ByTheWay: Boost Your Text-to-Video Generation Model to Higher Quality in a Training-free Way

Supplementary Material

In the supplementary material, we present additional qualitative results (Section A), more ablation experiments (Section B), details of our user study and MLLM assessment (Section C), more discussion on Temporal-Self Guidance (Section D), the proof of Fourier-based Motion Enhancement (Section E), as well as the limitation of our method (Section F), as a supplement to the main paper.

A. Additional Qualitative Results

More Results on AnimateDiff. We present more results for video motion enhancement (Fig. 7) and structure enhancement (Fig. 8) on AnimateDiff.

More Results on VideoCrafter2. We present more results for video motion enhancement (Fig. 9) and structure enhancement (Fig. 10) on VideoCrafter2.

Application on DiT-based T2V model. We validated the potential of ByTheWay on a DiT-based T2V backbone, CogVideoX-2B [7]. As depicted in Fig. 3, ByTheWay effectively facilitates enhanced structure and improved dynamics. Implementation details are listed below. *Temporal Self-Guidance.* For each DiT block (except the first one), we modify its attention map by fusing it with that in previous block, improving alignment in motion modeling for enhanced temporal consistency. *Fourier-based Motion Enhancement.* It is achieved by amplifying the high-frequency component of attention map, enabling amplified energy for enhanced motion dynamics.

Visual comparison with other baselines. As depicted in Fig. 4, both UniCtrl [2] and FreeU [6] fail to improve video temporal consistency. I4VGen [3] employs static image for initialization, suppressing motion dynamics. In comparison, the proposed ByTheWay effectively enhances temporal consistency and dynamic degree.

B. Additional Ablation Study

Choice of Guidance Anchor. Fig. 5 (a) demonstrates that `up_blocks.1` is the bottleneck in video motion modeling, injecting the temporal attention map of `up_blocks.1` into subsequent decoder blocks helps align motion modeling across different levels of diffusion U-Net, thus enhance temporal consistency. In contrast, injecting information from later blocks fails to achieve this goal. Note that using `up_blocks.3` as the anchor implies the absence of Temporal Self-Guidance (vanilla result).

Number of Operation Steps. Fig. 6 reveals that the initial 20% of sampling steps play a crucial role in shaping video motion, making the application of ByTheWay operations beyond this point have minimal effect on the generation quality. Moreover, when ByTheWay operations are applied only during 20% to 80% of the sampling steps, the generated video appears almost identical to the original video, which can be attributed that video motion is mainly determined by the early denoising stage.

Does More Sampling Steps Help? As shown in Fig. 5 (b), the vanilla T2V backbone with $5\times$ sampling steps is inferior to incorporating the ByTheWay-enhanced backbone with only $1\times$ sampling steps, this demonstrates that ByTheWay is not equivalent to simply increasing the DDIM sampling steps.

C. Details of User Study & MLLM Assessment

User Study Details. In our user study, each participant receives 50 videos synthesized by Vanilla T2V backbones and 50 videos synthesized by ByTheWay-enhanced backbones. These videos are sampled from the same random seeds to ensure fair comparison. For each video pair from Vanilla and Vanilla + ByTheWay, participants are required to select the video they perceive as superior based on overall *Video Quality*, considering both structure coherence and motion magnitude, and cast their vote accordingly. The videos were presented in a randomized order to reduce potential bias, and participants were allowed ample time to review each pair before making their selections.

MLLM Assessment Prompt. Here, we present the prompt used in the MLLM assessment.

"""

You are provided with two sets of video frames, each containing 4 representative frames, along with a shared textual prompt that was used to generate both videos. Your task is to perform a comparative evaluation of the two videos, focusing on their structure rationality / motion consistency.

Here is the frame data of Video_1.

Here is the frame data of Video_2.

Based on your evaluation of motion consistency, choose the video set you find to be superior. If you determine that the first set of frames (Video_1) is better, respond with "A". If the second set (Video_2) is superior, respond with "B". Return only "A" or "B" based on your assessment.

