Supplementary Document for *OmniVDiff*

Implementation Details

Modality embeddings The modality embeddings are two learnable vectors of dimension 16, matching the channel size of 3D VAE latent patches. They are directly added to the noisy latents before entering the transformer.

Training strategy We adopt a two-stage training strategy to enable a smooth transition from single-modality video generation to multi-modal controllable video synthesis. In the first stage, we fine-tune a pre-trained video generation model, CogVideoX, to generate multiple visual modalities (e.g., depth, segmentation, canny edges) from text conditions. This step helps the model learn a strong prior for multi-modal video synthesis, ensuring alignment between different modalities within a unified latent space. In the second stage, we introduce our adaptive modality control strategy, allowing the model to perform X-conditioned video generation and multi-modal video understanding. This stage enables the model to distinguish between generation and conditioning roles for different modalities, enhancing its flexibility across tasks.

Through this progressive learning approach, the model evolves from single rgb video generation to multi-modal controllable video generation, enabling seamless transition between various video generation and understanding tasks.

Model architecture Given input modalities including rgb, depth, segmentation, and canny, we first encode each modality independently into latent representations using a pretrained 3D Encoder (Yang et al. 2024b). These latent features are then concatenated along the channel dimension. To accommodate this multi-modal input, we modify the patch embedding linear layer accordingly, while initializing the extended portions of the weights with zeros. The outputs from the Transformer are then passed through *modality-specific projection heads* (MSPH), each consisting of an AdaLayerNorm followed by a Linear layer. The parameters of these heads are initialized by copying those from the rgb branch. Finally, each modality-specific representation is decoded back into color space using a pretrained 3D Causal Decoder (Yang et al. 2024b). Compared to using a single projection head—which expands the output dimension of the linear layer to match the number of modalities, modality-specific projection heads allow the model to better capture the unique characteristics of each modality.

Synthetic data annotation To supplement our training data with high-quality ground-truth labels, we construct a small synthetic dataset designed for depth and segmentation supervision. The synthetic dataset is composed of two sources: one rendered from the 3D-FRONT dataset (Fu et al. 2021), and the other from the ABO dataset (Collins et al. 2022). For rendering, we randomly select from a set of eight predefined camera trajectories. For each rendering sequence, a random initial position is chosen within a scene. We render 49 frames per sequence using the Cycles ray-tracing renderer provided by BlenderProc2 (Denninger et al. 2023), with 128 samples per pixel (spp) for each frame. In total, we generate 10,000 synthetic video clips.

Evaluation Details

VBench evaluation protocol Following the evaluation protocol in Aether (Team et al. 2025), we adopt VBench as the evaluation metric for our video generation tasks. Considering the differences in input settings between *OmniVDiff* and other baselines, we evaluate the generated videos using a customized configuration of VBench across six key dimensions:

1. **Subject Consistency:** Evaluates the temporal consistency of the main subjects.
2. **Background Consistency:** Measures the stability and coherence of the scene background.
3. **Motion Smoothness:** Assesses the naturalness and fluidity of motion.
4. **Dynamic Degree:** Quantifies the level of motion and activity within the video.
5. **Aesthetic Quality:** Evaluates the visual appeal and artistic merit of the generated content.
6. **Imaging Quality:** Measures the technical fidelity and visual clarity of the video output.

The final VBench score is computed as a weighted average of these dimensions using the official weights:

- Subject Consistency: 1.0
- Background Consistency: 1.0

- Motion Smoothness: 1.0
- Dynamic Degree: 0.5
- Aesthetic Quality: 1.0
- Imaging Quality: 1.0

For each evaluation metric, we randomly selected 2,048 samples from the validation set for testing.

Depth evaluation protocol Following the evaluation protocol in (Chen et al. 2025), we assess the geometric accuracy of our short video depth predictions. Consistent with the approach in (Hu et al. 2024), we first apply a global scale and shift alignment between the predicted depth maps and the ground truth across each video. To quantify geometric accuracy, we compute the Absolute Relative Error (AbsRel) and δ_1 metrics (Hu et al. 2024; Yang et al. 2024a). All evaluations are conducted on the ScanNet dataset (Dai et al. 2017). Following the (Chen et al. 2025), we set the maximum video length to 90 frames for short video evaluation.

