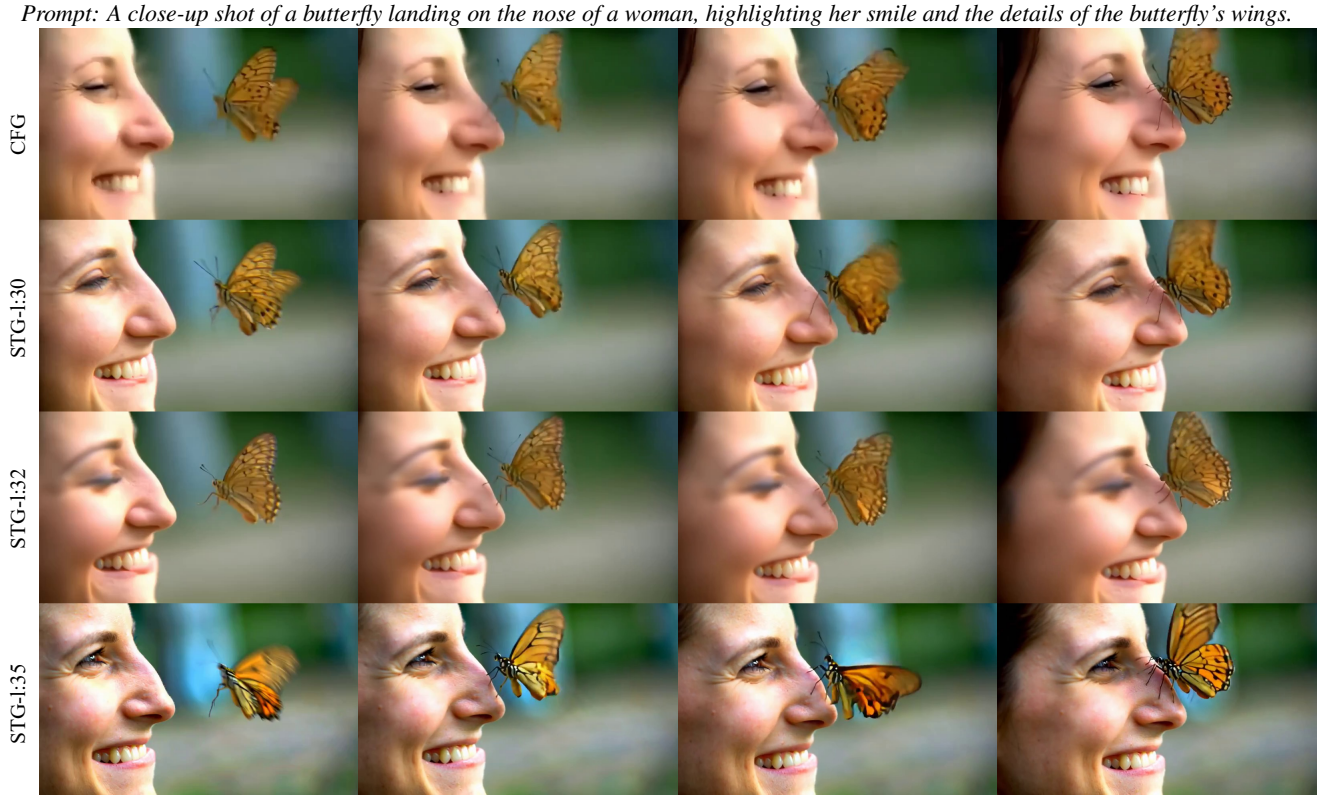


Figure 10. Ablation study on the effect of skipping different layers for STG on Mochi [33]. Our results indicate that skipping later layers has a greater impact on quality improvements, as these layers primarily contribute to texture details. For all experiments, we consistently skip layer 35 (denoted as STG-l:35).

(Image condition is given for SVD.)

Figure 11. Qualitative Comparison of CFG, SPAG, and STG on SVD [4]. PAG applied only to spatial layers is referred to as SPAG. The results show that while CFG and SPAG fail to preserve object clarity under motion, STG successfully achieves this.

Prompt: Brown chicken hunting for its food.

Figure 12. Qualitative Comparison of CFG, SPAG, and STG on Open-Sora [38]. The results show CFG fails to generate the object's head accurately, and SPAG struggles with the legs, whereas STG successfully generates all components correctly.

(Image condition is given for SVD.)

Figure 13. Qualitative Comparison of Object Structure in SVD [4] with and without Spatial Guidance. Spatial Guidance is represented by SPAG, which applies PAG only to the spatial layer. The results indicate that while CFG struggles to maintain clear object structures, leading to blurry videos, SPAG effectively enhances object structure and improves clarity.

(Image condition is given for SVD.)

