WaTeRFlow: Watermark Temporal Robustness via Flow Consistency

Supplementary Material

Overview

This supplementary material presents additional experimental results and further details of our proposed method, WaTeRFlow.

- Sec. A presents additional experimental results of WaTeRFlow.
- Sec. B provides additional experiments and explanations for the ablation study.
- Sec. C describes additional details of the experimental setup for the experiments presented in the main paper and the supplementary material.
- Sec. D provides the limitations of the proposed WaTeRFlow through failure cases.
- Sec. E provides additional qualitative evaluation results.

A. Additional Results

A.1. Robustness to Video Dynamics

In this section, we experimentally analyze how the bit accuracy changes with the degree of motion in videos generated via image-to-video (I2V) from watermarked images. We use the XT variant of Stable Video Diffusion (SVD) [1] as the video generation model, and vary the hyperparameter `motion_bucket_id`, which determines the degree of dynamics in the video, from its default value of 127 to 180 when generating videos from watermarked images. Larger values of this parameter lead to videos with more dynamic motion. Fig. 1a shows that, as `motion_bucket_id` increases, the overall height of the graph gradually decreases, which corresponds to the results of our proposed method, WaTeRFlow. Furthermore, Fig. 2 provides qualitative examples illustrating how dynamic the generated videos are when `motion_bucket_id` is set to the default value of 127 and to 180, respectively. In both cases, the frames are generated from watermarked images using WaTeRFlow.

A.2. Bit Accuracy with Long Video

In Fig. 4c of the main paper, we present the bit accuracy for frames generated by CogVideoX [8] as a plot. This result is obtained using only the first 25 frames out of the 49 frames generated by CogVideoX, so that it can be easily compared with the bit accuracy curve of SVD-XT shown in Fig. 4a of the main paper. Meanwhile, Fig. 1b shows, for WaTeRFlow and the baselines, the bit accuracy at the frames with even indices among all 49 generated frames. These results demonstrate that WaTeRFlow consistently achieves higher bit accuracy than the baselines over the entire range of frames, including those after the 25th frame.

A.3. Commercial Video Generation Service

In Fig. 4 of the main paper, we evaluated the robustness of each watermarking method by measuring bit accuracy using U-Net based SVD-XT [1] and Diffusion Transformer (DiT) based CogVideoX [8]. In this section, instead of open-source video generation models such as SVD-XT and CogVideoX, for which the papers and code are publicly available, we present results obtained with a commercial, closed-source AI video generation platform. Specifically, we conduct experiments using the Vidu Q2 model provided by Vidu [7]. Fig. 1c shows the qualitative evaluation results for the frames generated by Vidu. The 100 original images used in this experiment are identical to those used to produce the plots in Fig. 4 of the main paper. Fig. 1c shows that our proposed WaTeRFlow achieves the highest bit accuracy in the generated frames compared to baselines, even on a real commercial video generation platform.

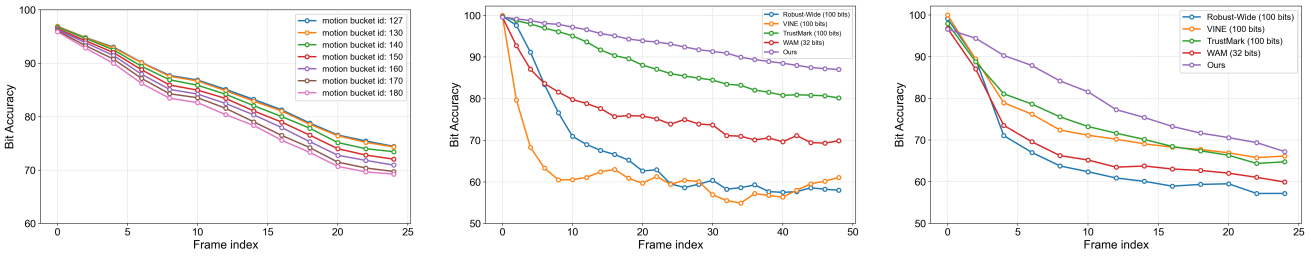
A.4. Analysis of Watermarked Image Quality

In the second None column of Tab. 1 in the main paper, we report the average bit accuracy over video frames generated by SVD-XT [1] for each image watermarking method. In this setting, excluding our proposed WaTeRFlow, VINE [5] achieves the highest average bit accuracy. However, in Tab. 2 of the main paper, VINE exhibits a relatively low PSNR. Figure 3 explains this by showing that embedding a watermark with VINE introduces strong noise along the image boundary, which in turn results in low PSNR. If this boundary region is treated as removable and a center crop is applied, PSNR increases significantly, but the watermark then becomes almost unrecoverable. This boundary-focused watermark signal also causes visible noise around frame borders in video frames generated from watermarked images, as seen in Fig. 3 in the main paper. In addition, further experiments with center and inverse center crops, reported in Tab. 1, show that baseline methods tend to hide watermark signals primarily near the center or the boundary of the image.