Segment evaluation protocol Our segmentation process begins by estimating the first-frame mask using SemanticSAM (Li et al. 2023a), which is then propagated across the video using SAM2 (Ravi et al. 2024) for consistent object tracking. For the initial segmentation, we set the granularity level to 2. Consequently, the quality of the first-frame mask is critical for overall segmentation accuracy. Given this data generation process, we focus on comparison experiments based on the first frame. Following the SemanticSAM protocol, we adopt the Single-Granularity Interactive Segmentation setting throughout all evaluations.

To process the segmentation results, we follow the procedure proposed in DICEPTION (Zhao et al. 2025), where K-Means clustering is applied to generate class-specific masks. For each predicted mask, we compute the Intersection over Union (IoU) with the corresponding ground-truth mask from the COCO 2017 Val dataset (Lin et al. 2015).

Additional Experiments

Expert modality vs. estimated modality in X-conditioned video generation Since our model is designed for omni-controllable video generation, it is essential to evaluate how well it can perform controllable generation using its own estimated modalities instead of relying on external expert inputs. We conduct a comprehensive evaluation to assess the differences between expert-provided modalities and our estimated modalities in controllable video generation. Specifically, for reference videos, we use expert models to generate corresponding depth and segmentation maps as pseudo labels. In contrast, our method uses the reference video as input to perform video understanding, estimating depth, segmentation, and canny edges simultaneously within a single diffusion process. These estimated modalities are then used as conditioning inputs for controllable video generation, allowing us to compare generation quality between expert-provided labels and our model's self-predicted pseudo labels.

Table 1 presents a quantitative comparison between video generation results conditioned on these two sources. Across all modalities, the performance of *ours-est* closely matches that of the expert-conditioned inputs, demonstrating the accuracy and robustness of our modality estimation pipeline. These findings highlight the capability of *OmniVDiff* to perform both video understanding and controllable video generation without relying on any expert models, thereby enabling a fully unified Omni Video Diffusion framework.

Qualitative comparison of canny-conditioned video generation As shown in Figure 1, we present qualitative comparisons of video generation under Canny-conditioned guidance. Compared to existing approaches, including VideoX-Fun (aigc-apps 2024), ControlVideo (Zhang et al. 2023), Control-A-Video (Chen et al. 2023), and CogVideoX+ControlNet (TheDenk 2024), our method produces the most visually faithful results, demonstrating superior alignment with the provided edge-based guidance.

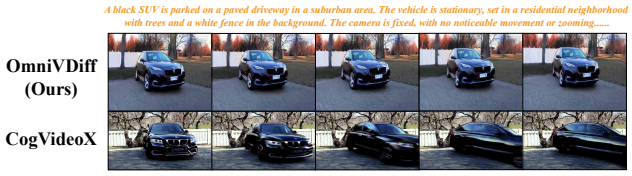


Figure 2: **Qualitative Comparison with CogVideoX.** Our method preserves better semantic content and structural alignment described in the prompt. (e.g., "A black SUV is parked on a paved driveway in a suburban area").

Qualitative comparison of text-conditioned video generation with CogVideoX Figure 2 presents a qualitative comparison between our method and CogVideoX. Note that for this evaluation, we focus on the rgb modality to ensure a fair comparison, as CogVideoX does not support multi-modal outputs. *OmniVDiff* generates videos that more faithfully follow the prompt (e.g., maintaining a static camera), while also delivering superior visual quality. Although our approach is designed for multi-modal training, the joint optimization may offer stronger regularization than using rgb alone, potentially resulting in more coherent and consistent predictions.

Segmentation evaluation In the main paper, we provide quantitative comparisons with baseline methods. Here, we further present qualitative results and analyze the impact of training with synthetic data for segmentation understanding. As shown in Table 3 and Figure 3, *OmniVDiff* achieves performance comparable to expert models but still shows inconsistencies in background segmentation, largely due to noise in the pseudo labels generated by SemanticSAM. For example, in Figure 3(b), SemanticSAM mistakenly segments the background into multiple regions.

To assess the impact of synthetic data on segmentation estimation, we additionally compare with the OmniVDiff-Syn setting. With the inclusion of 10k synthetic samples, *OmniVDiff*-Syn effectively mitigates background inconsistencies, as shown in Figure 3(d). This demonstrates the model's ability to leverage small amounts of high-quality data for noticeable improvements in segmentation performance.