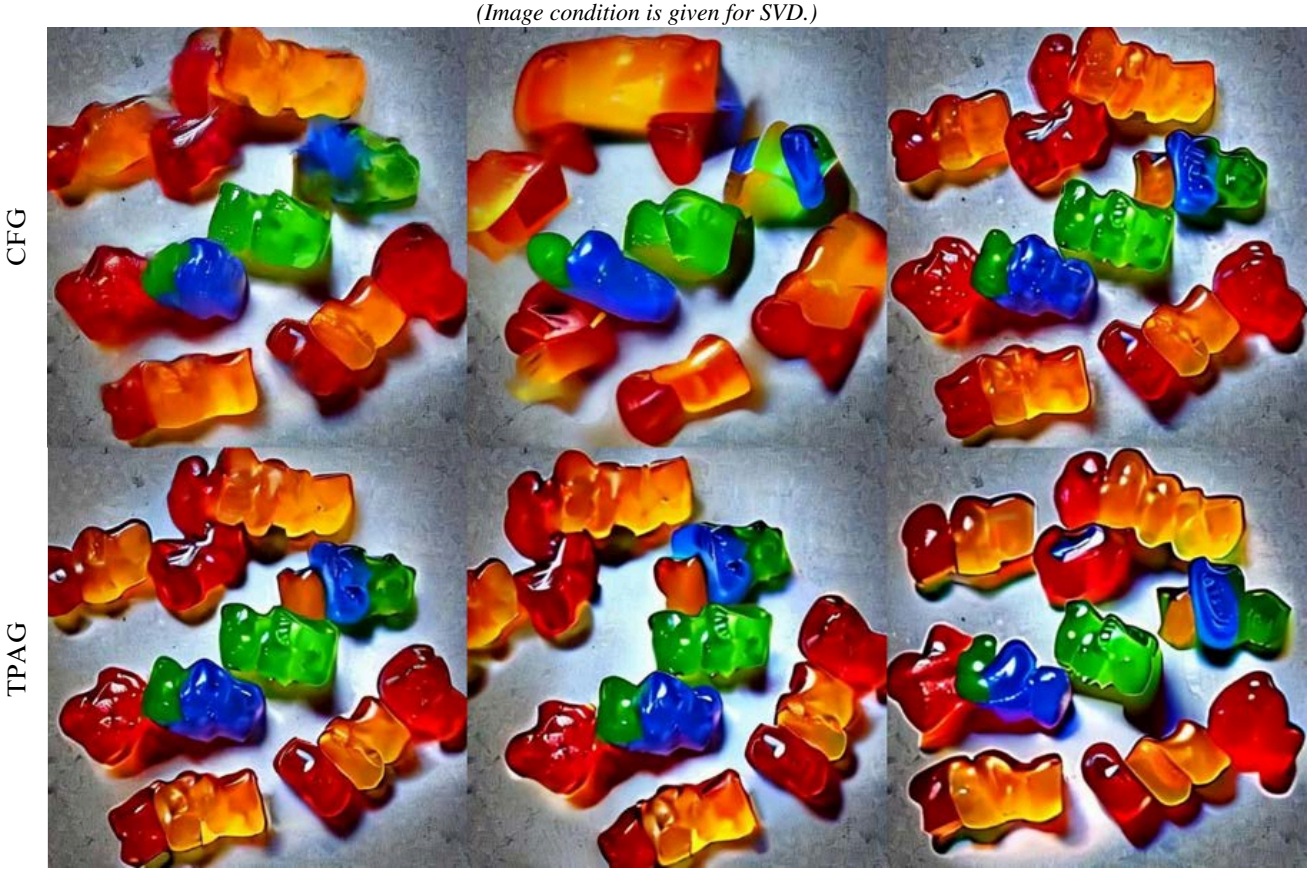
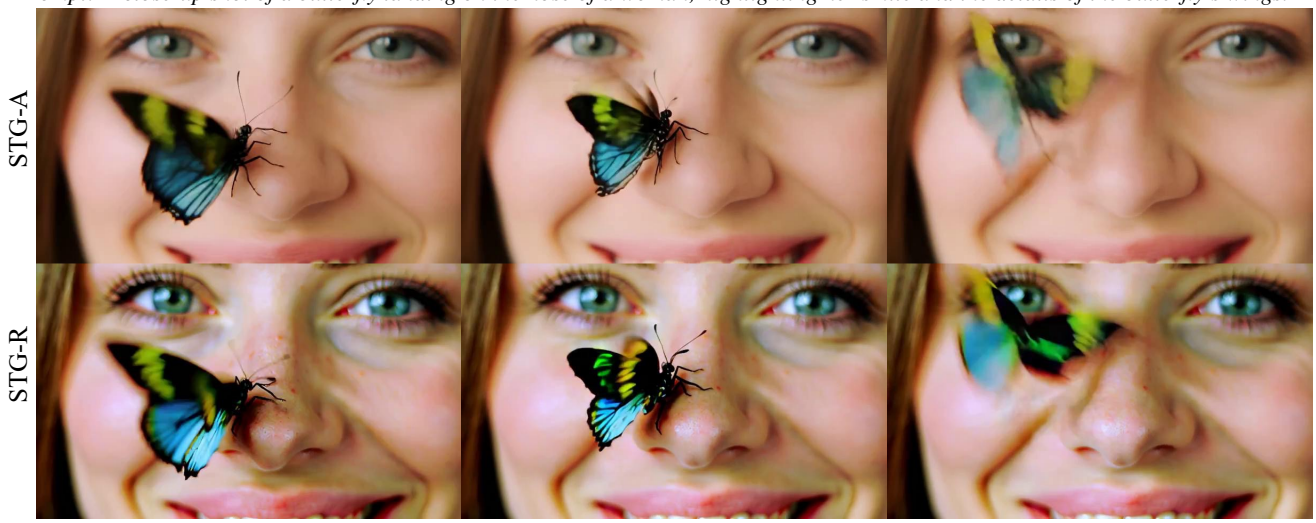


Figure 14. Qualitative Comparison of Temporal Consistency in SVD [4] with and without Temporal Guidance (TPAG). The results reveal that CFG struggles to ensure frame-to-frame consistency, with the shape and color of the jelly varying noticeably across frames, leading to a disjointed video. In contrast, TPAG effectively preserves the jelly's appearance throughout the sequence, creating a more cohesive video and significantly improving Temporal Consistency.

Prompt: A close-up shot of a butterfly landing on the nose of a woman, highlighting her smile and the details of the butterfly's wings.



Prompt: Cinematic 8k scene of a couple dancing under warmly glowing string lights in an intimate backyard setting, ...