B. Additional Ablation Study

B.1. Semantic Preservation Loss

We conducted an ablation study related to the semantic preservation loss in Sec. 5.3 of the main paper. In this section, we present a more detailed analysis of how the bit accuracy of the first generated frame varies with the value of λ_{sem}. We ablate the weight λ_{sem} of the semantic preservation loss in Eq. 9 of the main paper. As shown in Table 2,



(a) Bit accuracy of WaTeRFlow on SVD-XT vs. `motion_bucket_id`

(b) I2V with CogVideoX

(c) I2V with Vidu [7]

Figure 1. **Per-frame bit accuracy and I2V robustness.** Each plot shows the bit accuracy over the generated frames. From left to right, we present the results of I2V generation from WaTeRFlow-watermarked images while varying `motion_bucket_id`, followed by the results obtained by feeding images watermarked by WaTeRFlow and the baselines into CogVideoX. Finally, we show the results of generating frames with Vidu using images watermarked by WaTeRFlow and the baselines.

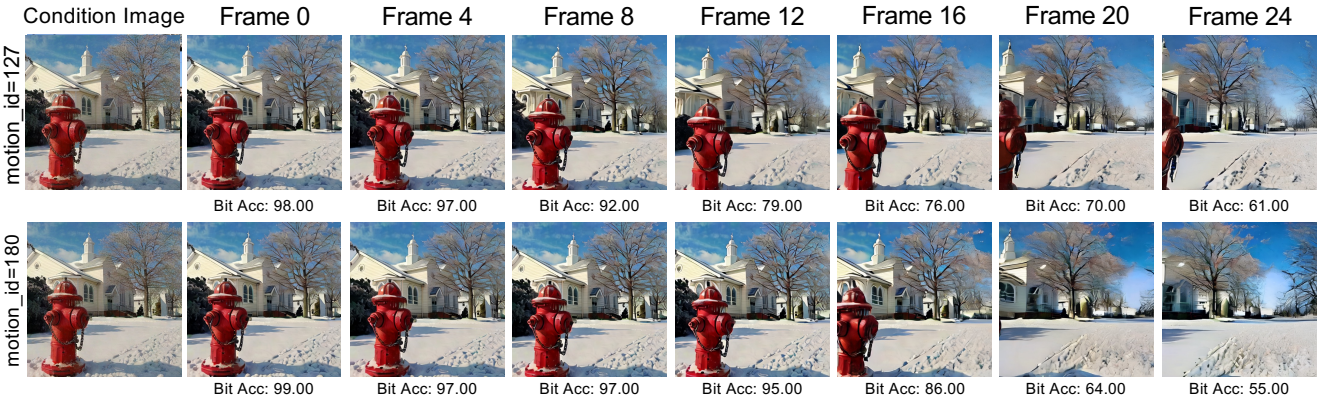


Figure 2. **Qualitative results with respect to `motion_bucket_id`.** The first and second rows show frames generated from watermarked images using our proposed WaTeRFlow with different motion_bucket_id values. Top: Frames generated with motion_bucket_id $= 127$. Bottom: Frames generated with motion_bucket_id $= 180$, in which the fire hydrant has completely disappeared in the last frame.

disabling this term, $\lambda_{\text{sem}} = 0$, reduces the first frame bit accuracy after I2V to 89.63%, whereas any non-zero weight in the range $[5 \times 10^{-5}, 10^{-3}]$ consistently yields 94.82–96.93% accuracy, i.e., an absolute gain of about 5–7 percentage points. This confirms that preserving the semantics between the original and watermarked images is important for retaining the embedded watermark in the first frame generated by the I2V model.