"""

D. Discussion on Temporal-Self Guidance

Why operate in the upsampling blocks? Previous works like MasaCtrl [1], FreeControl [5], and MotionClone [4] have validated that the per-frame structure of generated videos is mainly modeled by upsampling blocks in diffusion U-Net architecture. Given that video motion consists of per-frame structure, we apply temporal self-guidance in the upsampling blocks.

What problems Temporal-Self Guidance can help? Temporal-Self Guidance can address the temporal inconsistency issues arising from inconsistent modeling between different temporal attention blocks, such as implausible structural changes between frames, as shown in Fig. 1. However, Temporal-Self Guidance may fail in fixing structural anomalies caused by model's inherent cognitive limitation(e.g., a deformed ball), and struggles to handle real video issues while maintaining fidelity since ByTheWay mainly operates in the early denoising stage.

Figure 1. **Temporal-Self Guidance reduces temporal inconsistency and structural changes between frames.**

Motivation of Temporal-Self Guidance. In Fig. **??**(a)-(b), the x-axis is the denoising step, and the y-axis is the L2 difference of temporal attention maps across different upsampling blocks (lines represent mean and variance). It can be observed that implausible structures and temporal inconsistencies are often associated with significant differences of attention maps across up_blocks, **which motivates us to reduce such discrepancy** for enhanced consistency in structural modeling at different upsampling blocks.

E. Fourier-based Motion Enhancement Proof

In this section, we provide a detailed proof of how Fourier-based Motion Enhancement alters the energy of the temporal attention map in ByTheWay operations.

E.1. Frequency Components Manipulation

Given a temporal attention map $\mathcal{A} \in \mathbb{R}^{(B\times H\times W)\times F\times F}$ with batch size B, spatial resolution $H\times W$ and frame number F, since we treat it as a batch of 1D attention sequences, we will next discuss the operations performed on a single softmax sequence $x[n]$ of length F.

Mathematically, the operation of mapping the sequence $x[n]$ to the frequency domain is performed by the Discrete Fourier Transform (DFT):

$$X[k] = \sum_{n=0}^{F-1} x[n] \cdot e^{-j\frac{2\pi}{N}kn},\ k = 0, 1, \ldots, F-1. \quad (1)$$

Parseval's theorem states that the energy of a sequence is preserved under frequency domain transformation, meaning that the energy E_x of sequence $x[n]$ is the same in both the time and frequency domains. This theorem can be expressed as follows:

$$E_x = \sum_{n=0}^{F-1} x[n]^2 = \frac{1}{F}\sum_{k=0}^{F-1} X[k]^2. \quad (2)$$

As mentioned in the main paper, Fourier-based Motion Enhancement uses a threshold index τ to separate the high-frequency and low-frequency components of the sequence, scaling the high-frequency components by a factor of β. This operation can be expressed as:

$$X^{'}[k] = \begin{cases} \beta \cdot X[k] & k \in [\frac{F}{2}-\tau, \frac{F}{2}+\tau], \\ X[k] & \textit{otherwise}, \end{cases} \quad (3)$$

after applying this manipulation, the energy $E_x^{'}$ of current attention sequence $x^{'}[n]$ is given by:

$$E_x^{'} = \frac{1}{F}[\sum_{k\notin[\frac{F}{2}-\tau,\frac{F}{2}+\tau]} X^2[k] + \beta^2 \sum_{k\in[\frac{F}{2}-\tau,\frac{F}{2}+\tau]} X^2[k]], \quad (4)$$

thus the energy change amount ΔE caused by Fourier-based Motion Enhancement can be computed as:

$$\begin{aligned} \Delta E &= E_x^{'} - E_x \\ &= \frac{(\beta^2-1)}{F} \sum_{k\in[\frac{F}{2}-\tau,\frac{F}{2}+\tau]} X^2[k]. \end{aligned}$$

Clearly, in the scenario where $\beta > 1$, Fourier-based Motion Enhancement will lead to an increase in the energy of the attention sequence ($\Delta E > 0$), while the opposite will result in a decrease in energy ($\Delta E < 0$), which elucidates the mechanism by which Fourier-based Motion Enhancement effectively enhances motion magnitude in synthesized videos.

