A. Implementation Details

Data Collection. We collect paired data across six representative tasks, including generation tasks (1) scribble map transfer, (2) Van Gogh style transfer and (3) camera movement transfer and perception tasks (4) depth map prediction, (5) semantic segmentation prediction, and (6) salient object tracking. For task (1)(4)(6), we collect the source clips (A, B) from the Señorita-2M dataset [6], and the annotated clips (A', B') are obtained using preprocessing tools from VACE [2] code repository[1]. For task (2)(3), we use the source videos from Señorita-2M dataset [6] and edit them using TokenFlow [4] and computer software CapCut[2], respectively. For task (5), we source data from VSPW dataset [3].
Fine-tuning Details. We employ Wan2.1-T2V-1.3B[3] as the backbone video generation model for our experiments. The model is trained for 20~40 epochs for each vision tasks, with each epoch consisting of 200 iterations. For co-training across all vision tasks and contexts, we train the model for 20 epochs, with 1,200 iterations per epoch. Training on two A800 GPUs, each epoch takes about 12 minutes for 200 iterations and roughly one hour for 1200 iterations.
Experimental Details. In fine-tuning with mixed vision contexts, we randomly choose the vision contexts of each training sample. For tasks (1)(2)(4)(5)(6), we sample vision contexts I and II each with probability $p = 0.3$, context III with $p = 0.4$. For task (3), since the transformation pertains to the temporal dimension, sampling is limited to contexts I and IV, each with probability $p = 0.5$. In the ablation study investigating the impact of text, we utilize prompts at multiple levels of granularity, including detailed, rough, and null texts. The prompt template is illustrated in Fig. 1. For all other experiments, we consistently use the detailed text prompt.

B. Comparison between LVM and video generation model

As summarized in Table 1, the training data required by LVM [1] is complex to construct, while the video generation model Wan is pretrained only on raw images and videos. Although Wan uses more total training tokens than LVM, it achieves higher visual quality by employing an 8× downsampling encoder, compared to LVM's 16× downsampling. Additionally, Wan has fewer parameters than the released version of LVM, resulting in lower computational costs.

C. Additional Experimental Results

We provide additional results related to four experiments presented in the main paper. Specifically, we show the results of each vision task across all contexts in Section C.2, the performance of each task under mixed-context fine-tuning in Section C.3, the impact of text prompts in Section C.5, and the results of co-training with all tasks and contexts in Section C.4.

Text Prompts

Detailed Text: "[clip1] is the original source video, and [clip2] is its corresponding segmentation map. [clip3] is another, different source video. In [clip4], the segmentation map transformation applied from [clip1] to [clip2] is similarly applied to [clip3]."

Rough Text: "[clip1] and [clip2] form an editing pair. Apply the same transformation observed from [clip1] to [clip2] to [clip3], and generate [clip4]."

Null Text: ""

Figure 1. **Text Prompts at multiple levels of granularity.**

	LVM	Wan
Dataset Composition	1. Single images 2. Image sequences 3. Images with annotations 4. Image sequences with annotations	Raw images and videos
Dataset Scale	420B tokens	$\mathcal{O}(1)$T tokens [5]
Downsample Ratio	$16X$	$8X$
Parameters	7B (released version)	1.3B

Table 1. **Comparison between LVM [1] and the video generation model Wan [5].**

C.1. Conditional Generation Tasks

As shown in Fig. 2, the video generation model is capable of performing conditional generation tasks based on depth maps or masked videos.

C.2. Performance Across Different Contexts

We present the performance of each vision task across different contexts in Fig. 3. The results demonstrate that the fine-tuned video generation model effectively handles not only image and video tasks, but also cross-modal and cross-data-source tasks.

[1] https://github.com/ali-vilab/VACE
[2] https://www.capcut.com/
[3] https://github.com/Wan-Video/Wan2.1

