

Supplementary Material for TEMPOCONTROL: Temporal Attention Guidance for Text-to-Video Models

This supplementary material includes algorithmic pseudocode, ablation studies, experimental design details, application-specific configurations, extended quantitative and qualitative results, and the human-evaluation setup. Videos and code are also provided in the supplementary package.

The sections of the supplementary material are as follows:

1. Algorithm pseudocode (Section [.1](#)).
2. Experimental setup and hyperparameters (Section [.2](#)).
3. Configuration details for each application setting (Section [.3](#)).
4. Human evaluation protocol and interface (Section [.4](#)).
5. Latency analysis and trade-off (Section [.5](#)).
6. Comparison to Dense Diffusion (Section [.6](#)).
7. Optimization steps and update frequency (Section [.7](#)).
8. Extended multi-object evaluation (Section [.8](#)).
9. Ablation of loss components (Section [.9](#)).
10. Comparison to cosine similarity objective (Section [.10](#)).
11. Limitations of explicit temporal cues in text-to-video models (Section [.11](#)).
12. Additional qualitative results (Section [.12](#)).

.1. Algorithm Pseudo-code

Algorithm 1 outlines our inference-time optimization procedure. Starting from noise z_T , we iteratively steer the latent code during the first k denoising steps of the diffusion process. At each step t , we compute cross-attention maps from the video-text model DiT and derive the scalar temporal attention a_i^t for each token p_i .

The optimization loss \mathcal{L} combines three components: correlation alignment \mathcal{L}_c^t , magnitude modulation $\mathcal{L}_{\text{enr}}^t$, entropy regularization and $\mathcal{L}_{\text{ent}}^t$. The latent code is updated via gradient descent for up to l steps per denoising timestep, or until the Pearson correlation exceeds a threshold τ_{corr} . After optimization, the standard denoising process resumes to produce the final latent z_0 .

Algorithm 1: TEMPOCONTROL: Temporal Attention-Guided Inference-Time Steering

Input: Prompt \mathcal{P} , tokens p_i with temporal masks m_i , total diffusion steps T , optimized steps k , max updates per step l , correlation threshold τ_{corr} , model DiT

Output: Final latent z_0

```

1: Initialize  $z_T \sim \mathcal{N}(0, I)$ 
2: for  $t = T, T-1, \dots, T-k+1$  do
3:   for  $u = 1$  to  $l$  do
4:     Extract cross-attentions  $A^t$  from DiT( $z_t, \mathcal{P}, t$ )
5:     Average heads and layers:  $\bar{A}^t$ 
6:      $\mathcal{L} \leftarrow 0$ 
7:     for each  $p_i$  with  $m_i$  do
8:        $a_i^t \leftarrow [\langle \bar{A}_{j,i}^t \rangle_{x,y}]_{j=1}^{T'}$ 
9:        $\mathcal{L}_c^t \leftarrow -\text{Corr}(m_i, \text{minmax}(a_i^t))$ 
10:       $\mathcal{L}_\oplus^t \leftarrow \frac{1}{T'} \sum_{j=1}^{T'} \mathbb{1}[m_{i,j} > \tau] \cdot a_{i,j}^t$ 
11:       $\mathcal{L}_\ominus^t \leftarrow \frac{1}{T'} \sum_{j=1}^{T'} \mathbb{1}[m_{i,j} \leq \tau] \cdot a_{i,j}^t$ 
12:       $\mathcal{L}_{\text{mag}}^t \leftarrow \mathcal{L}_\ominus^t - \mathcal{L}_\oplus^t$ 
13:       $\mathcal{L}_{\text{ent}}^t \leftarrow \frac{1}{T'} \sum_{j=1}^{T'} \mathbb{1}[m_{i,j} > \tau] \cdot \mathcal{H}(\bar{A}_{j,i}^t)$ 
14:       $\mathcal{L} \leftarrow \mathcal{L} + \mathcal{L}_c^t + \lambda_1 \mathcal{L}_{\text{enr}}^t + \lambda_2 \mathcal{L}_{\text{ent}}^t$ 
15:     end for
16:      $z'_t \leftarrow z_t - \alpha_t \nabla_{z_t} (\mathcal{L}/N)$ 
17:     if All  $\mathcal{L}_{\text{corr}}^t \leq \tau_{\text{corr}}$  then
18:       break
19:     end if
20:      $z_t \leftarrow z'_t$ 
21:   end for
22:    $z_{t-1} \leftarrow \text{DiT}(z'_t, \mathcal{P}, t)$ 
23: end for
24: Continue denoising using the standard denoising diffusion process for  $t = T-k, \dots, 1$ 
25: return  $z_0$ 

```

.2. Experimental Setup

All experiments use Wan 2.1 as the backbone model. We apply a classifier-free guidance scale of 6 and a sample shift of 3 during generation.

For single-object scenes, we use $\lambda_1=0.3$, $\lambda_2=10$, a learning rate of 5×10^{-4} , a Pearson stopping threshold of $\tau_{\text{corr}}=0.9$, and perform optimization over the first $l=5$ denoising steps with up to $k=10$ gradient updates per step.

For two-object scenes, which demand stronger guidance, we increase the learning rate to 10^{-3} to improve convergence. For motion-centric prompts, we relax the Pearson threshold to $\tau_{\text{corr}}=0.85$ to enable more flexible optimization.

.3. Applications Setup

Single-Object Temporal Control. To evaluate temporal grounding for individual objects, we construct a dataset based on the 80 object categories detectable by the YOLO detector. These categories cover common entities such as animals, vehicles, household items, and everyday objects. The complete class list is provided below:

person, bicycle, car, motorcycle, airplane, bus, train, truck, boat, traffic light, fire hydrant, stop sign, parking meter, bench, bird, cat, dog, horse, sheep, cow, elephant, bear, zebra, giraffe, backpack, umbrella, handbag, tie, suitcase, frisbee, skis, snowboard, sports ball, kite, baseball bat, baseball glove, skateboard, surfboard, tennis racket, bottle, wine glass, cup, fork, knife, spoon, bowl, banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake, chair, couch, potted plant, bed, dining table, toilet, tv, laptop, mouse, remote, keyboard, cell phone, microwave, oven, toaster, sink, refrigerator, book, clock, vase, scissors, teddy bear, hair drier, toothbrush.

Videos are generated from prompts of the form ‘*An empty scene. Suddenly, $\langle o_1 \rangle$ appears out of nowhere, drawing all attention.*’ Object classes are uniformly sampled across different temporal intervals.