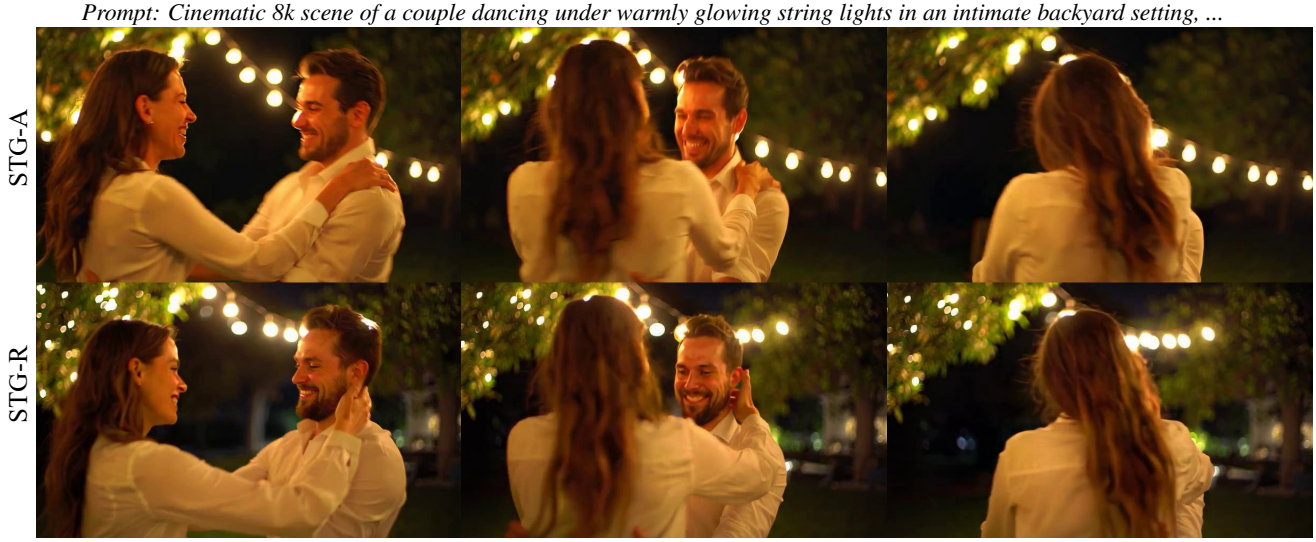


Figure 15. Comparison of attention skip (STG-A) and residual skip (STG-R) in Mochi [33]. The results indicate that STG-R delivers greater qualitative improvements for Mochi.

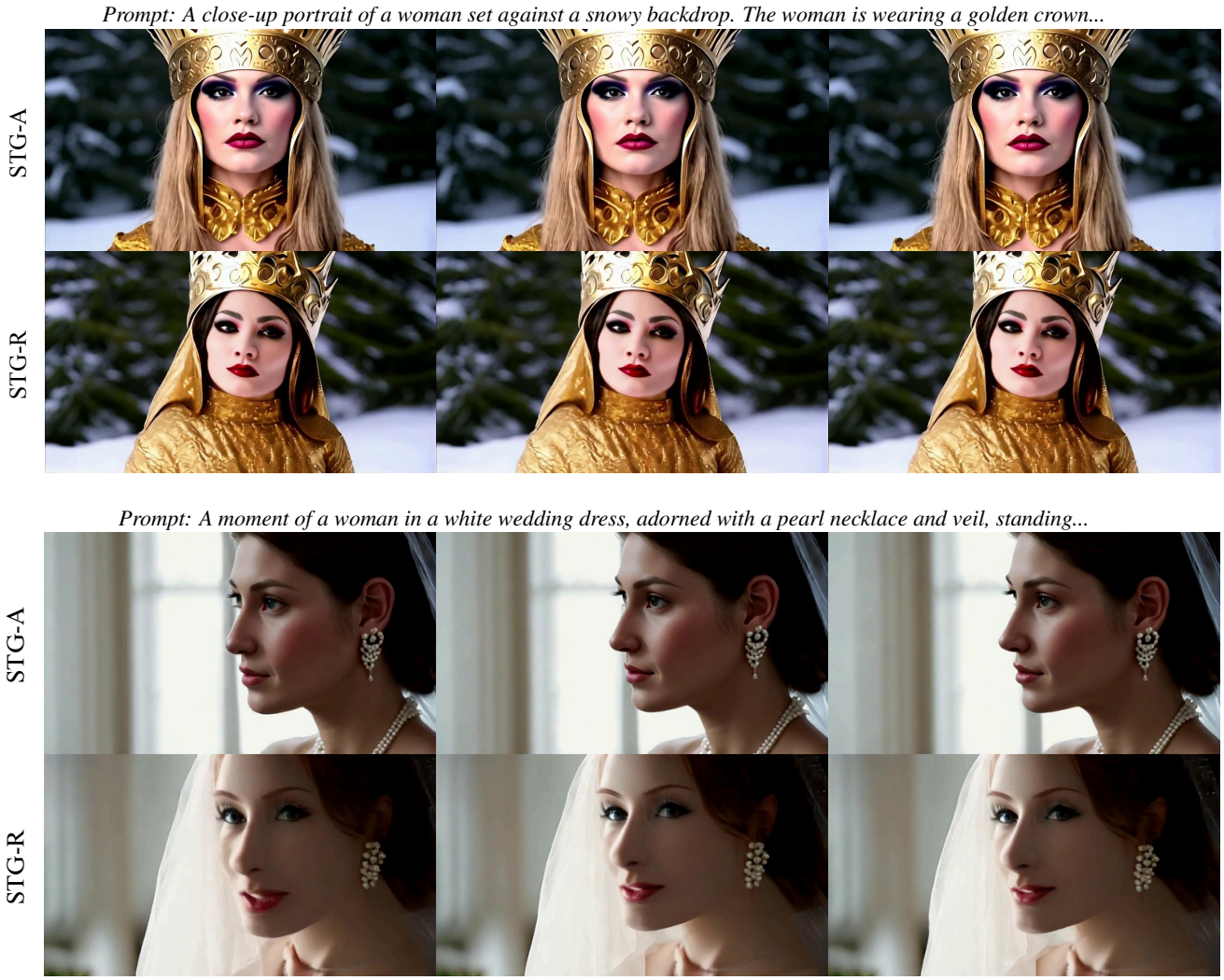


Figure 16. Comparison of attention skip (STG-A) and residual skip (STG-R) in Open-Sora [38]. The results indicate that STG-A delivers greater qualitative improvements for Open-Sora.