Model	subject consistency	b.g. consistency	motion smoothness	dynamic degree	aesthetic quality	imaging quality	weighted average
text+depth							
OmniVDiff(VideoDepthAnything)	**97.96**	**96.66**	**99.18**	**53.32**	**52.95**	**67.26**	**73.45**
OmniVDiff(ours-est)	96.92	95.96	99.17	52.48	51.59	66.97	72.80
text+canny							
OmniVDiff(openCV)	**97.84**	**95.55**	**99.23**	**53.53**	**52.34**	**67.14**	**73.14**
OmniVDiff(ours-est)	96.98	95.41	99.19	52.92	51.39	67.01	72.74
text+segment							
OmniVDiff(SemanticSAM& SAM2)	**97.97**	**95.81**	**99.31**	**53.18**	**53.37**	**67.51**	**73.42**
OmniVDiff(ours-est)	96.96	95.70	99.23	52.70	51.76	66.91	72.82

Table 1: **VBench metrics comparing video generation conditioned on expert-provided modalities and our estimated modalities.** *Ours* denotes using modalities generated by expert models as conditions, while *ours-est* uses modalities estimated by our model for conditioning.

Method	COCO Val 2017(Lin et al. 2015)	
	Point (Max) 1-IoU ↑	Point (Oracle) 1-IoU ↑
Semantic-SAM (T)(Li et al. 2023b)	54.5	73.8
Semantic-SAM (L)(e)(Li et al. 2023b)	**57.0**	**74.2**
OmniVDiff(ours)	<u>56.0</u>	<u>73.9</u>
OmniVDiff-Syn(ours)	<u>56.9</u>	**74.4**

Table 2: Comparison with prior methods on point-based interactions, evaluated on COCO Val2017. "Max" selects the prediction with the highest confidence score, while "Oracle" uses the one with highest IoU against the target mask.

Method	AbsRel ↓	δ_1 ↑
Aether(Team et al. 2025)	0.117	0.869
VDA-S (e)(Chen et al. 2025)	<u>0.110</u>	<u>0.876</u>
OmniVDiff-Syn(Ours)	**0.100**	**0.894**

Table 3: **Zero-shot video depth estimation results.** We compare our method with representative single-image and video depth estimation models. "VDA-S(e)" denotes the expert model with a ViT-Small backbone. The **best** and <u>second-best</u> results are highlighted.

Depth evaluation Recently, a concurrent work Aether (Team et al. 2025) proposes a unified framework for world model modeling, supporting both video understanding and joint multi-modal generation across rgb, depth, and camera pose. Although its primary focus lies in geometric world modeling, it shares the same video understanding task on depth modality. Therefore, we provide comparisons with it for a comprehensive evaluation. Given Aether is trained on high-quality synthetic data with GT depth supervision, we compare it to the setting *OmniVDiff*-Syn, trained on the Koala dataset augmented with 10k synthetic samples. As shown in Table 3, our model achieves superior results. This improvement may be attributed to the combination of large-scale real data and synthetic samples used in our training. Our model leverages large scale real data to inherit robust depth generation capabilities from Video Depth Anything (Chen et al. 2025), while synthetic data further refines its depth estimation accuracy. Figure 4 presents a visual comparison, where *OmniVDiff*-Syn demonstrates better generalization and geometric fidelity.

More visual results Figure 5 illustrates qualitative results of our model under two different generation settings: (a) shows video synthesis conditioned solely on text, while (b) demonstrates segmentation-conditioned generation. Figure 6 presents results under depth- and canny-guided generation: (a) corresponds to depth-conditioned outputs, and (b) to canny-conditioned ones.Finally, Figure **??** highlights the video understanding capabilities of *OmniVDiff*, showcasing its accurate prediction of auxiliary modalities from rgb inputs.

Applications

Video-to-video style control *OmniVDiff* can be seamlessly applied to the downstream task of video-to-video style control. In Figure 7, we show more diverse scene styles, such as winter, autumn, and sunset, while faithfully maintaining the original scene structure.

3D scene reconstruction Additionally, since *OmniVDiff* estimates the corresponding depth, which captures the 3D geometric structure of the scene, we can directly reproject the depth video into a 3D space to reconstruct the scene. This reconstructed 3D world can then be rendered over time from novel viewpoints, enabling dynamic scene visualization from different perspectives, as illustrated in Figure 8.