Two-Object Temporal Control. We evaluate TEMPOCONTROL on 82 prompts derived from the VBench protocol, each consisting of a pair of objects. The complete numbered list of pairs is provided below:

1. **a bird and a cat**
2. **a cat and a dog**
3. **a dog and a horse**
4. **a horse and a sheep**
5. **a sheep and a cow**
6. **a cow and an elephant**
7. **an elephant and a bear**
8. **a bear and a zebra**
9. **a zebra and a giraffe**
10. **a giraffe and a bird**
11. **a chair and a couch**
12. **a couch and a potted plant**
13. **a potted plant and a tv**
14. **a tv and a laptop**
15. **a laptop and a remote**
16. **a remote and a keyboard**
17. **a keyboard and a cell phone**
18. **a cell phone and a book**
19. **a book and a clock**
20. **a clock and a backpack**
21. **a backpack and an umbrella**
22. **an umbrella and a handbag**
23. **a handbag and a tie**
24. **a tie and a suitcase**
25. **a suitcase and a vase**
26. **a vase and scissors**
27. **scissors and a teddy bear**
28. **a teddy bear and a frisbee**
29. **a frisbee and skis**
30. **skis and a snowboard**

31. **a snowboard** and **a sports ball**
32. **a sports ball** and **a kite**
33. **a kite** and **a baseball bat**
34. **a baseball bat** and **a baseball glove**
35. **a baseball glove** and **a skateboard**
36. **a skateboard** and **a surfboard**
37. **a surfboard** and **a tennis racket**
38. **a tennis racket** and **a bottle**
39. **a bottle** and **a chair**
40. **an airplane** and **a train**
41. **a train** and **a boat**
42. **a boat** and **an airplane**
43. **a bicycle** and **a car**
44. **a car** and **a motorcycle**
45. **a motorcycle** and **a bus**
46. **a bus** and **a traffic light**
47. **a traffic light** and **a fire hydrant**
48. **a fire hydrant** and **a stop sign**
49. **a stop sign** and **a parking meter**
50. **a parking meter** and **a truck**
51. **a truck** and **a bicycle**
52. **a toilet** and **a hair drier**
53. **a hair drier** and **a toothbrush**
54. **a toothbrush** and **a sink**
55. **a sink** and **a toilet**
56. **a wine glass** and **a chair**
57. **a cup** and **a couch**
58. **a fork** and **a potted plant**
59. **a knife** and **a tv**
60. **a spoon** and **a laptop**
61. **a bowl** and **a remote**
62. **a banana** and **a keyboard**
63. **an apple** and **a cell phone**
64. **a sandwich** and **a book**
65. **an orange** and **a clock**
66. **broccoli** and **a backpack**
67. **a carrot** and **an umbrella**
68. **a hot dog** and **a handbag**
69. **a pizza** and **a tie**
70. **a donut** and **a suitcase**
71. **a cake** and **a vase**
72. **an oven** and **scissors**
73. **a toaster** and **a teddy bear**
74. **a microwave** and **a frisbee**
75. **a refrigerator** and **skis**
76. **a bicycle** and **an airplane**
77. **a car** and **a train**
78. **a motorcycle** and **a boat**
79. **a person** and **a toilet**
80. **a person** and **a hair drier**
81. **a person** and **a toothbrush**
82. **a person** and **a sink**

Prompts follow the structure: ‘*The video begins with a serene view centered on $\langle o_1 \rangle$, with no sign of $\langle o_2 \rangle$. In the second half, $\langle o_2 \rangle$ unexpectedly appears, altering the dynamic of the scene.*’

A static control signal (all ones) is used for $\langle o_1 \rangle$. We sample 20 evenly spaced frames and require $\langle o_1 \rangle$ to be present throughout. A frame is counted as successful if both objects are detected when $\langle o_2 \rangle$ is active, and only $\langle o_1 \rangle$ is detected when $\langle o_2 \rangle$ is inactive.

Motion Temporal Control. To evaluate fine-grained temporal modulation of actions, we generate 100 videos using prompts of the form: ‘A video of $\langle s \rangle$ $\langle v \rangle$ $\langle a \rangle$ with a strong movement at the $\langle t \rangle$ second,’ where $\langle s \rangle$, $\langle v \rangle$, $\langle a \rangle$, and $\langle t \rangle$ denote the subject, verb, adverb, and time reference, respectively.

The full list of subject–verb–adverb triplets is given below:

1. **a cat** pouncing **silently**
2. **a dancer** twirling **elegantly**
3. **a bird** flapping **frantically**
4. **a lion** roaring **loudly**
5. **a wave** crashing **violently**
6. **a ballerina** leaping **lightly**
7. **a helicopter** spinning **noisily**
8. **a snake** slithering **smoothly**
9. **a bear** growling **deeply**
10. **a man** running **fast**
11. **a woman** laughing **brightly**
12. **a child** crawling **slowly**
13. **a monkey** swinging **playfully**
14. **a firework** exploding **suddenly**
15. **a wolf** howling **hauntingly**
16. **a dolphin** diving **swiftly**
17. **a plane** ascending **steadily**
18. **a kangaroo** hopping **energetically**
19. **a fox** darting **quickly**
20. **a shadow** moving **mysteriously**
21. **a squirrel** scampering **nervously**
22. **a truck** crashing **loudly**
23. **a child** screaming **unexpectedly**
24. **a penguin** waddling **adorably**
25. **a frog** croaking **rhythmically**
26. **a leaf** drifting **slowly**
27. **a shark** circling **menacingly**
28. **a comet** streaking **brightly**
29. **a dancer** collapsing **dramatically**
30. **a boxer** punching **fiercely**
31. **a tree** swaying **gently**
32. **a man** lifting weights **powerfully**
33. **a windmill** rotating **steadily**
34. **a girl** skipping **cheerfully**
35. **a car** drifting **dangerously**
36. **a train** accelerating **fast**
37. **a snake** striking **quickly**
38. **a fire** burning **intensely**
39. **a glacier** cracking **slowly**
40. **a bee** buzzing **constantly**
41. **a deer** sprinting **fearfully**
42. **a volcano** erupting **violently**
43. **a runner** collapsing **from exhaustion**
44. **a gymnast** flipping **smoothly**
45. **a rocket** launching **thunderously**
46. **a drummer** hitting **rapidly**
47. **a magician** vanishing **mysteriously**
48. **a spider** crawling **delicately**
49. **a tiger** pacing **restlessly**
50. **a surfer** balancing **skillfully**
51. **a leopard** growling **softly**
52. **a swimmer** diving **gracefully**
53. **a baby** clapping **happily**

54. **a mime** gesturing **expressively**
55. **a crow** cawing **sharply**
56. **a goat** headbutting **suddenly**
57. **a girl** blowing bubbles **gently**
58. **a chef** chopping **rapidly**
59. **a horse** shaking its mane **proudly**
60. **a robot** malfunctioning **erratically**
61. **a meteor** falling **fast**
62. **a snake** coiling **tightly**
63. **a rabbit** thumping **nervously**
64. **a man** collapsing **dramatically**
65. **a car** braking **suddenly**
66. **a hawk** diving **precisely**
67. **a wolf** snarling **viciously**
68. **a plane** landing **smoothly**
69. **a firetruck** speeding **urgently**
70. **a chimpanzee** clapping **playfully**
71. **a dancer** stomping **rhythmically**
72. **a snail** inching **slowly**
73. **a lioness** crouching **quietly**
74. **a storm cloud** rolling **ominously**
75. **a bee** landing **precisely**
76. **a magician** pulling a rabbit **suddenly**
77. **a glacier** melting **steadily**
78. **a jellyfish** floating **gracefully**
79. **a tree** falling **loudly**
80. **a drummer** drumming **fiercely**
81. **a torch** blazing **brightly**
82. **a woman** spinning **dramatically**
83. **a child** running **barefoot**
84. **a fox** sniffing **cautiously**
85. **a man** throwing a ball **forcefully**
86. **a bear** shaking off water **heavily**
87. **a lion** walking **regally**
88. **a horse** rearing **suddenly**
89. **a falcon** gliding **silently**
90. **an elephant** raising its trunk **highly**
91. **a dog** barking **forcefully**
92. **a man** sneezing **powerfully**
93. **a child** jumping **joyfully**
94. **a tiger** leaping **fiercely**
95. **a woman** spinning **gracefully**
96. **a robot** marching **stiffly**
97. **a horse** galloping **wildly**
98. **an eagle** soaring **majestically**
99. **a flame** flickering **rapidly**
100. **a skateboarder** flipping **expertly**

The verb is modulated using a binary temporal mask (set to 1 only at time $\langle t \rangle$), while the subject remains active throughout the video.