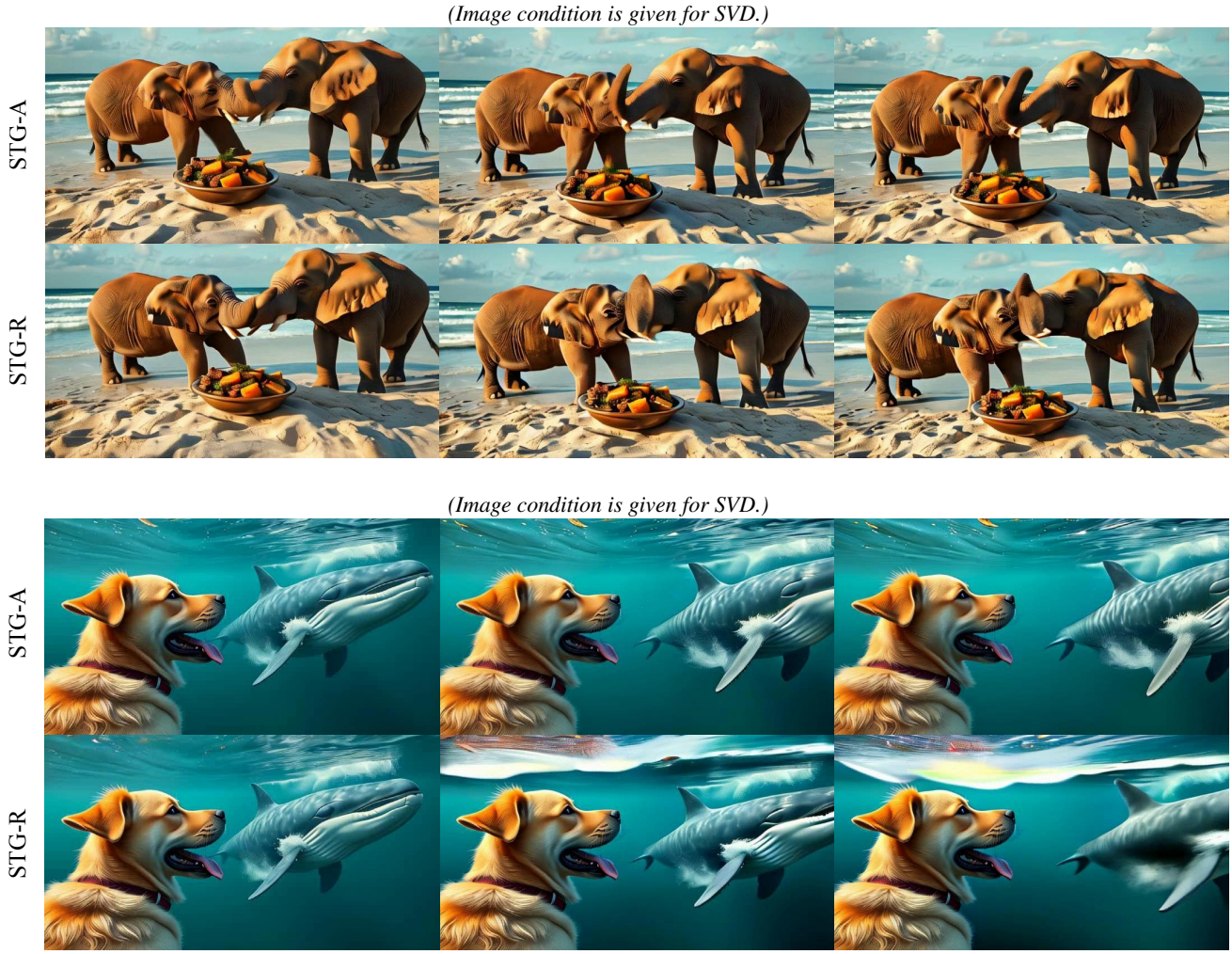


Figure 17. Comparison of attention skip (STG-A) and residual skip (STG-R) in SVD [4]. The results indicate that STG-A delivers greater qualitative improvements for SVD.

(Image condition is given for SVD.)