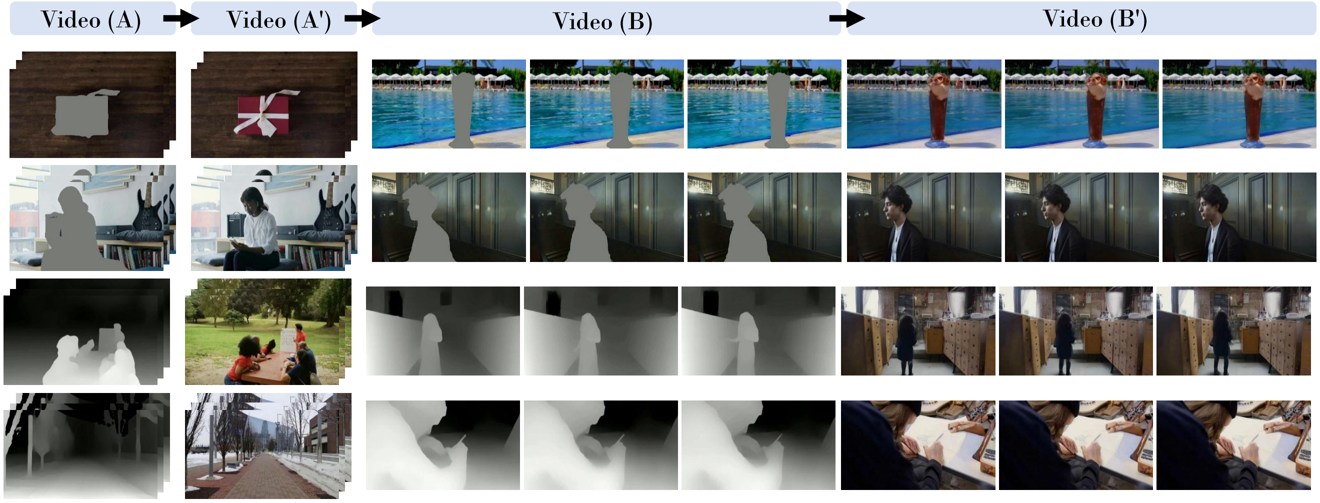


Figure 2. **Results of conditional generation tasks.**

C.3. Performance under Mixed-Context Fine-Tuning

As shown in Fig. 4, Fig. 5 and Fig. 6 , when fine-tuned on mixed vision contexts, the model can automatically adjust to each vision context, demonstrating strong generalization ability.

C.4. Co-training with All Vision Tasks and Contexts

As shown in Fig. 11, when co-trained on all vision tasks under mixed contexts, the model achieves consistent performance across tasks and contexts, demonstrating robust generalization ability with limited supervision.

C.5. Impact of Texts Across Different Vision Contexts

As shown in Fig. 7, Fig. 8, Fig. 9, and Fig. 10, the model effectively learns the relationships among the four clips across different vision contexts without explicit textual guidance, demonstrating strong in-context learning capabilities in the temporal dimension.

D. Broader Impacts

Our work offers a resource-efficient approach to building unified vision models, which benefit applications in low-resource settings by reducing the reliance on large annotated datasets. However, the broad adaptability of such models also raises potential risks, including misuse for generating deceptive content. Careful evaluation and responsible deployment are essential to mitigate the concern.

References

[1] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22861–22872, 2024.

[2] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025.

[3] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4133–4143, 2021.

[4] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. *arXiv preprint arXiv:2412.03069*, 2024.

[5] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[6] Bojia Zi, Penghui Ruan, Marco Chen, Xianbiao Qi, Shaozhe Hao, Shihao Zhao, Youze Huang, Bin Liang, Rong Xiao, and Kam-Fai Wong. Señorita-2m: A high-quality instruction-based dataset for general video editing by video specialists. *arXiv preprint arXiv:2502.06734*, 2025.