Furthermore, it can be demonstrated that the attention sequence processed by Fourier-based Motion Enhancement remains a softmax sequence. This property is preserved because the direct current (DC) component $X[0]$ of the attention sequence, which determines the sum of the sequence, is not modified throughout the operation. By plugging $k = 0$ into Eq. 1, we can ascertain this property:

$$X[0] = \sum_{n=0}^{F-1} x[n] = \sum_{n=0}^{F-1} x^{'}[n] = 1. \quad (5)$$

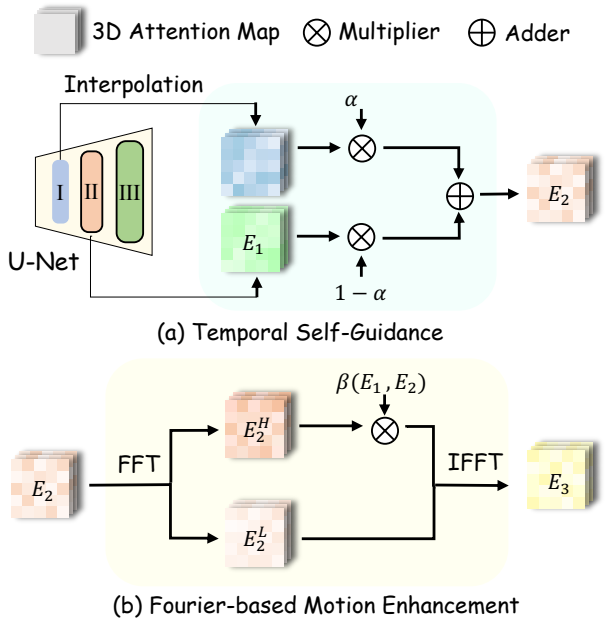

Figure 2. **ByTheWay Operations.**

E.2. Adaptive β in ByTheWay Operations

As depicted in the Fig. 2, let E_1 denote the the energy of the temporal attention map before applying ByTheWay operations, E_2 the energy after Temporal Self-Guidance, and E_3 the energy after Fourier-based Motion Enhancement. Here, we demonstrate that using the adaptive β as defined in Eq. 6 ensures that $E_3 \geq E_1$.

$$\beta(E_1, E_2) = max\{\beta_0, \sqrt{\frac{E_1 - E_2^L}{E_2^H}}\}, \tag{6}$$

Based on the separation of high-frequency and low-frequency components in the sequence as described in Section E.1, we can compute the energy of the high-frequency and low-frequency parts of the sequence $x[n]$, denoted as E_x^H and E_x^L, respectively:

$$\begin{aligned} E_x^H &= \frac{1}{F} \sum_{k \in [\frac{F}{2}-\tau, \frac{F}{2}+\tau]} X^2[k], \\ E_x^H &= \frac{1}{F} \sum_{k \notin [\frac{F}{2}-\tau, \frac{F}{2}+\tau]} X^2[k]. \end{aligned} \tag{7}$$

According to Eq. 2 and Eq. 7, it is evident that the following relationship holds:

$$E_x = E_x^H + E_x^L. \tag{8}$$

Furthermore, we can concisely express the energy manipulation performed by Fourier-based Motion Enhancement described in Section E.1, as follows:

$$E_x^{'} = \beta^2 E_x^H + E_x^L, \tag{9}$$

which indicates:

$$E_3 = \beta^2 E_2^H + E_2^L. \tag{10}$$

Therefore, to ensure $E_3 \geq E_1$, it is necessary to ensure that β adheres to the following condition:

$$\beta^2 E_2^H + E_2^L \geq E_1, \tag{11}$$

the critical value of β, denoted as β_c, that satisfies this condition is:

$$\beta_c = \sqrt{\frac{E_1 - E_2^L}{E_2^H}}. \tag{12}$$

In ByTheWay operations, the user-specified β, denoted as β_0, will be compared with the critical value β_c, and the larger of the two will be selected as the actual β value in Fourier-based Motion Enhancement:

$$\beta = max\{\beta_0, \beta_c\}. \tag{13}$$

By adopting such a adaptive β value, it can be theoretically guaranteed that the energy of the temporal attention map is increased during ByTheWay operations, thereby enhancing the motion magnitude in synthesized videos.