C. Further Experimental Details

C.1. CogVideoX Text Prompts

This section describes the text prompts used when generating videos with CogVideoX [8]. Tab. 4 shows the text prompts used in the CogVideoX experiments throughout the main paper and the supplementary material. In particular, they correspond to the prompts associated with the 100 images used in the experiment shown in Fig. 4 of the main paper. The listed text prompts were applied identically to both our method and the baseline when generating videos from watermarked images using CogVideoX.

C.2. Resolution Scaling

TrustMark [2] and VINE [5] propose, as shown in Algorithm 1, a method to adapt any watermarking model so that it can handle arbitrary image resolutions. This approach is designed to preserve the visual quality of watermarked images while maintaining the model's robustness to image transformations it can inherently handle at the native resolution used during training. In all experiments reported in the main paper and the Supplementary Material, we apply this resolution-scaling procedure uniformly to all image watermarking methods so that they can be fairly evaluated at a common resolution of 512×512.

D. Failure Case

We consider image regeneration as a representative attack in which an adversary attempts to remove embedded image watermarks. In this process, a given image is mapped to a noisy intermediate state along the diffusion trajectory and then denoised back to the image space by a trained diffusion model, producing an output that is visually similar but resampled at the pixel level. Because this procedure weakens fine textures and watermark signals, diffusion-based regen-

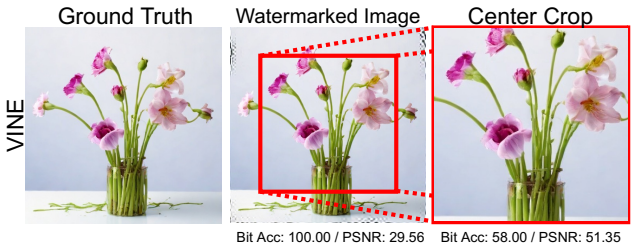


Figure 3. **Qualitative results of VINE.** This figure visually illustrates why images watermarked with VINE [5] yield low PSNR. From left to right, we show the original image, the watermarked image, and the image obtained by applying a center crop that retains 70% of the height and width. Below each image, we report the corresponding bit accuracy and PSNR.

Method	Cap. (Bits)	Watermarked Image		
		|Center − Inverse| (%) ↓	Bit Acc. (%) ↑	
			Center Crop (70%)	Inverse Center Crop (70%)
Robust-Wide [4]	100	3.26	82.33	79.07
VINE [5]	100	43.27	56.73	99.99
TrustMark [2]	100	29.57	97.98	68.41
WAM [6]	32	20.54	75.75	96.29
WaTeRFlow (Ours)	100	**1.35**	78.73	77.38

Table 1. **Comparison of center and inverse center cropping on watermarked images.** We report bit accuracy under center crop and inverse center crop on watermarked images and show the difference between the two cases. These results indicate that existing baseline methods tend to embed watermark signals primarily in the central or peripheral regions of the image.

eration constitutes a strong watermark removal attack. In Tab. 1 of the main paper, we regenerate watermarked images using a VP noise schedule [3] and then measure the average bit accuracy in the subsequent image-to-video generation. The graph in Fig. 4 shows the average bit accuracy of our method and the baselines over frames generated by SVD-XT [1] under a stochastic regeneration setting, as we vary the noise step. Our WaTeRFlow achieves the highest average bit accuracy across all evaluated noise steps, though its accuracy falls below 80% once the noise step exceeds 200. As shown by the qualitative results in Fig. 4, such attacks using large noise can suppress watermarks but severely degrade the visual quality of the original image, which limits their practicality as realistic attack scenarios.

E. Additional Qualitative Results

In this section, we provide additional qualitative evaluations beyond those presented in the main paper.

E.1. Heatmap

In this section, we present the original images, the watermarked images, and the heatmaps between them. The heatmaps are generated as follows. First, we compare the original and watermarked images pixel by pixel and combine the changes in the RGB channel values at each location into a single scalar value to calculate a grayscale difference.