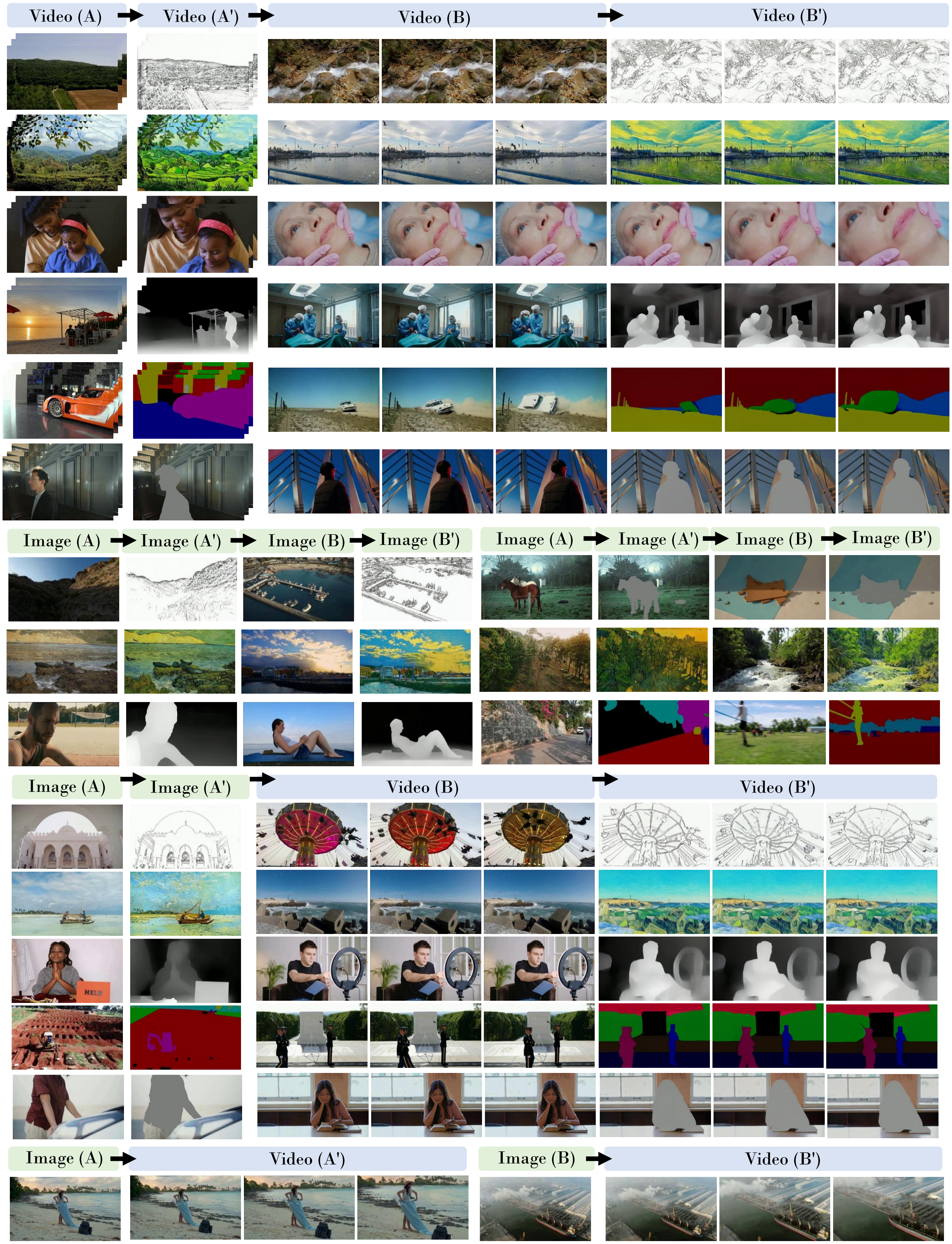
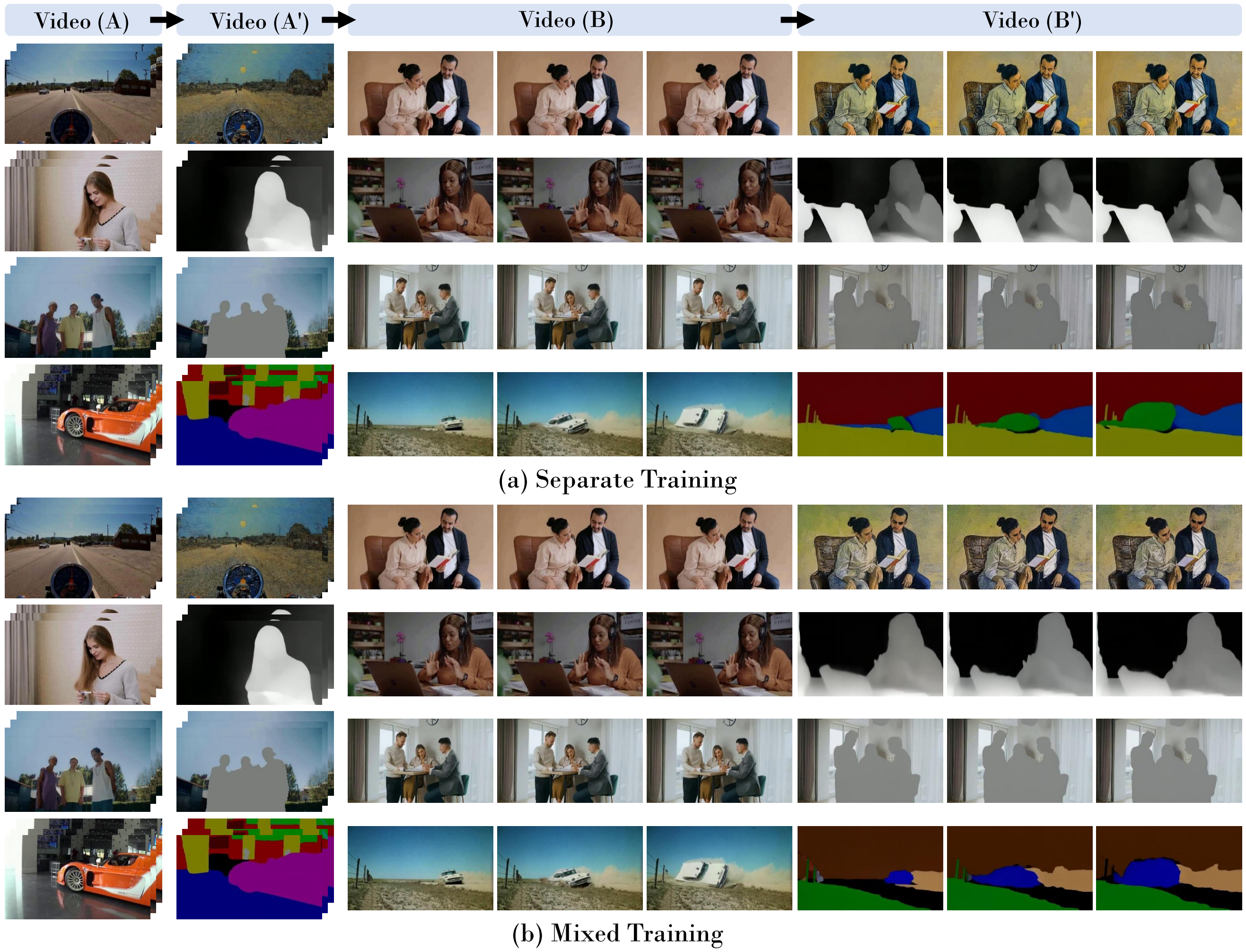