Adaptability to new modalities and tasks To evaluate the adaptability of our model to previously unseen modalities and applications, we conduct experiments on two representative tasks: video deblurring and video super-resolution.

We fine-tune *OmniVDiff* for 2k steps using 5k training samples, with a learning rate of 2e-5. During fine-tuning, we repurpose the slot originally used for the Canny modality to accommodate a new input—either a blurred or low-resolution rgb video—depending on the task. At inference time, the model treats the modified input as a conditioning signal and generates high-quality rgb outputs, either deblurred or super-resolved.

As shown in Fig. 9, our model successfully adapts to both tasks, achieving visually pleasing results. This experiment highlights the strong flexibility of our unified framework, demonstrating that it can be extended to new modalities and tasks with minimal architectural or data adjustments.

Limitations and Future Work

Despite its effectiveness, *OmniVDiff* has several limitations. First, due to the scarcity of high-quality real-world data with paired modalities, we adopt a strategy that uses pseudo labels generated by expert models. However, these pseudo labels can introduce artifacts—especially in segmentation, where results may lack consistency, as shown in Figure 10. In this work, we take an initial step by incorporating a small set of synthetic data to improve performance. Scaling up the synthetic dataset in the future may further enhance model robustness and accuracy. Second, in this work, we primarily focus on demonstrating the effectiveness and practicality of our approach for video generation using four commonly used modalities: rgb, depth, segmentation, and canny edges. Extending *OmniVDiff* to support a broader range of modalities is another promising direction, which we leave for future work. Finally, our framework can benefit from leveraging more powerful pretrained backbones. Exploring the integration of recent large-scale video foundation models, such as WAN (Wan et al. 2025) and Hunyuan (Kong et al. 2024) Video, may help further improve the model performance.

References

aigc-apps. 2024. VideoX-Fun: A Video Generation Pipeline for AI Images and Videos. https://github.com/aigc-apps/VideoX-Fun. GitHub repository, accessed 2025-07-21.

Chen, S.; Guo, H.; Zhu, S.; Zhang, F.; Huang, Z.; Feng, J.; and Kang, B. 2025. Video Depth Anything: Consistent Depth Estimation for Super-Long Videos. *arXiv:2501.12375*.

Chen, W.; Ji, Y.; Wu, J.; Wu, H.; Xie, P.; Li, J.; Xia, X.; Xiao, X.; and Lin, L. 2023. Control-A-Video: Controllable Text-to-Video Diffusion Models with Motion Prior and Reward Feedback Learning. *arXiv preprint arXiv:2305.13840*.

Collins, J.; Goel, S.; Deng, K.; Luthra, A.; Xu, L.; Gundogdu, E.; Zhang, X.; Vicente, T. F. Y.; Dideriksen, T.; Arora, H.; Guillaumin, M.; and Malik, J. 2022. ABO: Dataset and Benchmarks for Real-World 3D Object Understanding. arXiv:2110.06199.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. arXiv:1702.04405.

Denninger, M.; Winkelbauer, D.; Sundermeyer, M.; Boerdijk, W.; Knauer, M.; Strobl, K. H.; Humt, M.; and Triebel, R. 2023. BlenderProc2: A Procedural Pipeline for Photorealistic Rendering. *Journal of Open Source Software*, 8(82): 4901.

Fu, H.; Cai, B.; Gao, L.; Zhang, L.-X.; Wang, J.; Li, C.; Zeng, Q.; Sun, C.; Jia, R.; Zhao, B.; et al. 2021. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10933–10942.

Hu, W.; Gao, X.; Li, X.; Zhao, S.; Cun, X.; Zhang, Y.; Quan, L.; and Shan, Y. 2024. DepthCrafter: Generating Consistent Long Depth Sequences for Open-world Videos. arXiv:2409.02095.

Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; Dai, Z.; Zhou, J.; Xiong, J.; Li, X.; Wu, B.; Zhang, J.; et al. 2024. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*.

Li, F.; Zhang, H.; Sun, P.; Zou, X.; Liu, S.; Yang, J.; Li, C.; Zhang, L.; and Gao, J. 2023a. Semantic-SAM: Segment and Recognize Anything at Any Granularity. *arXiv preprint arXiv:2307.04767*.