Algorithm 1: Resolution scaling

Input: Input image I, binary watermark w
Output: Watermarked image I_w
Model: Encoder E trained on the resolution of $U \times V$

$H, W \leftarrow \text{Size}(I)$
$I \leftarrow I/127.5 - 1$
$I' \leftarrow \text{interpolate}(I, (U, V))$
$\text{res}' \leftarrow E(I') - I'$
$\text{res} \leftarrow \text{interpolate}(\text{res}', (H, W))$
$I_w \leftarrow \text{clamp}(I + \text{res}, -1, 1)$
$I_w \leftarrow I_w \times 127.5 + 127.5$

λ_{sem}	Bit Acc. (%) First frame	PSNR ↑	SSIM ↑	LPIPS ↓
10^{-3}	96.93	38.83	0.9902	0.0291
5×10^{-4}	96.22	38.18	0.9873	0.0351
10^{-4}	95.99	38.19	0.9873	0.0358
5×10^{-5}	94.82	38.39	0.9906	0.0310
0 (no $\mathcal{L}_{\text{sem}}$)	89.63	40.44	0.9920	0.0299

Table 2. **Ablation on semantic preservation loss.** λ_{sem} denotes the weight of the semantic preservation loss introduced in Eq. 9 of the main paper. PSNR, SSIM, and LPIPS measure the quality of the watermarked images produced by WaTeRFlow. The row no $\mathcal{L}_{\text{sem}}$ indicates the setting where all other contributions of our method are kept, but the semantic preservation loss term is removed.

Next, based on the distribution of these difference values, we rescale them to the range $[0, 1]$, such that values with almost no change fall at the lower end of the range and very large changes move to the upper end. Finally, we apply a colormap to this normalized heatmap, rendering regions with small changes in blue and regions with large changes in red, thereby enabling an intuitive visualization of areas that are strongly affected by watermark insertion. The corresponding results can be found in Fig. 5.

E.2. Additional I2V Qualitative Results

For additional qualitative results on videos generated from watermarked images using SVD-XT [1] and CogVideoX [8], please refer to the project page provided together with the Supplementary Material.

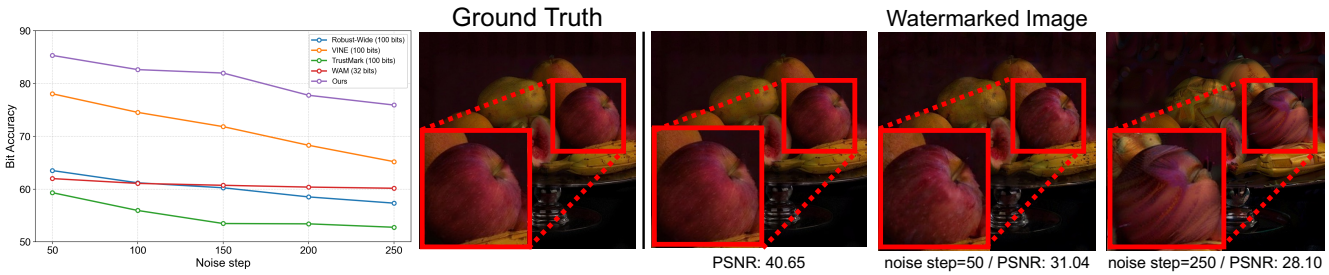


Figure 4. **Quantitative and Qualitative Results of Image Regeneration.** From left to right, we first plot the *average bit accuracy over frames* produced by the I2V generation after stochastically regenerating the watermarked image with an increasing number of noise steps. Next, we sequentially show the original image and the watermarked image generated by our proposed method, WaTeRFlow. For the watermarked image, we visualize two cases, one with, regeneration using a relatively low number of noise steps and the other with regeneration using a relatively high number of noise steps.