.4. Human Evaluation Protocol

To assess temporal alignment and visual quality, we conducted a human preference study using side-by-side video comparisons. All annotators were graduate students conducting research in computer vision or closely related areas.

For each prompt, two videos were shown: one generated by our method and one by the text-only baseline, with their order (Video A or Video B) randomized to avoid bias.

Annotators were instructed to read the prompt and answer two questions: (1) Which video is more temporally accurate (i.e., does the object appear at the correct time and remain visible)? (2) Which video is more visually appealing? A “Same (use sparingly)” option was available when no meaningful difference was perceived.

After the experiments, several annotators noted that judging temporal control could be challenging when the difference in timing was only 1–2 seconds. They reported that knowing the exact expected trigger time from the control mask would have helped them assess temporal accuracy more reliably. While we chose not to repeat the study, we include this observation as guidance for future temporal evaluation protocols.

Example screenshots of the evaluation interface are shown in Figures 7 and 8.

User Study- Temporal Control for Text to Video Diffusion Models

Welcome, and thank you very much for your willingness to participate!

This study is part of an academic research project and will take no more than a **few minutes** of your time.

The goal is to evaluate how well text-to-video diffusion models can generate **temporally accurate** videos. That is, whether objects appear at the exact time specified in the prompt.

Here's how it works:

You'll be shown **two videos at a time**, both generated from the **same prompt**.

Each prompt includes a **temporal instruction**, such as:

"A dog appears in the third second," or "In the second half, a bottle appears."

For each video pair, please answer two questions:

- **Which video is more temporally accurate?** (Does the object appear at the right time, and remain visible until the end of the video?)
- **Which video looks more visually appealing to you?** (This includes overall quality, consistency, and how much sense the video makes.)

Your answers will help us compare a baseline model with our new method.

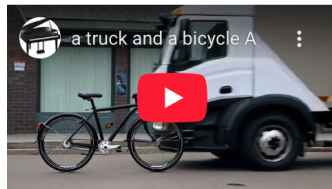
Thanks again, and enjoy the mini tour into the world of text-to-video models :)

Figure 7. Evaluation interface showing the prompt and two videos (A and B) for comparison.

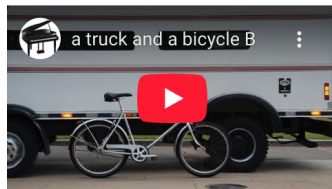
Prompt:

"The video begins with a serene view centered on the *bicycle*, with no sign of the *truck*. In the second half, the *truck* unexpectedly appears, altering the dynamic of the scene"

Video A



Video B



2. *

Which video is more temporally accurate?

(Does the object appear at the right time, and remain visible until the end of the video?)

- ☐ Video A
- ☐ Video B
- ☐ Same (Use sparingly!)

2. *

Which video looks more visually appealing to you?

- ☐ Video A
- ☐ Video B
- ☐ Same (Use sparingly!)

Figure 8. Second example of the human evaluation interface.

5. Latency Discussion

We analyze the tradeoff between inference latency and temporal control by varying the number of inference steps in TEMPOCONTROL. Figure 9 visualizes the latency–accuracy frontier, while Table 5 reports the corresponding temporal, absence, and presence accuracies.

As shown, TEMPOCONTROL consistently outperforms all baselines in comparable latency regimes. Even with only 2 inference steps, TEMPOCONTROL achieves 83.25% temporal accuracy, surpassing all baselines while operating within a similar latency range to mid-size models. Increasing the number of steps further improves performance, with 5 steps providing the best overall balance between latency and accuracy (83.56% temporal accuracy). Beyond this point, the gains are marginal, with the performance largely saturating after 5–10 steps.

Importantly, TEMPOCONTROL achieves superior temporal control while using a significantly smaller backbone (1.3B parameters), compared to baselines that scale up to 19B parameters. This demonstrates that strong temporal reasoning can be achieved efficiently without requiring large-scale models.

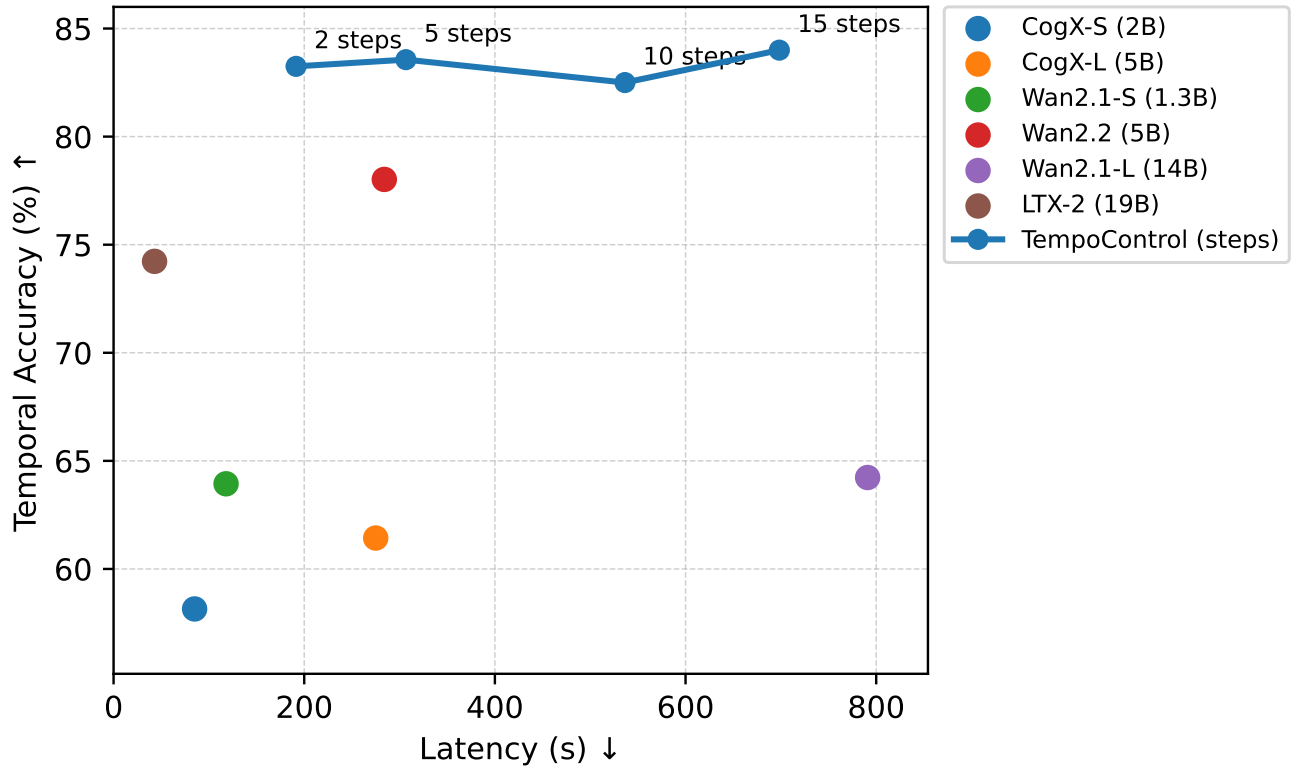


Figure 9. Latency vs. temporal accuracy. Each baseline is shown as a single point, while TEMPOCONTROL forms a tradeoff curve as the number of inference steps increases. Labels denote the number of inference steps.