Li, F.; Zhang, H.; Sun, P.; Zou, X.; Liu, S.; Yang, J.; Li, C.; Zhang, L.; and Gao, J. 2023b. Semantic-SAM: Segment and Recognize Anything at Any Granularity. *arXiv preprint arXiv:2307.04767*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C. L.; and Dollár, P. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Team, A.; Zhu, H.; Wang, Y.; Zhou, J.; Chang, W.; Zhou, Y.; Li, Z.; Chen, J.; Shen, C.; Pang, J.; and He, T. 2025. Aether: Geometric-Aware Unified World Modeling. arXiv:2503.18945.

TheDenk. 2024. cogvideox-controlnet: ControlNet Extensions for CogVideoX. https://github.com/TheDenk/cogvideox-controlnet. GitHub repository, commit `<YOUR-COMMIT-HASH>`, accessed 2025-07-21.

Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; Zeng, J.; Wang, J.; Zhang, J.; Zhou, J.; Wang, J.; Chen, J.; Zhu, K.; Zhao, K.; Yan, K.; Huang, L.; Feng, M.; Zhang, N.; Li, P.; Wu, P.; Chu, R.; Feng, R.; Zhang, S.; Sun, S.; Fang, T.; Wang, T.; Gui, T.; Weng, T.; Shen, T.; Lin, W.; Wang, W.; Wang, W.; Zhou, W.; Wang, W.; Shen, W.; Yu, W.; Shi, X.; Huang, X.; Xu, X.; Kou, Y.; Lv, Y.; Li, Y.; Liu, Y.; Wang, Y.; Zhang, Y.; Huang, Y.; Li, Y.; Wu, Y.; Liu, Y.; Pan, Y.; Zheng, Y.; Hong, Y.; Shi, Y.; Feng, Y.; Jiang, Z.; Han, Z.; Wu, Z.-F.; and Liu, Z. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*.

Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024a. Depth Anything V2. arXiv:2406.09414.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024b. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*.

Zhang, Y.; Wei, Y.; Jiang, D.; Zhang, X.; Zuo, W.; and Tian, Q. 2023. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*.

Zhao, C.; Liu, M.; Zheng, H.; Zhu, M.; Zhao, Z.; Chen, H.; He, T.; and Shen, C. 2025. DICEPTION: A Generalist Diffusion Model for Visual Perceptual Tasks. arXiv:2502.17157.