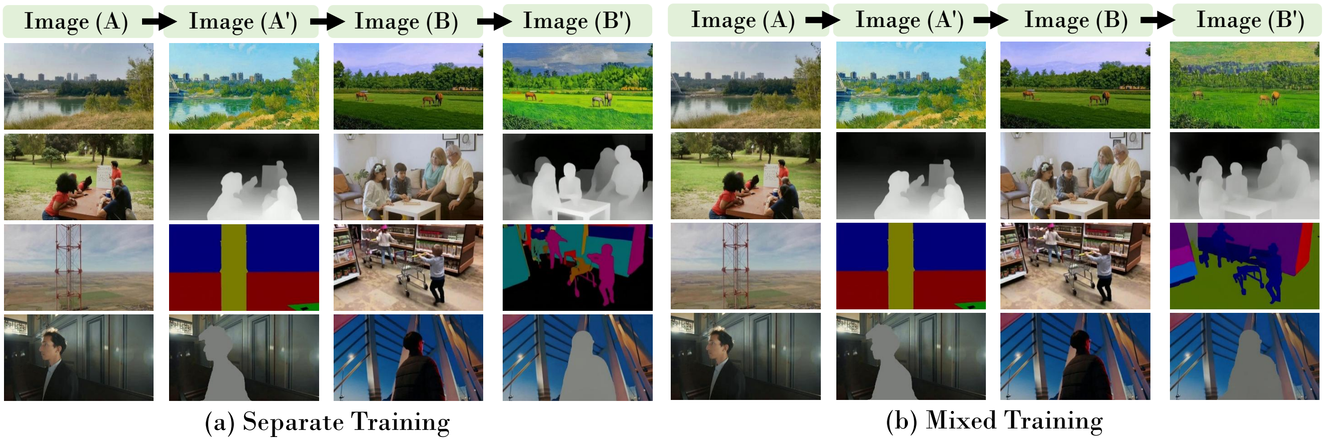
Figure 3. **Additional results of various vision tasks across context types.**