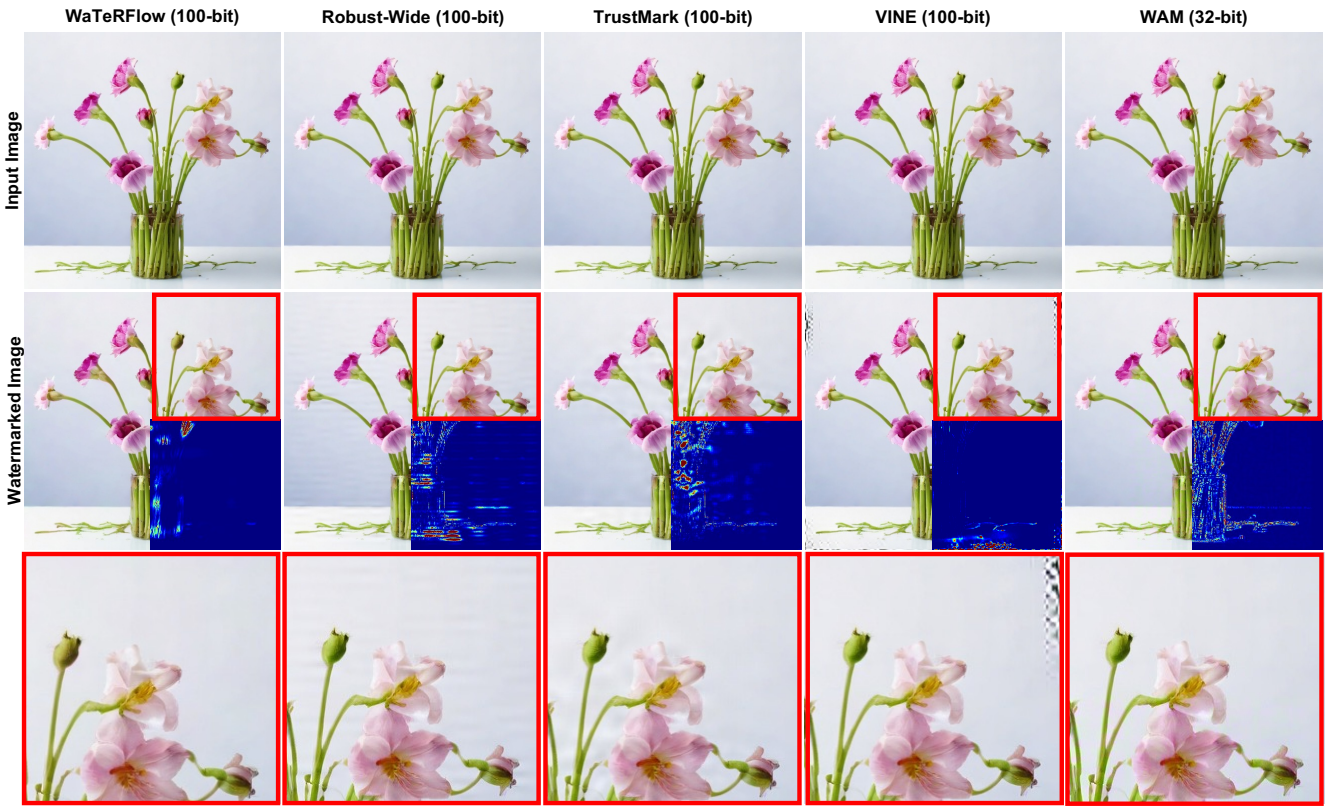


Figure 5. **Watermarked Image and Heatmap.** Along with the original image, we sequentially present the resulting images obtained by inserting watermarks using WaTeRFlow and the baselines, as well as the corresponding heatmaps that visualize the regions altered by the watermark as if seen through an X-ray. Top: Original image. Middle: Watermarked images generated by each method and their corresponding heatmaps. Bottom: Zoomed-in crops of the watermarked images.