F. Limitation

Although ByTheWay demonstrates the capability to unlock the synthesis potential of various T2V backbones, the synthesized videos remain confined within the sampling distribution of the original T2V backbone. Therefore, one limitation of our method is that its performance upper bound is still constrained by the original T2V backbone.

References

[1] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22560–22570, 2023. 2

[2] Xuweiyi Chen, Tian Xia, and Sihan Xu. Unictrl: Improving the spatiotemporal consistency of text-to-video diffusion models via training-free unified attention control. *arXiv preprint arXiv:2403.02332*, 2024. 1

[3] Xiefan Guo, Jinlin Liu, Miaomiao Cui, and Di Huang. I4vgen: Image as stepping stone for text-to-video generation. *arXiv preprint arXiv:2406.02230*, 2024. 1

[4] Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. Motionclone: Training-free motion cloning for controllable video generation. *arXiv preprint arXiv:2406.05338*, 2024. 2

[5] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. *arXiv preprint arXiv:2312.07536*, 2023. 2

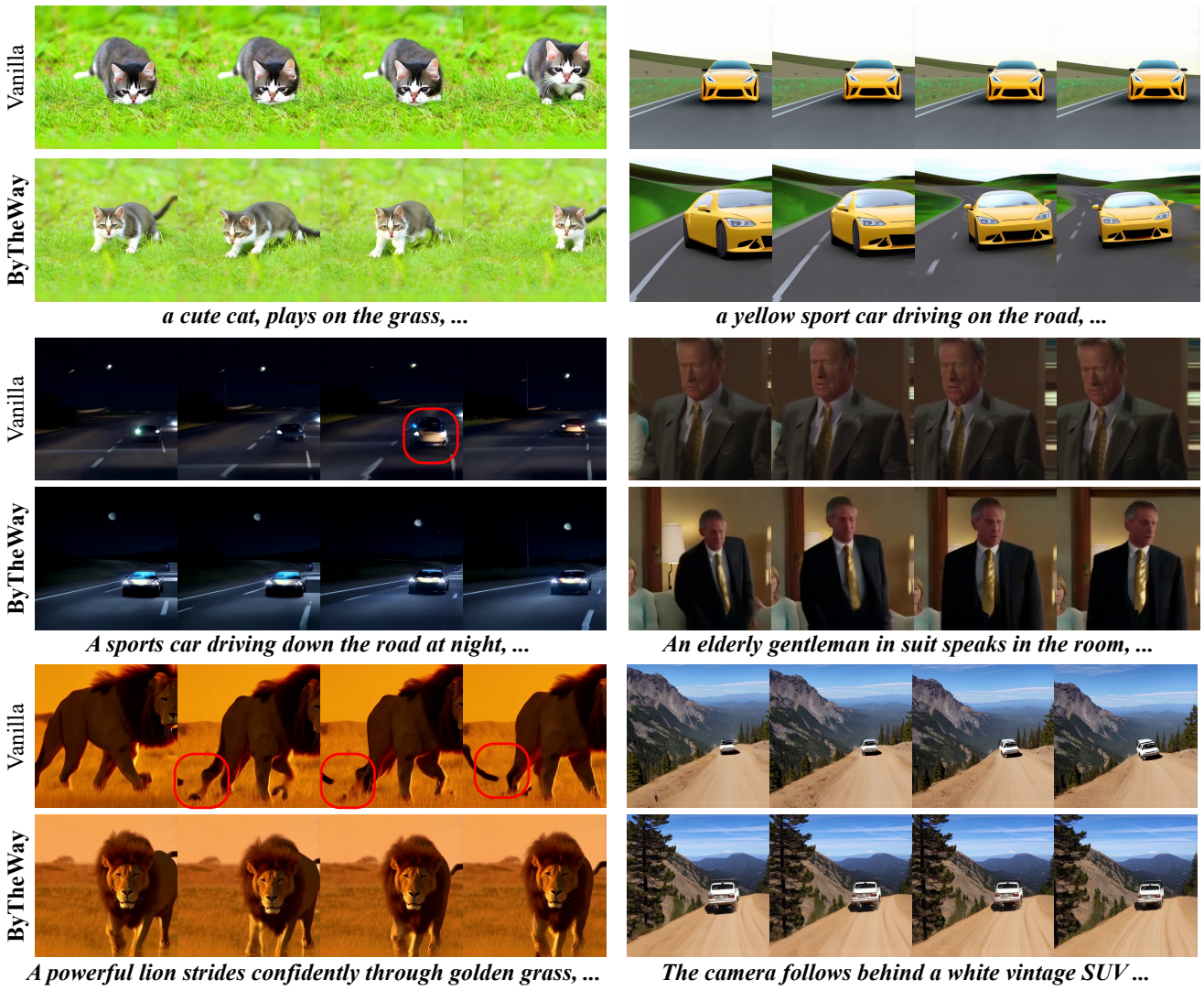

Figure 3. **Samples generated by CogVideoX-2B [7] with or without ByTheWay.**

[6] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *CVPR*, 2024. 1

[7] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1, 4

Figure 4. **Visual comparison with other training-free methods.**

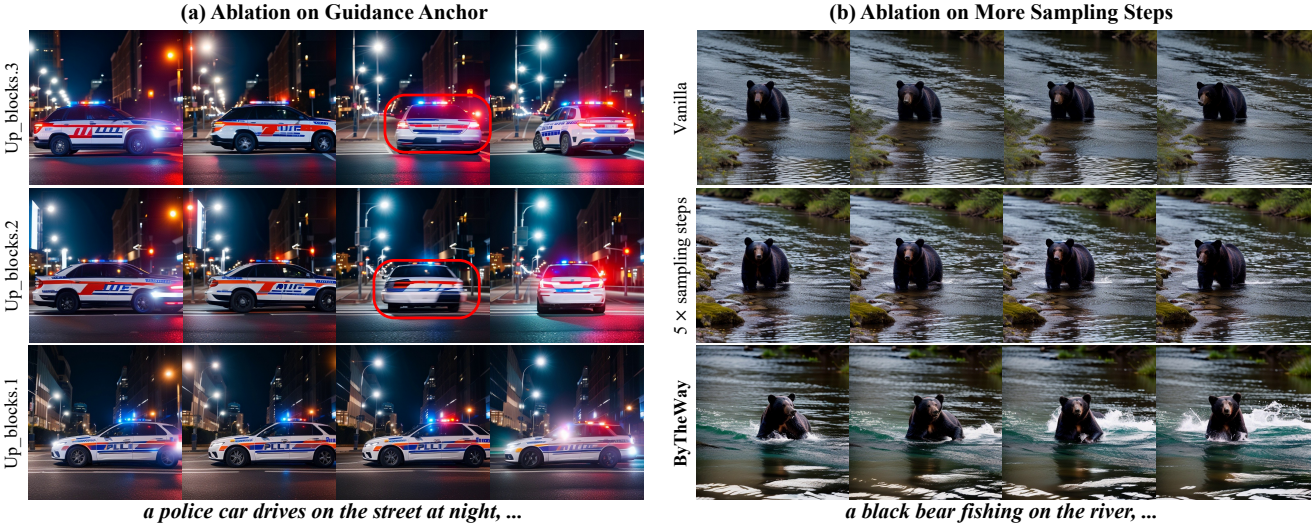

Figure 5. **Ablation on Guidance Anchor (a) and More Sampling Steps (b).**

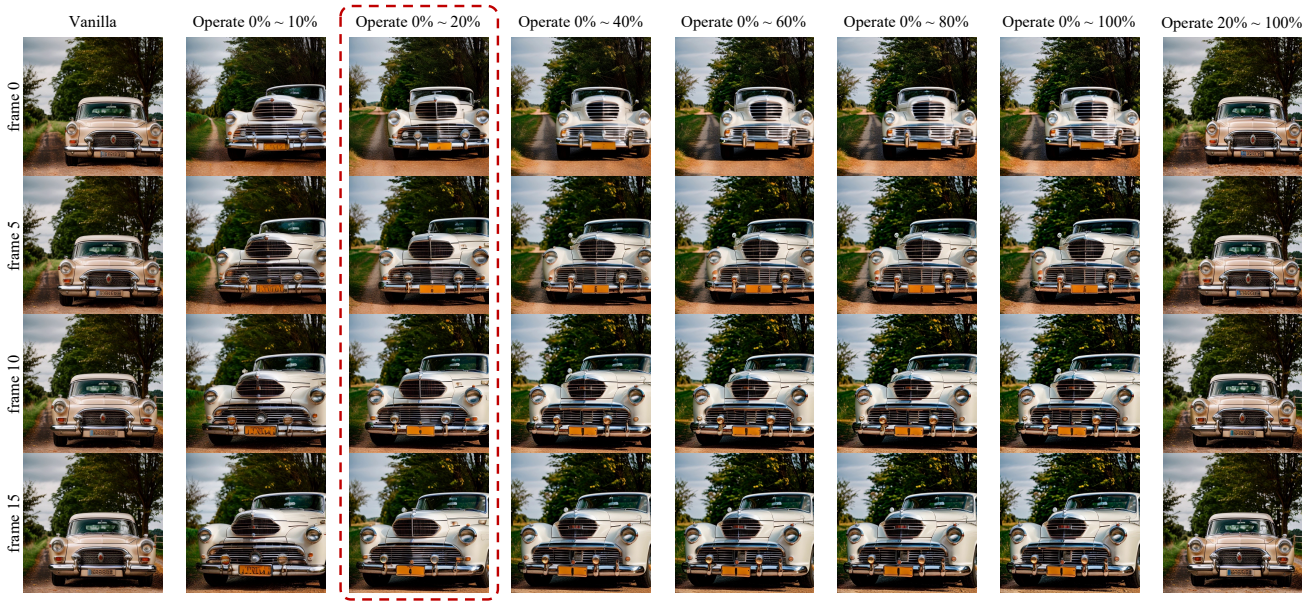

Figure 6. **Ablation on Operation Steps.** Prompt: *"a vintage car drives on a country road, ..."*