(a) Separate Training

(b) Mixed Training

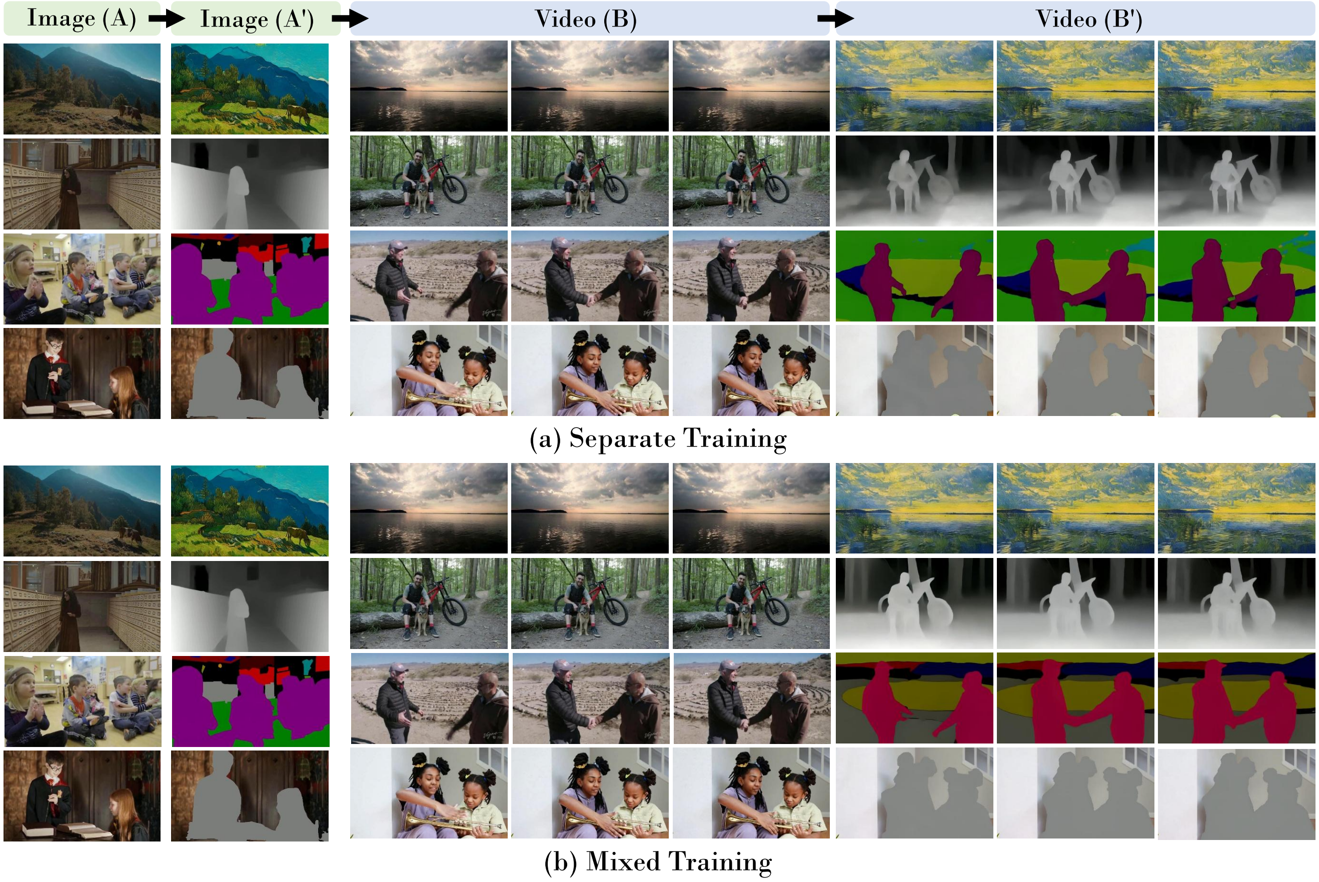
Figure 4. **Performance under separate and mixed training for context I.**



(a) Separate Training

(b) Mixed Training

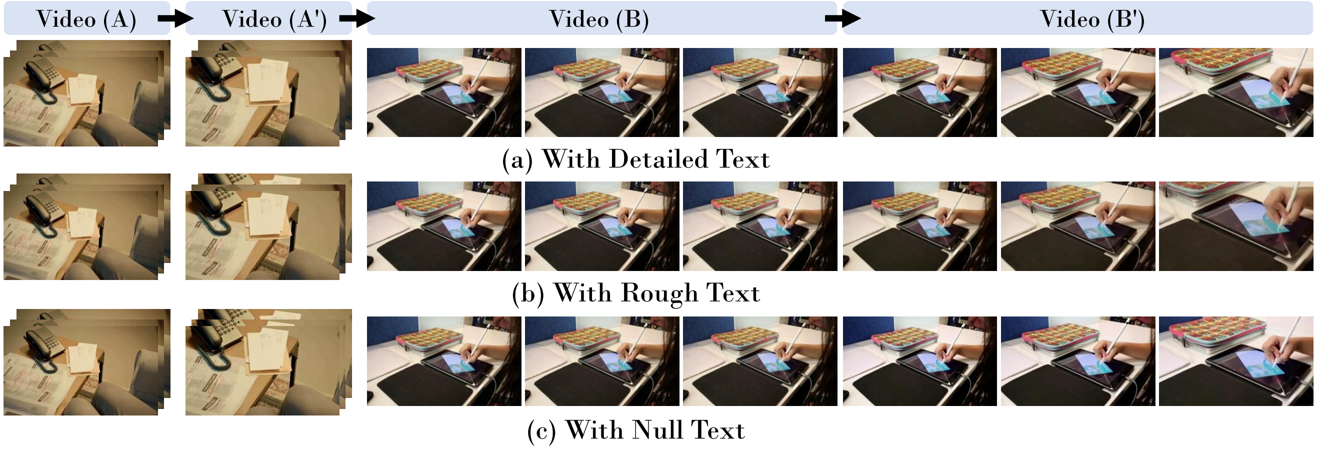
Figure 5. **Performance under separate and mixed training for context II.**