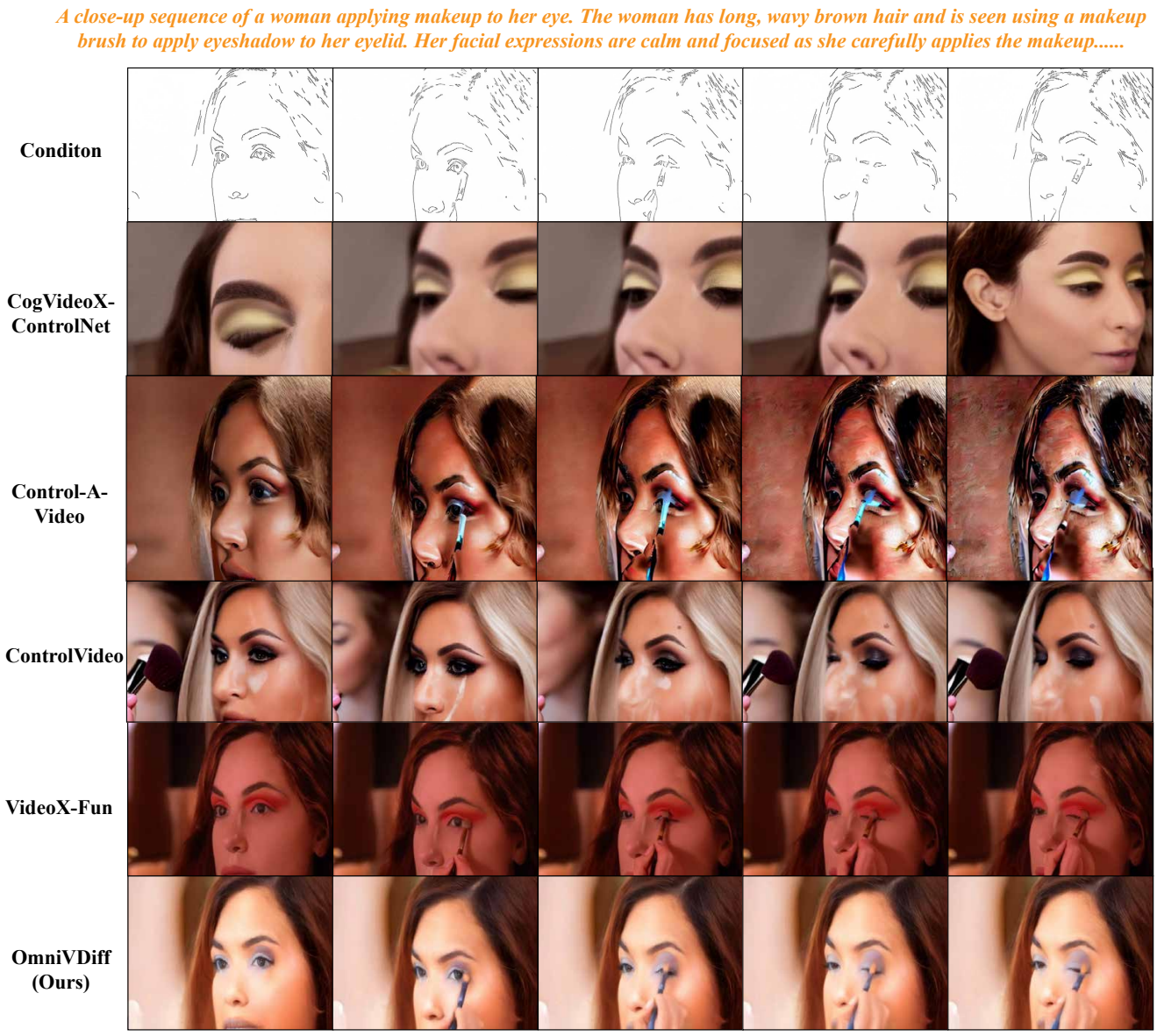


Figure 1: **Visual comparison for canny-guided video generation.**

Figure 3: **Qualitative comparison of video segment estimation.** Yellow arrows highlight frame-wise background segmentation inconsistencies.