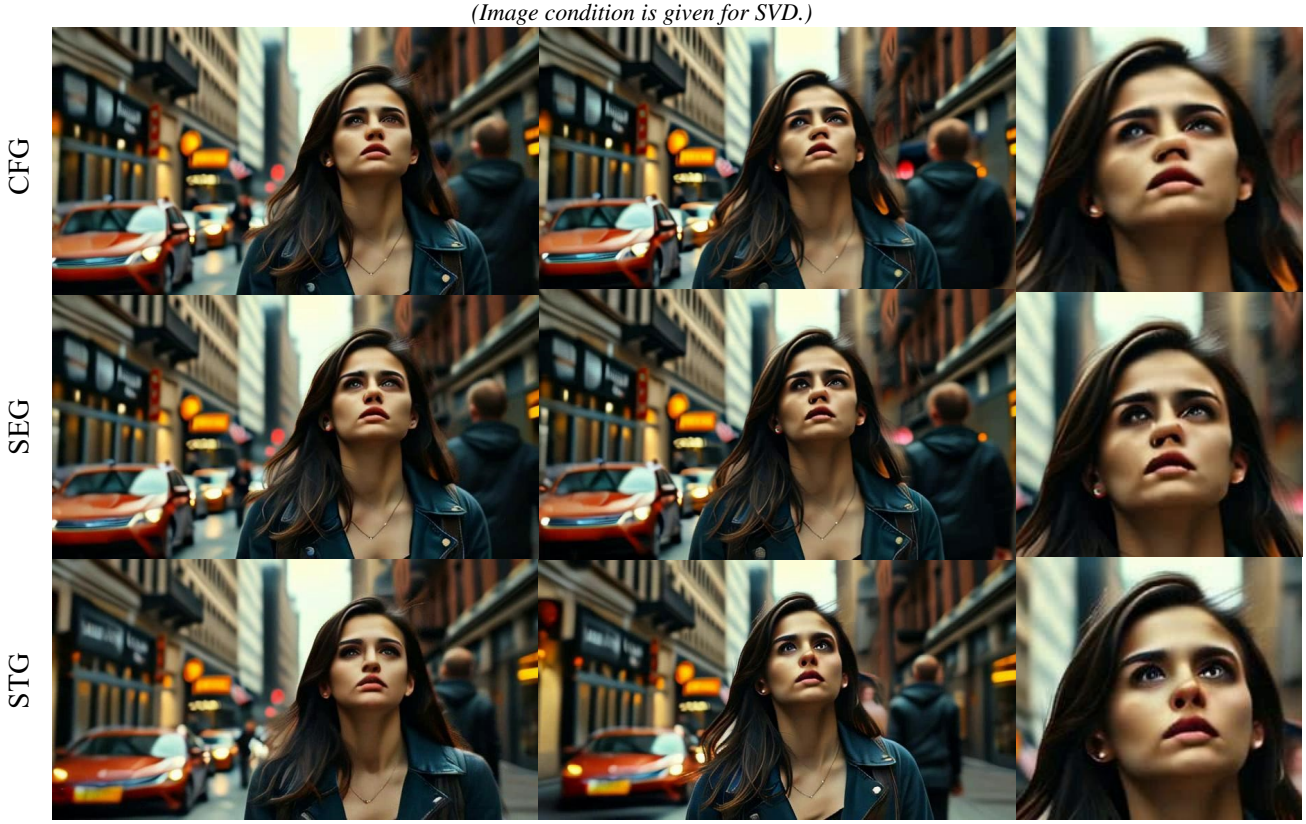


Figure 18. Comparison of CFG, SEG [13], and STG in SVD [4]. The results show that CFG and SEG generate an unnatural nose for the person, whereas STG successfully generates all components naturally.

Prompt: A family having a picnic under a shady tree in a large park.