(a) Separate Training

(b) Mixed Training

Figure 6. **Performance under separate and mixed training for context III.**



(a) With Detailed Text

(b) With Rough Text

(c) With Null Text

Figure 7. **Impact of texts under contexts I.**

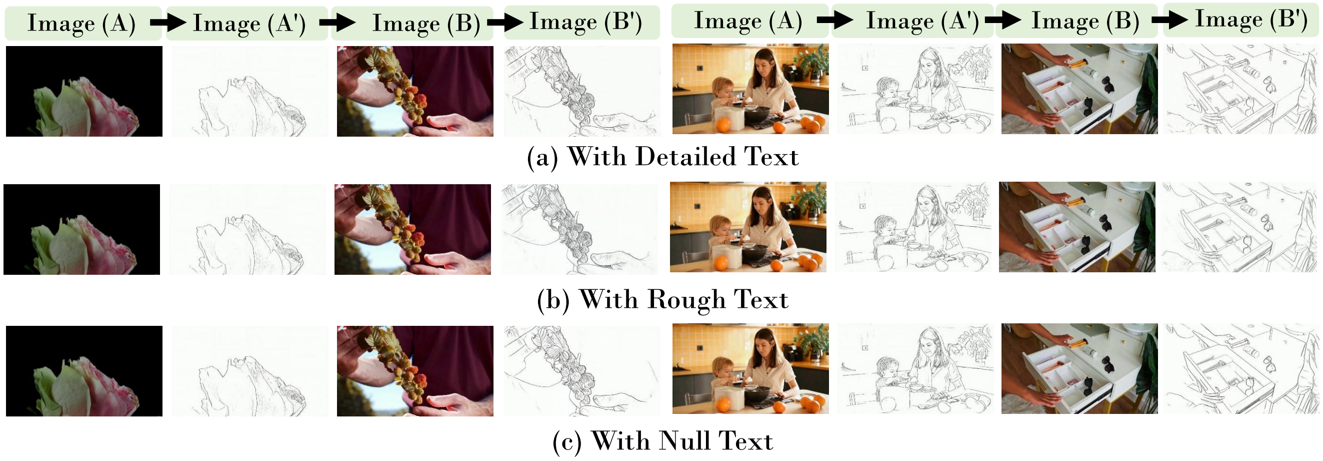


Figure 8. **Impact of texts under contexts II.**

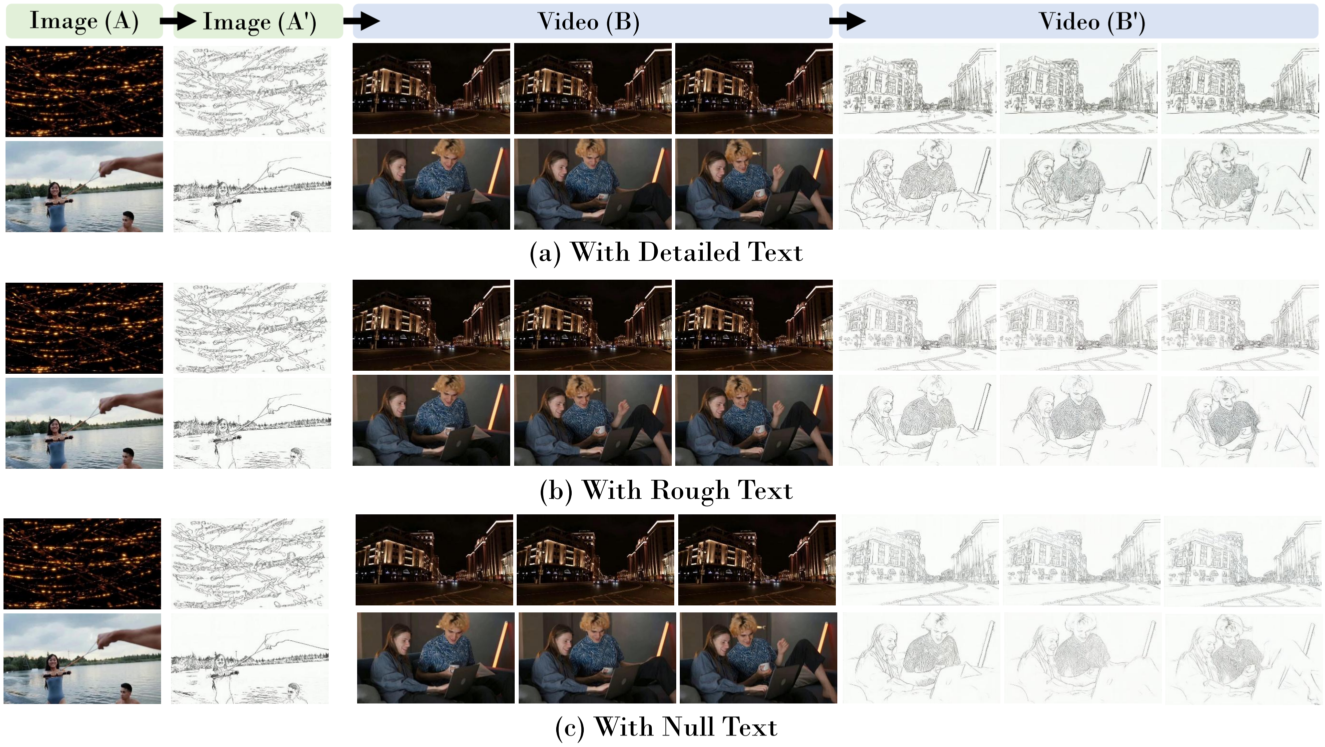


Figure 9. **Impact of texts under contexts III.**