Table 5. Latency–performance tradeoff. We compare TEMPOCONTROL (Wan 1.3B) under different inference steps against representative baselines.

Method	Latency (s) ↓	Temp ↑	Abs ↑	Pres ↑
CogX (2B)	~ 85	58.15	74.77	39.88
Wan2.1-S (1.3B)	~ 118	63.94	67.38	60.50
Wan2.2 (5B)	~ 284	78.02	87.42	68.00
Wan2.1-L (14B)	~ 791	64.23	55.34	74.00
LTX-2 (19B)	~ 43	74.23	85.23	62.12
Ours (2 steps)	~ 191	83.25	86.12	80.38
Ours (5 steps)	~ 307	83.56	87.38	79.75
Ours (10 steps)	~ 536	82.50	83.25	81.75
Ours (15 steps)	~ 698	84.00	86.38	81.62

.6. Dense Diffusion vs. TempoControl

We compare TempoControl to Dense Diffusion [15], which applies direct spatiotemporal masking $R_i(t, x, y)$ to cross-attention maps based on a temporal control signal, without explicitly optimizing temporal alignment.

While this enables temporal gating, it does not align the attention distribution $A_i(t)$ with the desired pattern $m_{i,t}$. Moreover, the lack of regularization leads to spatially incoherent attention, producing visual artifacts. These artifacts degrade object appearance and hinder detection, resulting in artificially inflated absence accuracy rather than correct temporal control.

In contrast, TempoControl explicitly optimizes temporal alignment via correlation, controls activation strength via a magnitude term, and enforces spatial coherence through entropy regularization. This yields temporally aligned and visually consistent generations, improving presence accuracy while maintaining image quality.

Method	Temp. Acc.	Abs. Acc.	Pres. Acc.	Img. Qual.
Dense diffusion	77.94	86.25	69.62	51.02
TempoControl	82.50	83.25	81.75	56.51

Table 6. Comparison between Dense Diffusion and TempoControl with Wan2.1 for the one object setting..

.7. Optimization Steps and Update Frequency

We analyze the effect of optimization depth for TEMPOCONTROL. We first ablate the number of gradient updates per step in the one-object setting (Table 7). Based on this, we fix the number of updates to 5 and vary the number of optimized diffusion steps across tasks and backbones (Table 8).

Increasing the number of gradient updates improves temporal alignment, but fewer than five updates per step leads to under-optimized attention, while additional updates yield diminishing returns.

The impact of optimizing different diffusion steps depends on when temporal structure is established in the underlying model. In Wan-based one- and two-object settings, performance saturates after 5 steps, as object appearance is determined early in the denoising process. In contrast, movement control and CogVideoX benefit from additional steps, since temporal structure emerges later. Overall, optimizing 5 steps provides a strong trade-off between performance and latency.

Table 7. Effect of gradient updates per step in the one-object setting. Bold denotes the best result.

Setting	Acc ↑	Abs ↑	Pres ↑	Img ↑
2 steps / 2 updates	72.00%	74.50%	69.50%	57.87%
2 steps / 5 updates	77.75%	79.00%	76.50%	57.41%
5 steps / 2 updates	74.12%	73.62%	74.62%	58.03%
5 steps / 5 updates	78.81%	81.50%	76.12%	57.43%
5 steps / 10 updates (Ours)	83.56%	87.38%	79.75%	56.92%

Table 8. Effect of optimized diffusion steps across tasks and backbones (5 updates). Bold denotes the best result within each block.

Setting	Acc ↑	Abs ↑	Pres ↑	Img ↑
One Object (Wan)				
2 steps	83.25%	86.12%	80.38%	56.72%
5 steps (Ours)	83.56%	87.38%	79.75%	56.92%
10 steps	82.50%	83.25%	81.75%	56.51%
One Object (CogVideoX)				
2 steps	60.12%	78.07%	40.38%	46.35%
5 steps (Ours)	63.75%	78.41%	47.62%	46.67%
10 steps	65.48%	77.27%	52.50%	47.88%
Two Objects (Wan)				
2 steps	50.98%	59.88%	42.07%	70.61%
5 steps (Ours)	53.17%	57.32%	49.02%	70.82%
10 steps	52.93%	58.05%	47.80%	71.09%
Movement (Wan)				
2 steps	38%	–	–	62.85%
5 steps (Ours)	54%	–	–	63.24%
10 steps	57%	–	–	63.38%

.8. Multi-Object Evaluation

We report extended results on multi-object text-to-video generation using the VBench benchmark. In this setting, prompts reference multiple entities (e.g., ‘A dog and a cat’), and the evaluation measures whether all specified objects appear consistently throughout the video. Each video is sampled at 16 evenly spaced frames, and a frame is counted as successful if both target objects are detected; the final score is the average per-frame success rate. In the context of our method, this setup corresponds to applying a constant mask of one for both objects across all frames.

For completeness, we briefly summarize the VBench metrics reported in Table 9:

- **Multiple-Objects Accuracy (GriT / YOLO).** Measures whether all referenced objects appear correctly in each sampled frame, using open-vocabulary detectors (GriT) or standard detectors (YOLO). Higher values indicate more reliable object visibility and grounding.
- **Subject Consistency.** Quantifies how consistently the appearance of the main subject is preserved across frames. It is computed using feature similarity both between consecutive frames and relative to the first frame of the video; higher values indicate a more stable and coherent subject identity.
- **Background Consistency.** Measures the stability of the background across the video. High scores indicate coherent scene structure without abrupt changes.
- **Dynamic Degree.** Captures the amount of motion or temporal change in the video. The metric is binary at the video level: a video is marked as dynamic if its average optical flow surpasses a predefined threshold. Higher values indicate more motion, while lower values reflect greater temporal stability.

Table 9 summarizes the results. Our method achieves substantial gains in object-centric metrics: multi-object accuracy improves from 74.1% to 76.4% (GriT) and from 61.5% to 65.7% (YOLO), demonstrating better grounding and more reliable visibility of both entities. We also observe improvements across consistency metrics, including subject consistency (from 97.2% to 97.8%) and background consistency (from 97.6% to 98.1%).

The Dynamic Degree decreases from 30.5% to 18.8%, reflecting a trade-off: our method promotes temporal stability, often leading to more consistent object appearance across the video, at the cost of reduced motion. This highlights a limitation of this evaluation protocol: VBench counts an object as correct only when it is present in *every* sampled frame; hence, the benchmark implicitly favors unnatural videos in which objects persist throughout the sequence.

Table 9. Full results for Multiple object benchmark results. Best results per column are in bold.