Index	Prompt
0	The image depicts a charming scene featuring three teddy bears interacting with a small birdhouse in an outdoor setting.
1	The image shows a section of railway tracks with a focus on the foreground where the tracks are clearly visible, showing signs of wear and tear such as rust and discoloration.
2	The image shows a dog dressed in a black leather jacket with a white collar, sitting on the seat of a red motorcycle.
3	The image shows two plush teddy bears sitting side by side against a plain white background.
4	The image shows a neatly made bed with a patchwork quilt featuring various patterns in shades of yellow, gray, and white.
5	The image shows two sandwiches placed inside an oven.
6	The image depicts an urban street scene with a focus on a stop sign in the foreground.
7	The image appears to be the cover of a National Geographic magazine.
8	The image shows two sandwiches being toasted in an oven.
9	The image depicts a serene marina scene with a prominent statue of a giraffe in the foreground.
10	The image depicts a man sitting in a modern office environment, holding a black tablet device in his hands.
11	The image depicts the interior of a modern train carriage.
12	The image depicts a young woman lying on her side on a bed, her head resting on a pillow and her body partially covered by a blue sheet or blanket.
13	The image depicts an indoor equestrian arena where several riders on horseback are engaged in activities, likely practicing or participating in a competition.
14	The image depicts a cozy living room decorated for the Christmas season.
15	The image shows a man with short, dark hair, dressed in a formal blue suit jacket and light blue dress shirt, sitting comfortably in a modern office environment.
16	The image shows a well-worn plush teddy bear sitting upright on a wooden surface in an antique shop or museum setting.
17	The image depicts a steaming cup of coffee placed on a saucer.
18	The image depicts a classroom setting with several students seated at desks.
19	The image depicts a young woman standing in a well-lit kitchen, placing or removing a baking dish from a stainless steel oven.
20	The image shows a medium-sized brown dog, possibly a mixed breed, standing on a windowsill.
21	The image shows a plush brown teddy bear sitting on a light-colored surface in front of a wall that has a height measurement chart painted on it.
22	The image depicts a classroom setting with several students seated in rows, concentrating on writing or taking an exam.
23	The image depicts a classroom setting with several students seated at desks, each working on assignments or taking notes.
24	The image depicts a construction worker wearing high-visibility clothing operating a hydraulic excavator on a construction site.
25	The image shows a busy classroom environment where multiple students are seated at desks, working intently on tasks or exams.
26	The image shows a young girl with long blonde hair, dressed in a white top and blue jeans, standing on a grassy lawn in a park or backyard setting.
27	The image depicts a group of school-aged children seated around a rectangular table in a classroom setting.
28	The image appears to be a scanned or photographed page from a book or academic text written in German.
29	The image depicts a man walking down a city street at night.
30	The image shows the interior of a classroom filled with students seated at desks arranged in rows.
31	The image shows an indoor ice hockey rink where a pickup game is taking place.
32	The image shows a close-up view of a person's torso and arms as they sit at a desk or table, using a pen to write in an open notebook or planner.
33	The image depicts an outdoor poolside scene at what appears to be a residential or private property.
34	The image shows a dog partially outside a window or door, holding a plush toy of a squirrel in its mouth.
35	The image shows a medium-sized light brown dog standing on a windowsill looking outside.
36	The image depicts a close-up of a teddy bear with long, furry brown hair.
37	The image depicts an indoor setting where several individuals are engaged in what appears to be a discussion or meeting.
38	The image appears to be a scanned or printed page from a German-language academic or technical book.

Index	Prompt
39	The image depicts a maritime scene featuring a large sailing ship docked at a harbor.
40	The image depicts a dog standing on the windowsill of a house, looking outside.
41	The image depicts an indoor scene where a man is sitting in a bar or pub setting.
42	The image shows a light brown teddy bear wearing a red sweater with the word London and the letter E on it, sitting in front of a scenic background featuring the Big Ben clock tower and the Palace of Westminster in London.
43	The image depicts a quaint village street lined with traditional stone cottages featuring chimneys and slate roofs.
44	The image shows a person performing a bicycle stunt in mid-air, likely in a skate park or similar recreational area.
45	The image shows a woman sitting at a table in a bright office or meeting room, engaged in a discussion or presentation.
46	The image shows a close-up of a can of Bunnaberg & Cola.
47	The image shows a group of young women sitting in what appears to be a classroom or training room, wearing bright green vests and smiling towards the camera.
48	The image depicts a panel discussion or interview setting involving three individuals seated on high chairs or stools in front of a bright yellow backdrop.
49	The image shows two action figures of Teenage Mutant Ninja Turtles, specifically Leonardo and Donatello, riding skateboards.
50	The image depicts a black folder laying open on a table with several documents neatly arranged on both sides.
51	The image shows a young woman in a white lab coat standing at a lab bench in a classroom or laboratory setting.
52	The image shows a group of individuals seated in what appears to be a classroom or training room setting.
53	The image depicts a large yellow truck, likely used for industrial or construction purposes, parked on a paved area near a modern building.
54	The image depicts a dog standing on the ocean shore during sunset.
55	The image shows a football player in action on the field during a game.
56	The image depicts a black and white portrait of a dog, possibly a mixed breed or a spaniel type.
57	The image shows a group of people sitting in a classroom or training room setting.
58	The image shows a stadium filled with a crowd of enthusiastic fans, many of whom are wearing yellow shirts or jerseys and waving yellow flags.
59	The image shows a professional ice hockey game in progress.
60	The image depicts a woman social worker or case manager visiting an elderly man in his home.
61	The image shows two brown teddy bears sitting on a couch or sofa.
62	The image depicts a close-up of a playful puppy standing on a beach or sandy area near a waterfront.
63	The image depicts a close-up of a calculator with a pencil resting on an open notebook or planner.
64	The image depicts a classic London street scene with an iconic red double-decker bus prominently featured in the foreground.
65	The image appears to be a scanned or photographed page from an academic book or textbook in German.
66	The image shows a serene beach scene with two wooden lounge chairs placed on a white sandy shore.
67	The image shows an outdoor swimming pool in a residential backyard or small resort setting.
68	The image shows a busy job interview or career fair setting.
69	The image appears to be a scanned or photocopied page from a German-language book.
70	The image depicts a modern office building with a glass facade located in an urban setting.
71	The image depicts a lively outdoor music concert or festival scene.
72	The image depicts a dog lying down and resting its head on what appears to be the leg or knee of a person.
73	The image shows a dog standing on a windowsill partially obscured by white curtains on either side.
74	The image depicts a panel discussion or interview setup in a studio environment.
75	The image depicts a motorcycle race on a racetrack, featuring two riders leaning into a curve at high speed.
76	The image shows a black and white photograph of a football locker room.
77	The image depicts a man standing in a kitchen or living room area, holding a yellow cleaning bucket in his right hand.
78	The image shows a dog standing on the windowsill of a house or building, looking outside through an open window.
79	The image shows a close-up view of an audio mixing console in a recording studio or live sound setting.
80	The image depicts a physical therapy or rehabilitation session taking place in a medical or clinical setting.