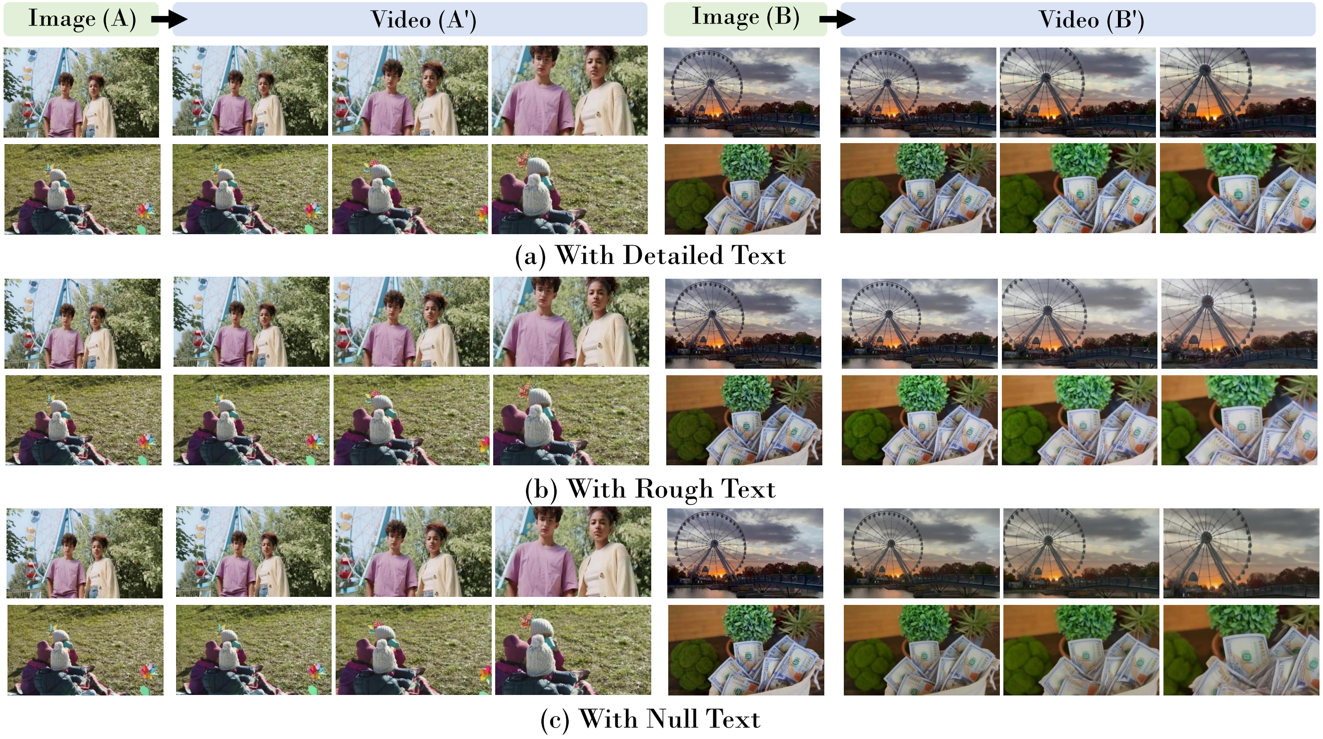


Figure 10. **Impact of texts under contexts IV.**

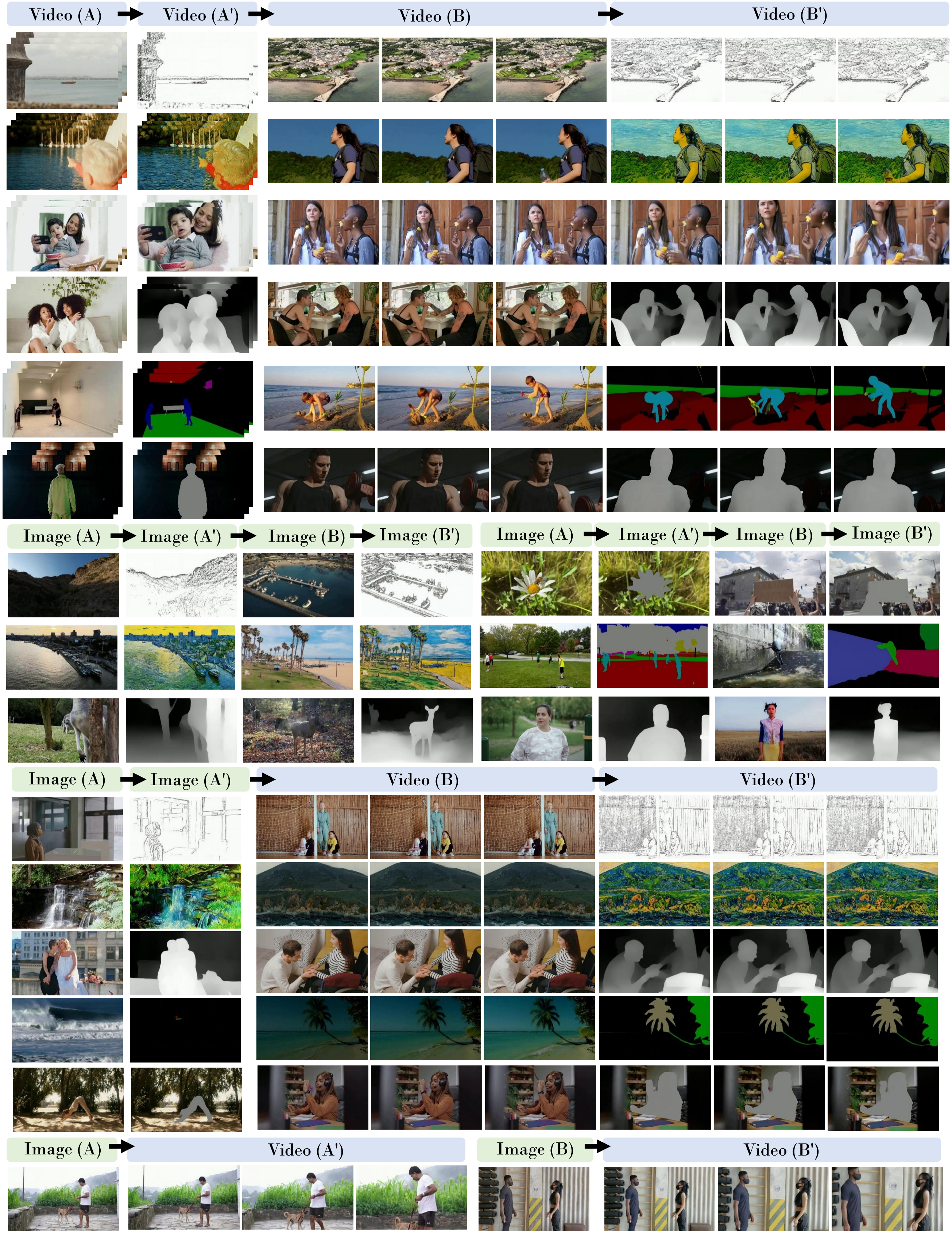


Figure 11. **Co-training with all vision tasks and contexts.**