Method	Multiple Object (GriT)	Multiple Object (YOLO)	Subject Consistency	Background Consistency	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality
Text	74.13%	61.54%	97.22%	97.56%	99.18%	30.49%	62.84%	70.25%
Ours	76.37%	65.73%	97.81%	98.10%	99.40%	18.78%	62.52%	70.21%

.9. Loss Component Ablation

We conduct an ablation study to analyze the influence of the magnitude weight λ_1 and the entropy weight λ_2 on temporal control and visual quality. The results reveal clear and complementary effects of the two loss components. Increasing the magnitude term λ_1 reduces Imaging Quality while improving T. Absence Accuracy. This suggests that stronger magnitude guidance more strictly enforces the temporal mask, but may also reduce overall visual fidelity.

In contrast, increasing the entropy term λ_2 improves Imaging Quality and raises T. Presence Accuracy, but it decreases T. Absence Accuracy, indicating that stronger entropy regularization encourages more persistent object appearance.

Overall, our chosen hyperparameters ($\lambda_1 = 0.3$, $\lambda_2 = 10$) provide a balanced tradeoff between these competing effects, achieving strong Temporal Accuracy while maintaining high Imaging Quality.

Table 10. Ablation results with Wan2.1-S for the one object setting. All variants use the same temporal Pearson term; we ablate the magnitude weight λ_1 and the entropy weight λ_2 . Best scores per column are in **bold**.

(a) Effect of magnitude weight λ_1 with fixed entropy weight $\lambda_2 = 10$.				
Method	Temporal Accuracy	T. Absence Accuracy	T. Presence Accuracy	Imaging Quality
Text baseline	63.94%	67.38%	60.50%	53.76%
Ours ($\lambda_1=0.3$, $\lambda_2=10$)	82.50%	83.25%	81.75%	56.51%
$\lambda_1=0.1$, $\lambda_2=10$	79.50%	80.62%	78.38%	56.37%
$\lambda_1=0.65$, $\lambda_2=10$	81.06%	86.50%	75.62%	53.94%
$\lambda_1=1.0$, $\lambda_2=10$	83.12%	89.00%	77.25%	53.74%

(b) Effect of entropy weight λ_2 with fixed magnitude weight $\lambda_1 = 0.3$.				
Method	Temporal Accuracy	T. Absence Accuracy	T. Presence Accuracy	Imaging Quality
Text baseline	63.94%	67.38%	60.50%	53.76%
Ours ($\lambda_1=0.3$, $\lambda_2=10$)	82.50%	83.25%	81.75%	56.51%
$\lambda_1=0.3$, $\lambda_2=1$	80.38%	91.25%	69.50%	52.01%
$\lambda_1=0.3$, $\lambda_2=5$	81.88%	88.50%	75.25%	54.04%
$\lambda_1=0.3$, $\lambda_2=15$	81.56%	81.62%	81.50%	56.31%

.10. Cosine Similarity vs. TempoControl

We evaluate a variant in which the pearson correlation term is replaced with cosine similarity, a standard measure of alignment between attention patterns.

Cosine similarity jointly captures alignment and magnitude, effectively entangling temporal synchronization with activation strength. This coupling allows improvements in the objective without necessarily correcting temporal misalignment, since increased activation can artificially boost similarity. In contrast, TempoControl explicitly separates correlation (for alignment) and magnitude (for visibility), enabling more precise and targeted optimization. This decoupling improves temporal accuracy and presence.

Method	Temp. Acc.	Abs. Acc.	Pres. Acc.	Img. Qual.
Cosine similarity	79.00	82.38	75.62	55.97
TempoControl	82.50	83.25	81.75	56.51

Table 11. Comparison between cosine similarity objective and TempoControl with Wan2.1-S for the one object setting.

Figure 10 illustrates ablation results for the entropy term, revealing the limitation of using the Pearson correlation term alone without entropy regularization. While Pearson-based optimization successfully forces the object to appear at the desired time, it often alters the object’s semantics, leading to noticeable inconsistencies. For example, in the “Cake” row, the Pearson-only variant transforms the cake into a noticeably different shape and texture; in the “Sandwich” row, the sandwich becomes distorted across frames; and in the “Book” row, the open book becomes a closed stack with altered proportions. Similar semantic drift is observed in the “Keyboard,” “Oven,” and “Stop sign” examples.

These results highlight that Pearson correlation alone can enforce temporal timing but does not sufficiently constrain the appearance space, resulting in unwanted semantic changes. The entropy regularization helps mitigate those artifacts.

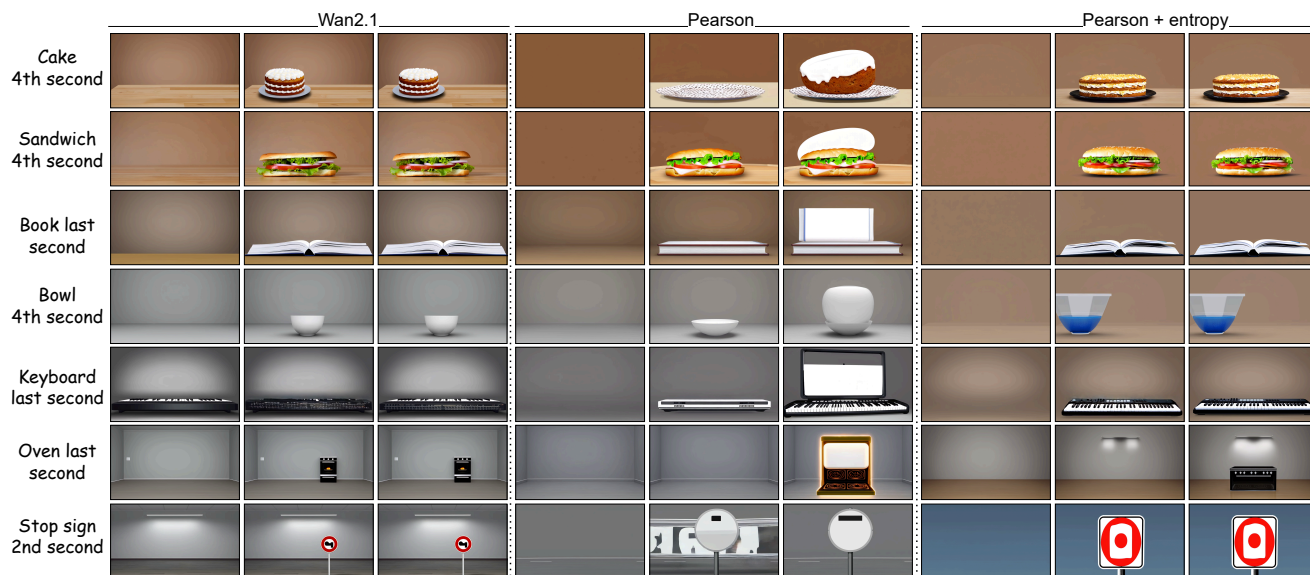


Figure 10. Comparison between Wan2.1 and Pearson-only optimization for single-object temporal control. Pearson correlation alone often satisfies the temporal constraint but alters object semantics, producing distorted or inconsistent appearances (e.g., the cake becoming misshaped, the sandwich warping, the book changing structure, or the keyboard turning into a laptop). Adding entropy regularization prevents these failures by stabilizing the object’s identity across time.