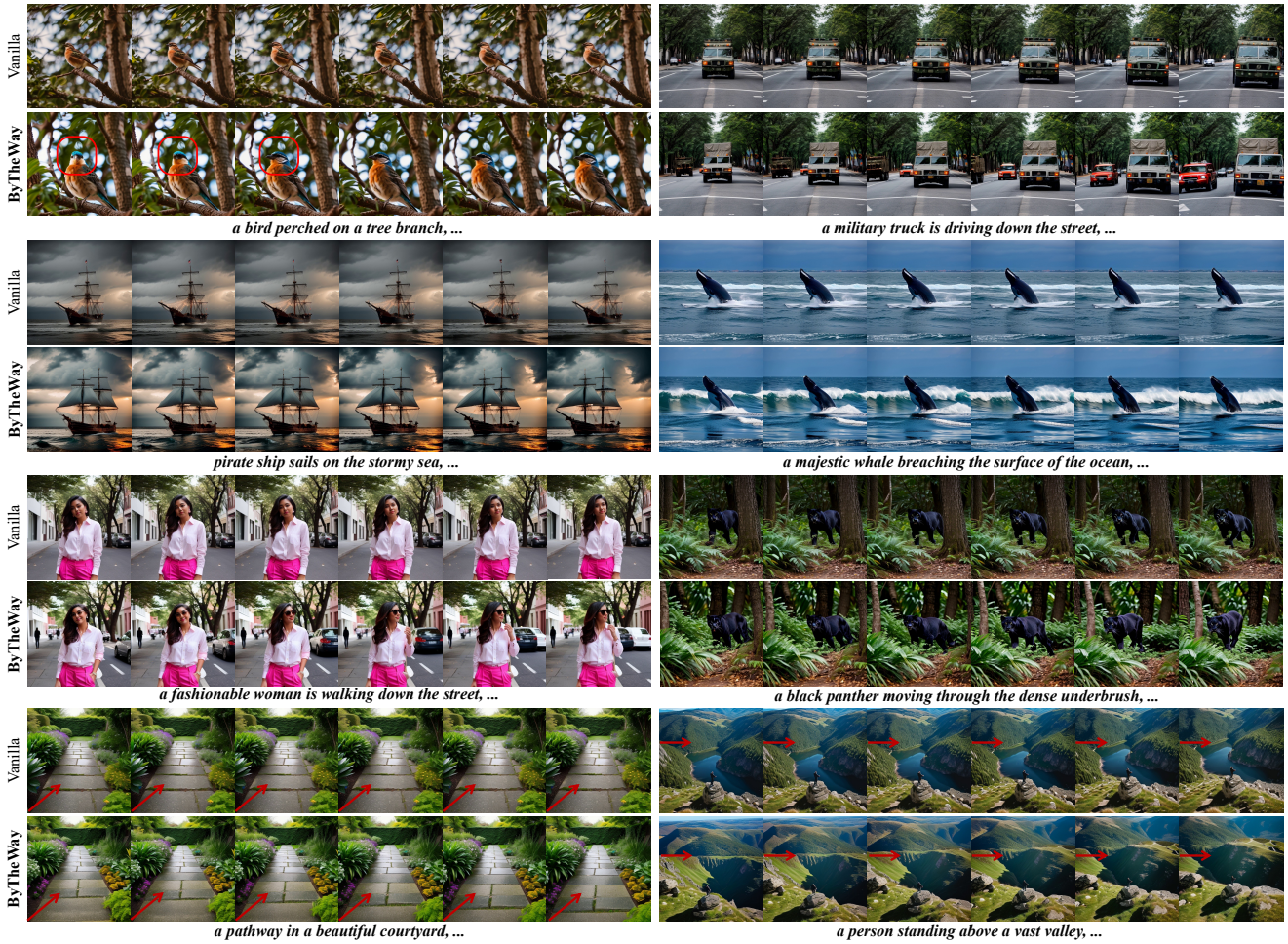

Figure 7. **More Results on AnimateDiff (Motion Enhancement).**

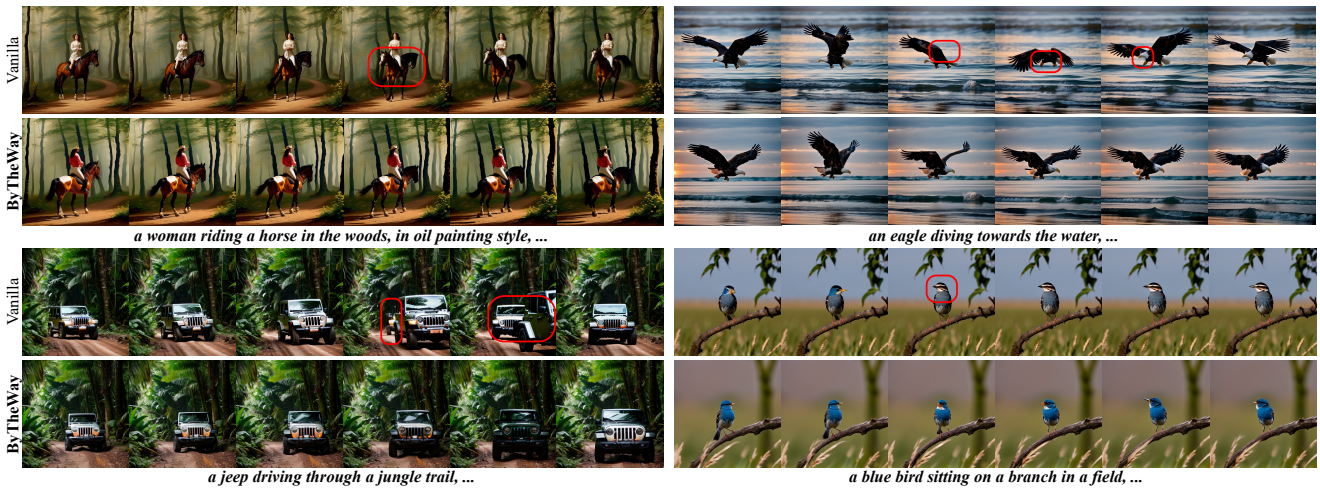

Figure 8. **More Results on AnimateDiff (Structure Enhancement).**

Figure 9. **More Results on VideoCrafter2 (Motion Enhancement).**

Figure 10. **More Results on VideoCrafter2 (Structure Enhancement).**