Index	Prompt
81	The image depicts a mother and her newborn baby in a hospital room shortly after birth.
82	The image shows three green military-style armored trucks parked in a line on a dirt or gravel surface.
83	The image depicts a pair of boots displayed in an artistic arrangement.
84	The image depicts a group of seven people in an office setting, all smiling and giving thumbs-up gestures towards the camera.
85	The image depicts a modern laboratory setting with several individuals engaged in scientific research or experimentation.
86	The image depicts a classic red double-decker bus, reminiscent of those commonly seen in London, England.
87	The image shows the facade of a brick building with a prominent clock mounted on it.
88	The image shows a British Rail Class 47 diesel-electric locomotive at a railway station platform.
89	The image depicts a smartphone with a digital representation of a credit card displayed on its screen.
90	The image captures a skateboarder performing a trick in an urban skate park setting.
91	The image depicts a clock tower situated in an urban environment.
92	The image shows the iconic clock face of Big Ben, a part of the Elizabeth Tower in London, England.
93	The image shows a bathroom with a toilet as the central focus.
94	The image depicts a scene from a theatrical performance or a play.
95	The image depicts a charming urban scene featuring a clock tower that stands prominently in the center.
96	The image shows a white vehicle parked on a street or driveway.
97	The image appears to be a promotional poster for an event or performance.
98	The image depicts an urban street scene at dusk or early evening, characterized by the illuminated windows of tall buildings in the background, suggesting a cityscape with modern architecture.
99	The image features a person standing in front of a red, vertically slatted background.

Table 4. **Text prompts used in the CogVideoX experiments.** Text prompts corresponding to the 100 images used in the experiment shown in Fig. 4 of the main paper, which are used to generate videos with CogVideoX [8] throughout the main paper and the Supplementary Material. The same prompts are applied to both our method and the baselines when generating videos from watermarked images.

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1, 3

[2] Tu Bui, Shruti Agarwal, and John Collomosse. Trustmark: Robust watermarking and watermark removal for arbitrary resolution images. In *IEEE International Conference on Computer Vision (ICCV)*, 2025. 2, 3

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 3

[4] Runyi Hu, Jie Zhang, Ting Xu, Jiwei Li, and Tianwei Zhang. Robust-wide: Robust watermarking against instruction-driven image editing. In *European Conference on Computer Vision*, pages 20–37. Springer, 2024. 3

[5] Shilin Lu, Zihan Zhou, Jiayou Lu, Yuanzhi Zhu, and Adams Wai-Kin Kong. Robust watermarking using generative priors against image editing: From benchmarking to advances. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2, 3

[6] Tom Sander, Pierre Fernandez, Alain Durmus, Teddy Furon, and Matthijs Douze. Watermark anything with localized messages. In *International Conference on Learning Representations (ICLR)*, 2025. 3

[7] Vidu. Vidu ai: AI video generator – text & image to video in seconds. https://www.vidu.com/, 2025. Online video generation platform; accessed Nov. 16, 2025. 1, 2

[8] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Xu Bin, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Representation Learning*, pages 83048–83077, 2025. 1, 2, 3, 7