.11. Limitations of Explicit Temporal Cues in State-of-the-Art Text-to-Video Models

We analyze the effect of using explicit temporal phrasing (e.g., ‘in the third second’, ‘in the second part of the video’) in the prompt for the one-object setup. Table 12 compares Wan 2.1, Wan 2.2, and our method, each evaluated with and without such phrasing.

For Wan 2.1, explicit temporal cues do not improve temporal grounding. In fact, removing them increases Temporal Accuracy (63.94% to 65.18%) and leads to a clear gain in Imaging Quality (53.76% to 59.99%), indicating that the model struggles to interpret precise timing expressions and performs better without them.

Wan 2.2 shows a more nuanced pattern: it handles temporal phrasing better than Wan 2.1, likely due to improved data or training techniques. However, the drop in quality when explicit timing is introduced is still *significant*. Imaging Quality decreases sharply from 57.78% (without time) to 48.13% (with time), showing that temporal phrasing introduces confusion even for stronger models. Presence Accuracy also falls (79.50% to 68%), further highlighting that explicit timing signals continue to degrade performance.

Our method remains the most robust and accurate across settings. Across all conditions, our method provides the most reliable and precise temporal control, regardless of prompt style. Interestingly, although our approach builds on Wan 2.1 as the base model, explicit temporal phrasing does increase the temporal accuracy. This suggests that our optimization procedure may curate the base model’s timing capabilities.

Table 12. Effect of temporal phrasing in prompts for the one-object setup. Removing explicit timing improves video quality.

Method	Temp. Acc.	T. Abs. Acc.	T. Pres. Acc.	Imaging Quality
Wan2.1 with time	63.94%	67.38%	60.50%	53.76%
Wan2.1 without time	65.18%	65.12%	65.25%	59.99%
Wan2.2 with time	78.02%	87.42%	68.00%	48.13%
Wan2.2 without time	75.24%	71.25%	79.50%	57.78%
Ours with time	82.50%	83.25%	81.75%	56.51%
Ours without time	78.75%	82.25%	75.25%	57.14%

.12. Additional Qualitative Examples

We provide qualitative results across the one-object, two-object, and motion-based setups. For each example, we show sampled frames from the generated videos, illustrating the temporal grounding of object appearances. Prompts are abbreviated for brevity, indicating the key object and its expected time of appearance. Note that once introduced, the object is expected to persist until the end of the video. **Complete videos are included in the supplementary material. Note that images and videos were compressed to meet the CVPR supplementary file size limitations.**

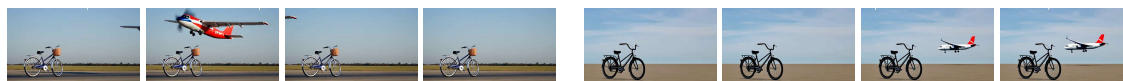
a baseball glove and a skateboard



a bicycle and a car



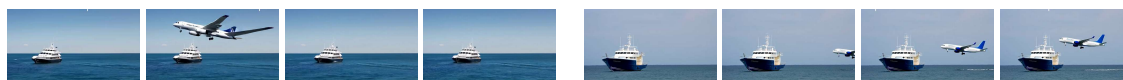
a bicycle and an airplane



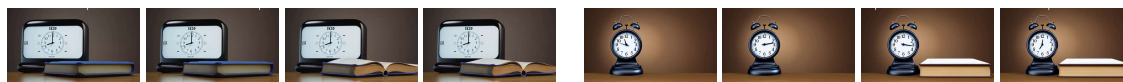
a bird and a cat



a boat and an airplane



a book and a clock



a bowl and a remote



a bus and a traffic light



a cake and a vase



a car and a motorcycle



a cat and a dog



a cell phone and a book



a donut and a suitcase



Figure 11. Two Objects Text vs Ours.

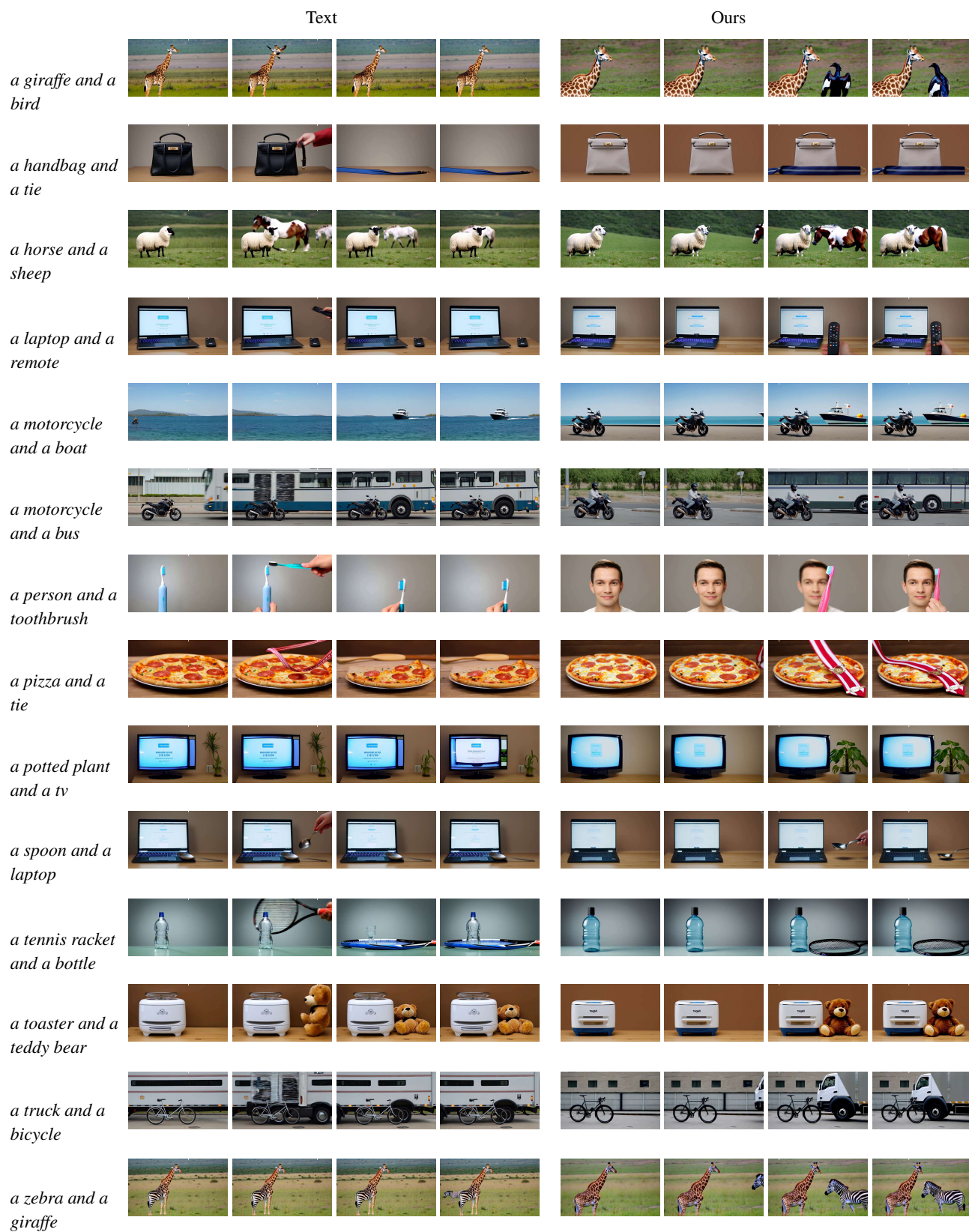


Figure 12. Two Objects Text vs Ours.