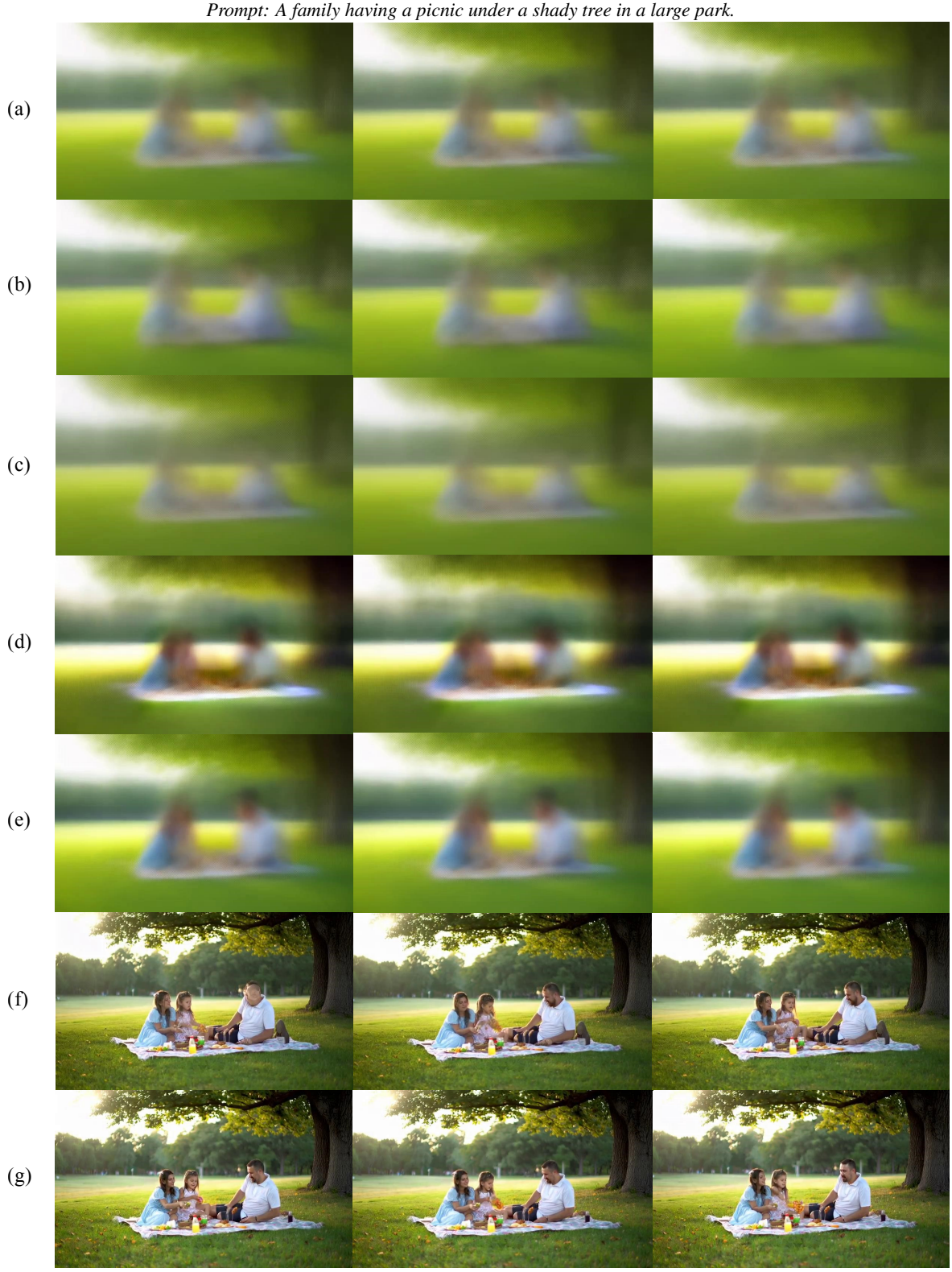


Figure 19. Weak model visualization for Mochi [33]. Generated video using one-step prediction from timestep 30 ($t = 30$). (a) $\epsilon_\theta(x_t)$, (b) $\epsilon_\theta(x_t|\phi)$, (c) $\epsilon_\theta^{s,t}(x_t)$, (d) CFG, (e) STG (f) Final video (CFG) (g) Final video (STG). The video predicted by CFG exhibits