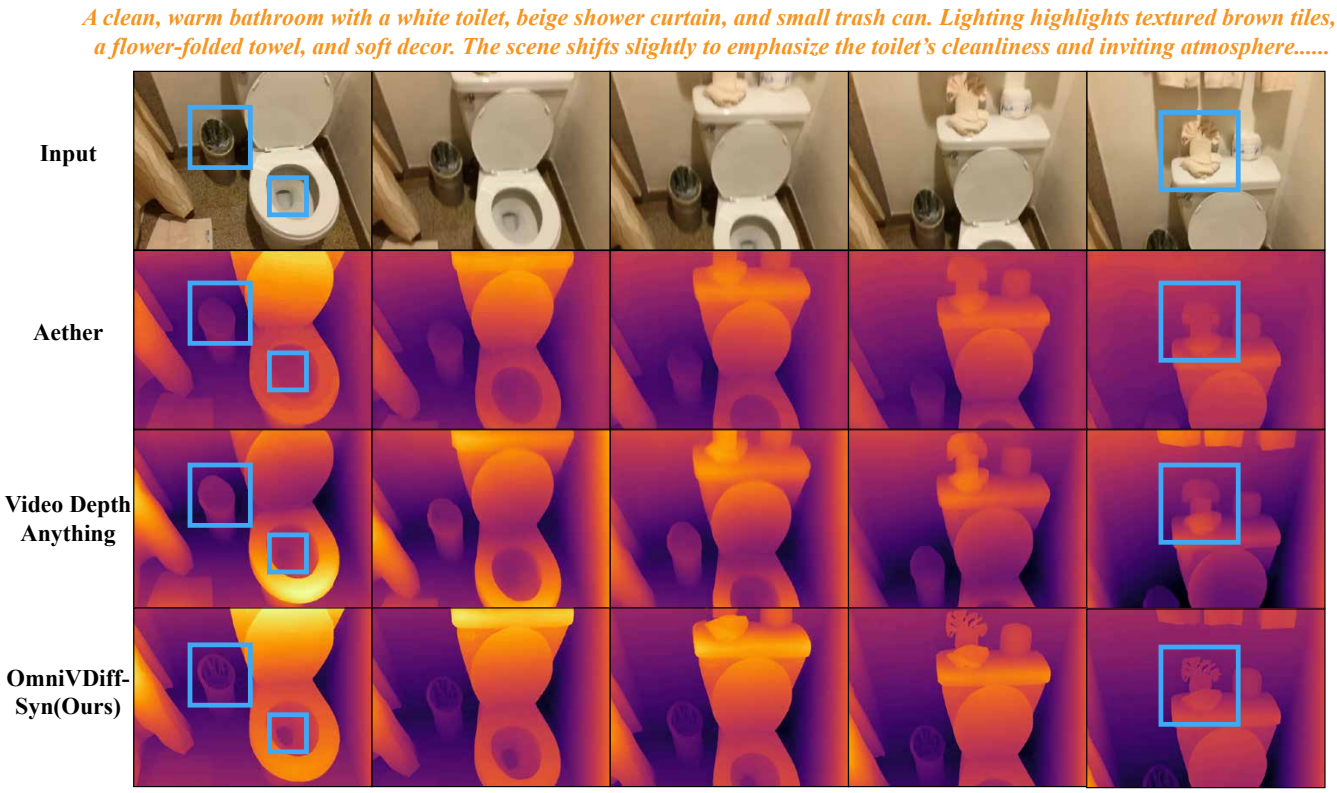


Figure 4: **Qualitative comparison of video depth estimation.** Blue boxes indicate *OmniVDiff*-Syn successfully captures sharper structural details and superior geometric fidelity.