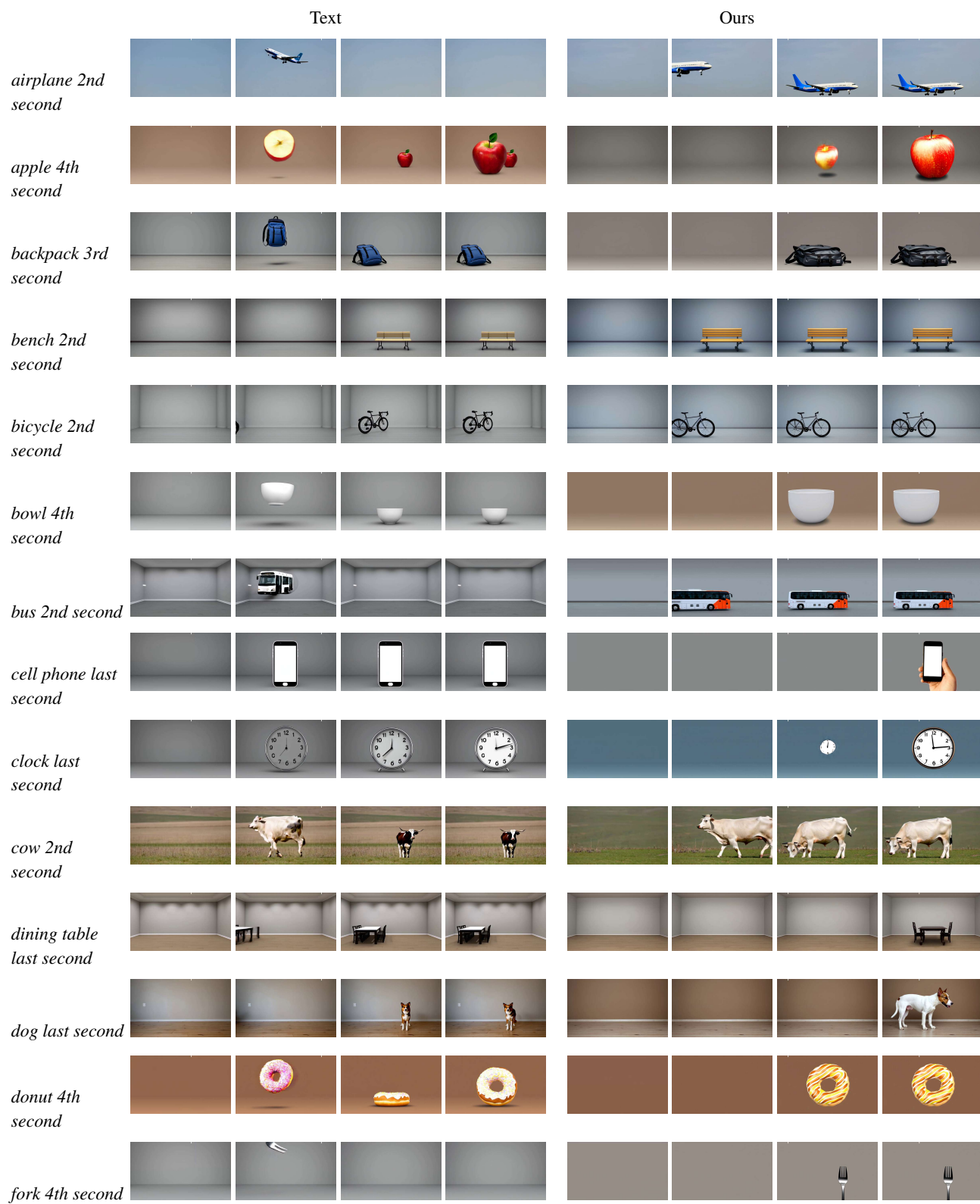


Figure 13. One Object Text vs Ours.



Figure 14. One Object Text vs Ours.

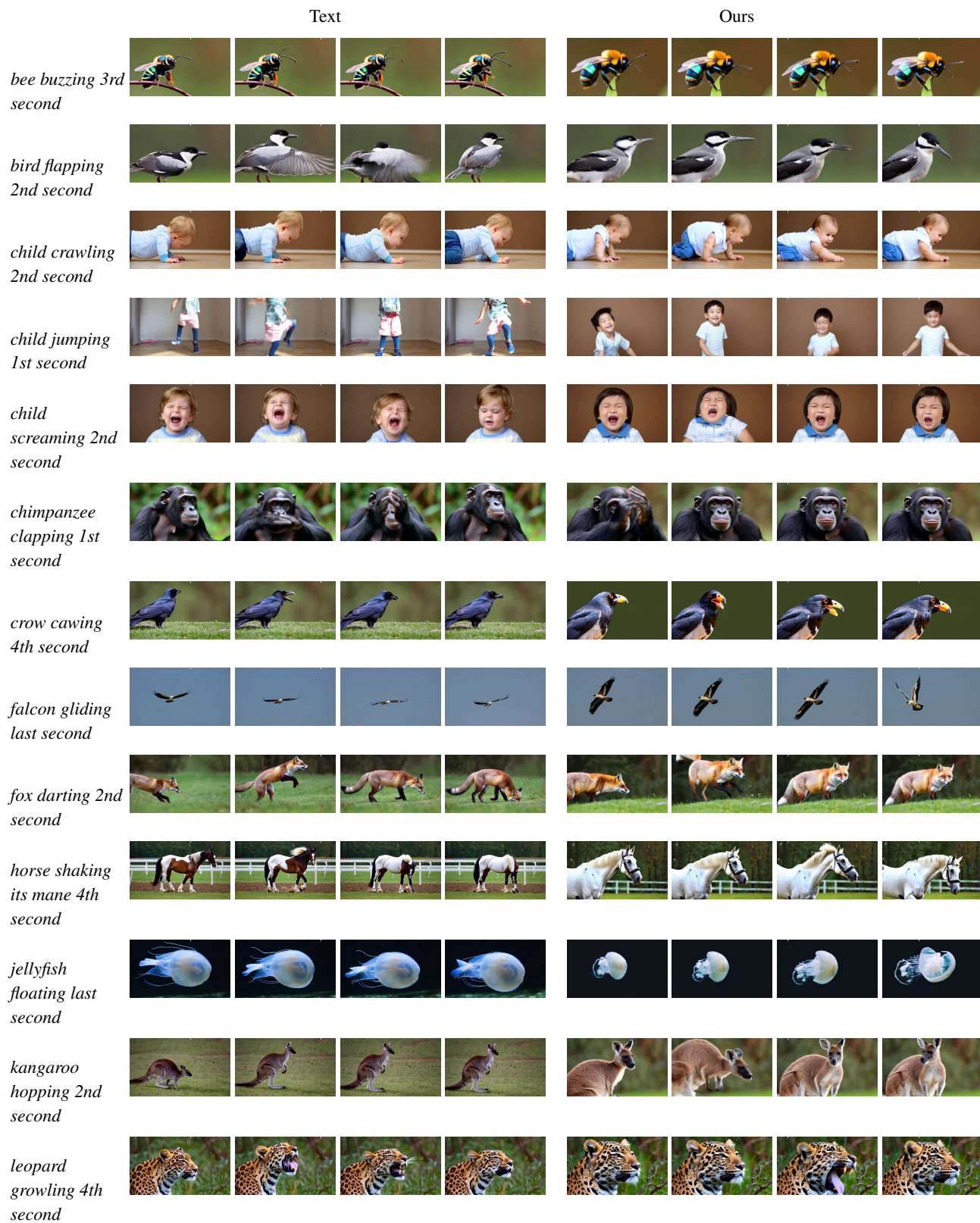


Figure 15. Movement Text vs Ours.