unnatural colors in certain areas and broken structures. In contrast, the video generated with STG demonstrates improved structural integrity and more natural color tones.

Prompt: A family having a picnic under a shady tree in a large park.

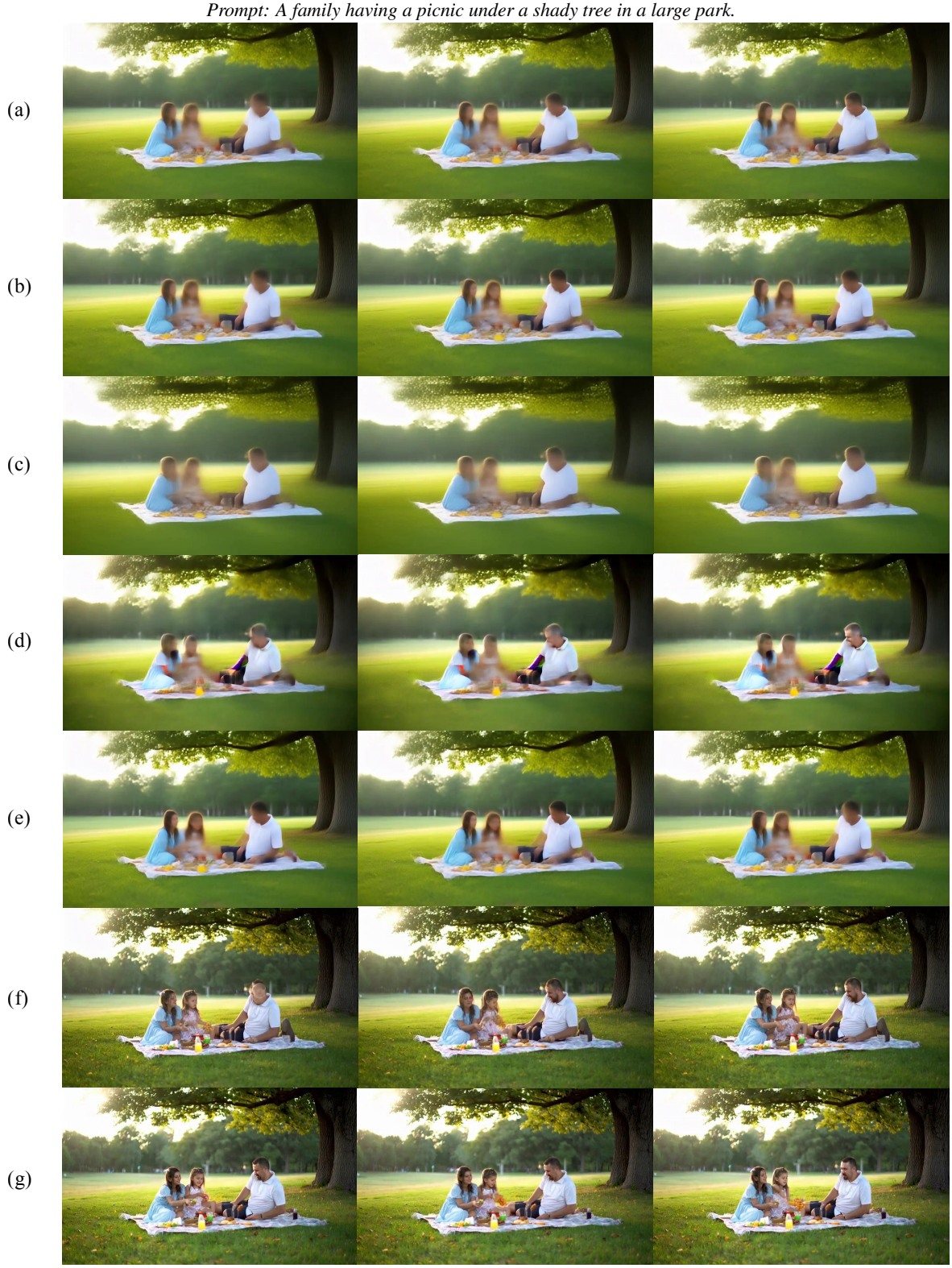
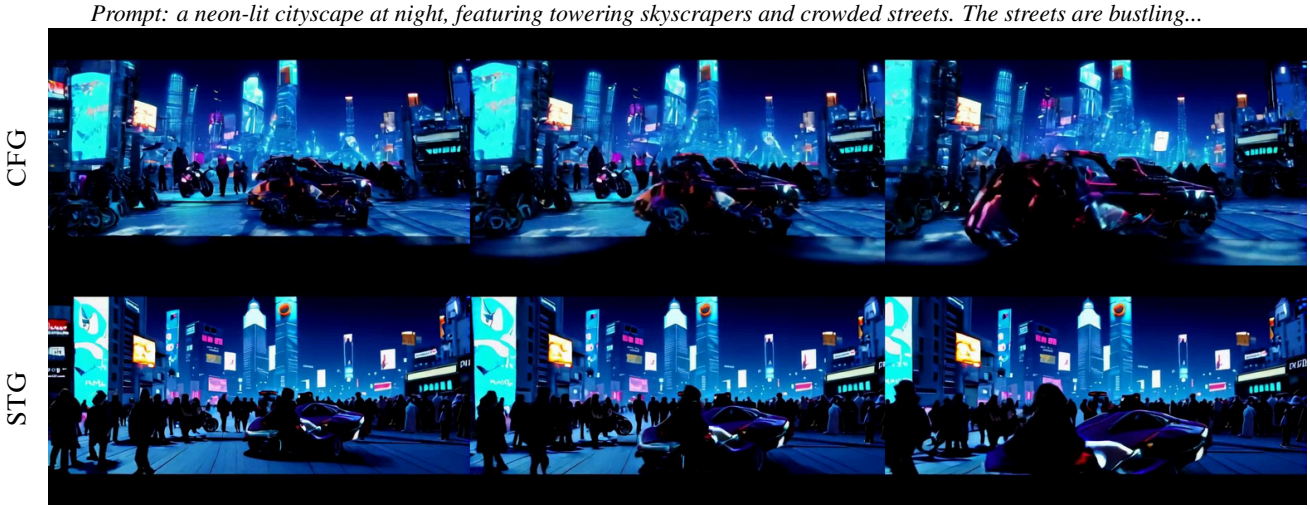
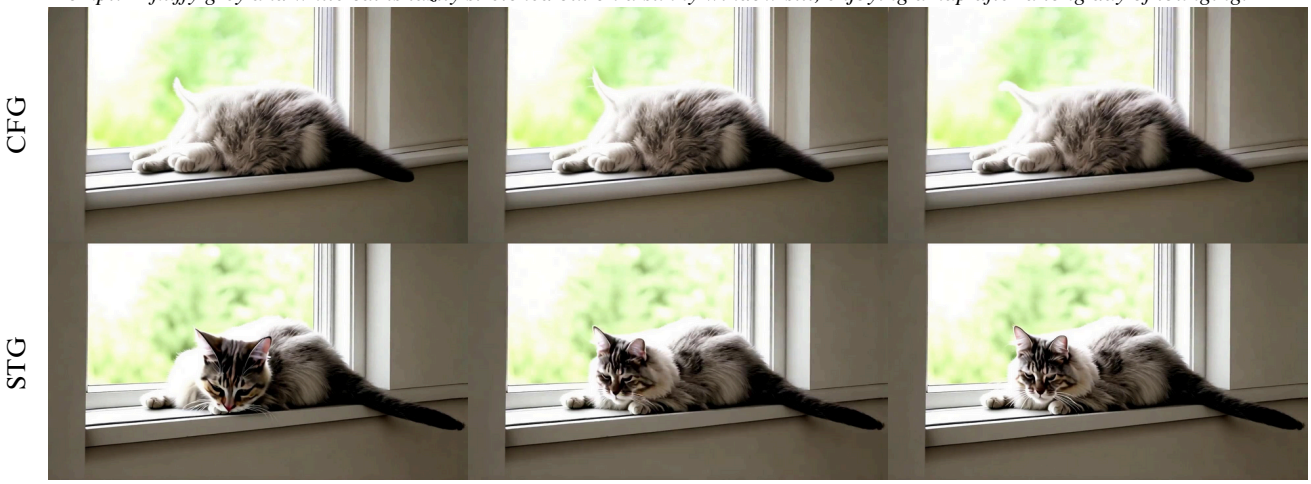