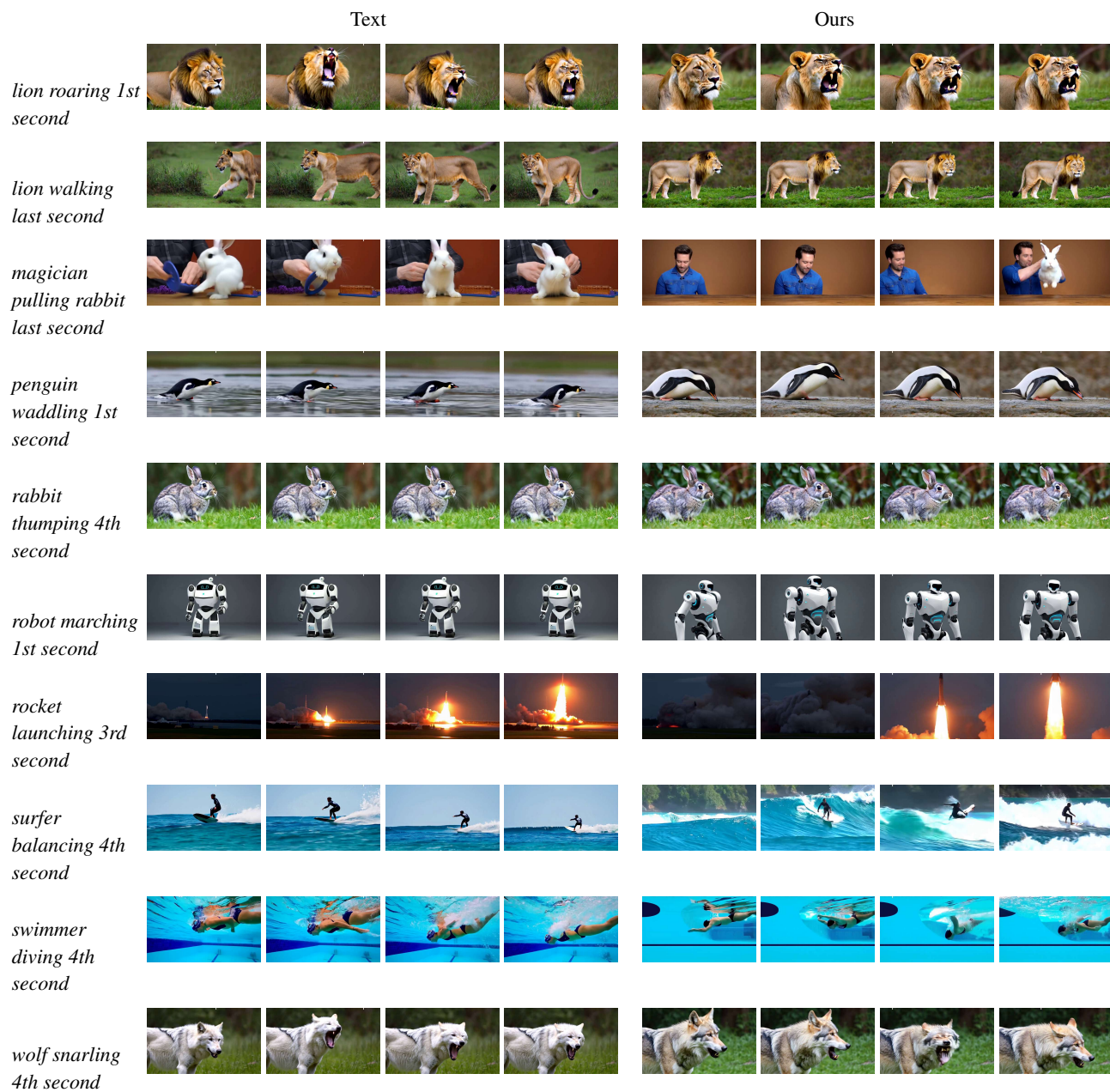


Figure 16. Movement Text vs Ours.

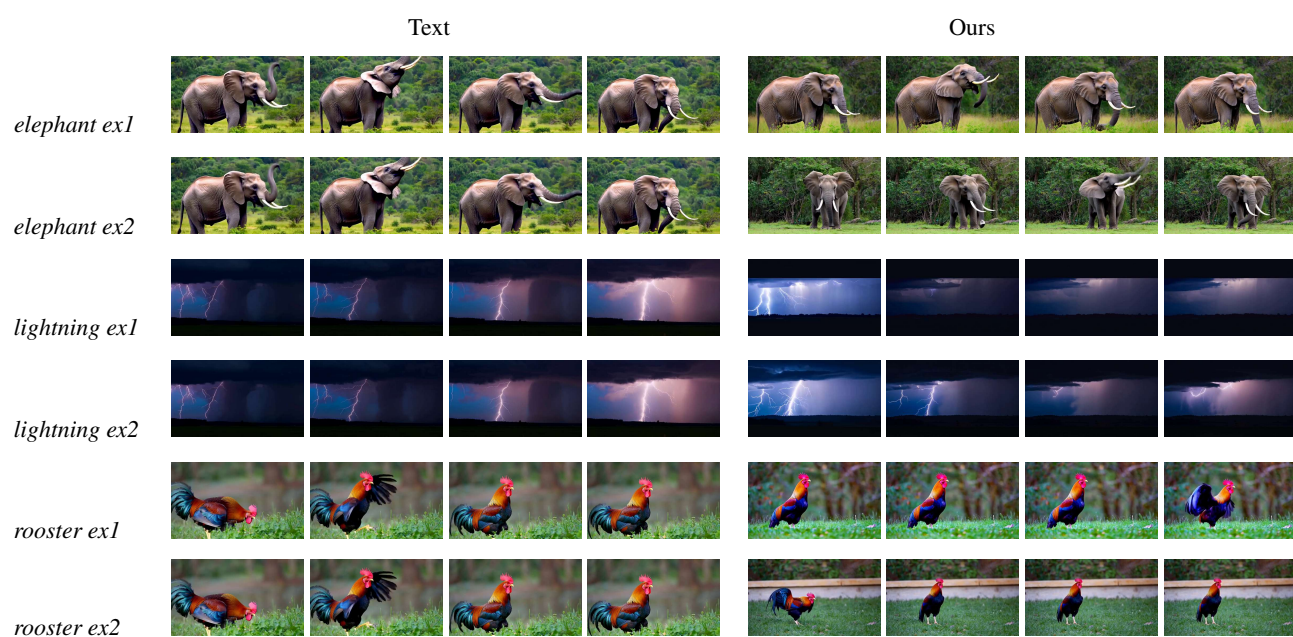


Figure 17. Audio-video Alignment Text vs Ours.