Figure 20. Weak model visualization for Mochi [33]. Video generated using one-step prediction from timestep 24 ($t = 24$). (a) $\epsilon_\theta(x_t)$, (b) $\epsilon_\theta(x_t|\phi)$, (c) $\epsilon_\theta^{s,t}(x_t)$, (d) CFG, (e) STG (f) Final video (CFG) (g) Final video (STG). The result demonstrates that STG effectively guides the model to maintain structural integrity and realistic color distribution while avoiding the unintended artifacts present in CFG predictions.

Prompt: a neon-lit cityscape at night, featuring towering skyscrapers and crowded streets. The streets are bustling...



Prompt: A fluffy grey and white cat is lazily stretched out on a sunny window sill, enjoying a nap after a long day of lounging.



Prompt: Iron Man is walking towards the camera in the rain at night, with a lot of fog behind him. Science fiction movie, close-up.

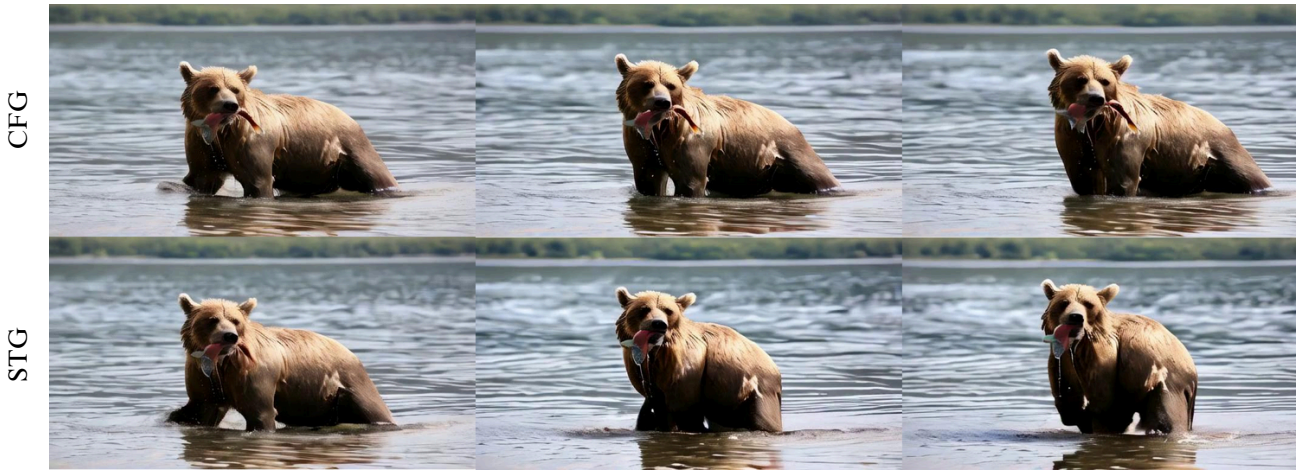
Figure 21. Qualitative comparison of video quality with and without STG applied on Open-Sora [38]. The results demonstrate that applying STG enhances the video's aesthetic appeal and fidelity.

Prompt: A dog wearing vr goggles on a boat.

Prompt: A cyborg standing on top of a skyscraper, overseeing the city, back view, cyberpunk vibe, 2077, NYC, intricate details, 4K.

Figure 22. Qualitative comparison of Temporal Flickering in Open-Sora [38] with and without STG. Without STG, temporal flickering is observed around the object, causing sudden bright flashes that disrupt the video experience. STG significantly reduces these artifacts, resulting in smoother and more cohesive motion.

(Image condition is given for SVD.)



(Image condition is given for SVD.)

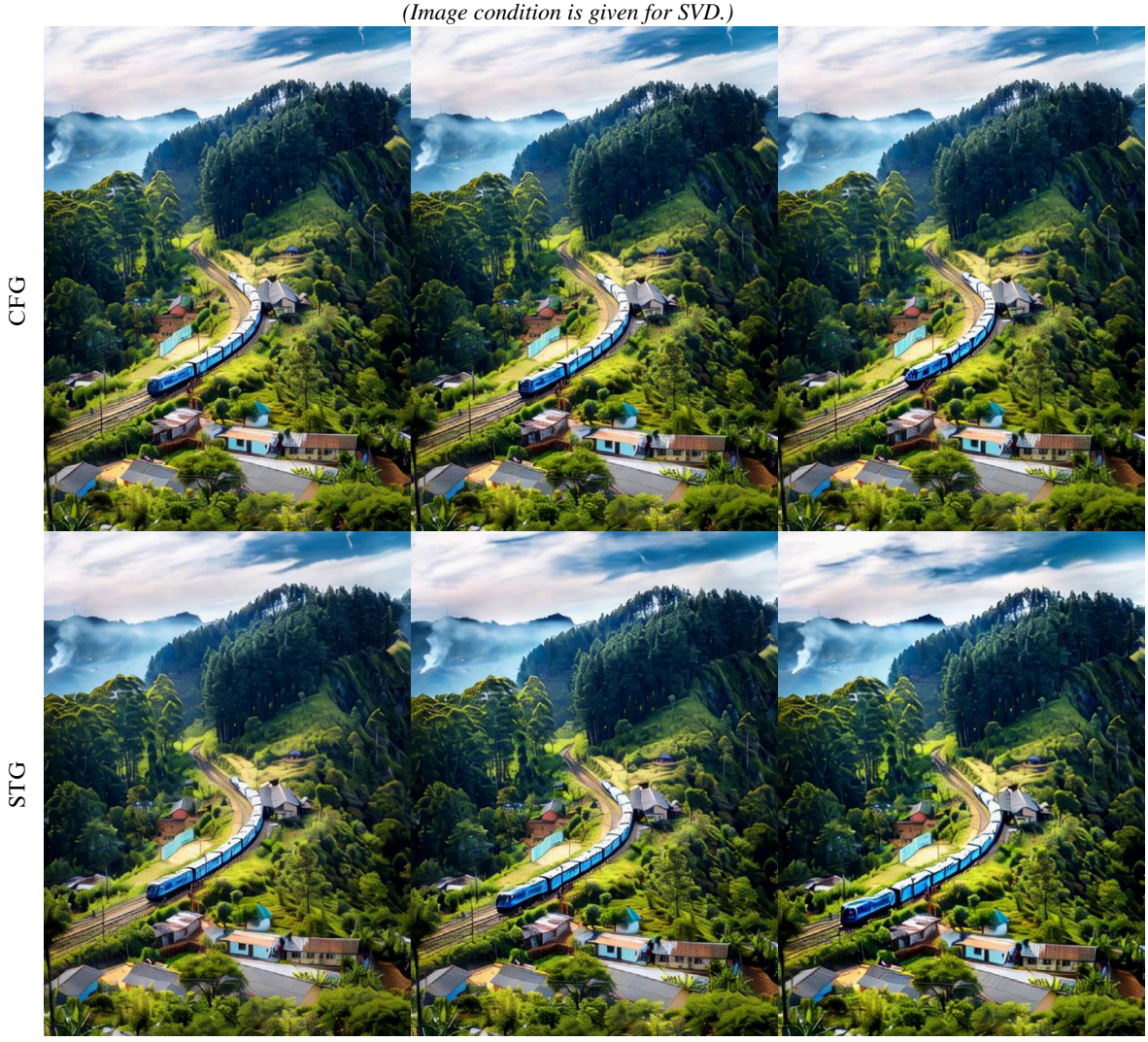


Figure 23. Qualitative Comparison of Dynamic Degree in SVD [4] with and without STG. The results show CFG results in limited object motion, whereas STG mitigates the restrictive effects of CFG, effectively enhancing the motion.