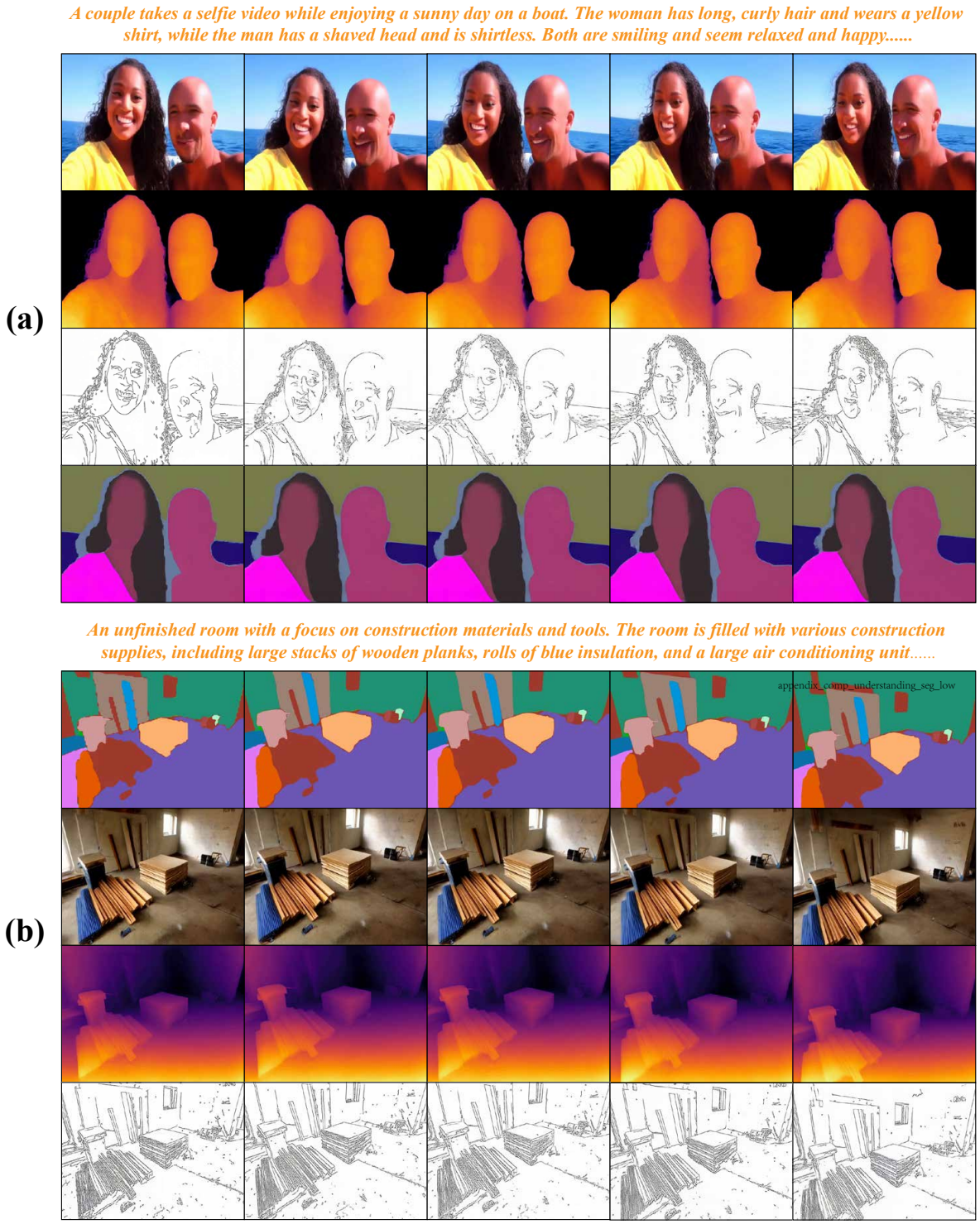
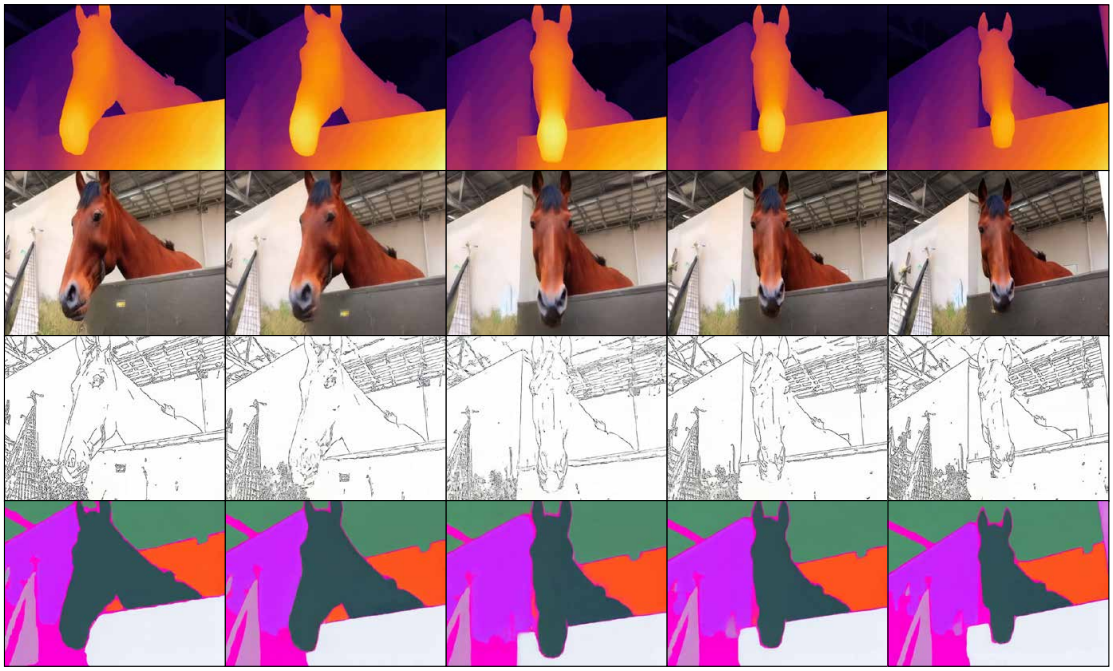


Figure 5: **Qualitative results of text- and segmentation-conditioned video generation.** (a) shows generation results conditioned only on text; (b) presents results conditioned on both segmentation masks and text, demonstrating improved spatial alignment and controllability.

(a)

A close-up view of a brown horse standing in a stable. The horse is calmly looking directly at the camera, with its head slightly tilted to the side. The stable is well-lit, and the horse's coat appears shiny and well-groomed.....



(b)

A young woman with long, light brown hair stands outdoors facing the camera. She wears a white t-shirt under a red and white plaid shirt and appears to speak or react, as her facial expressions subtly change throughout the video.....

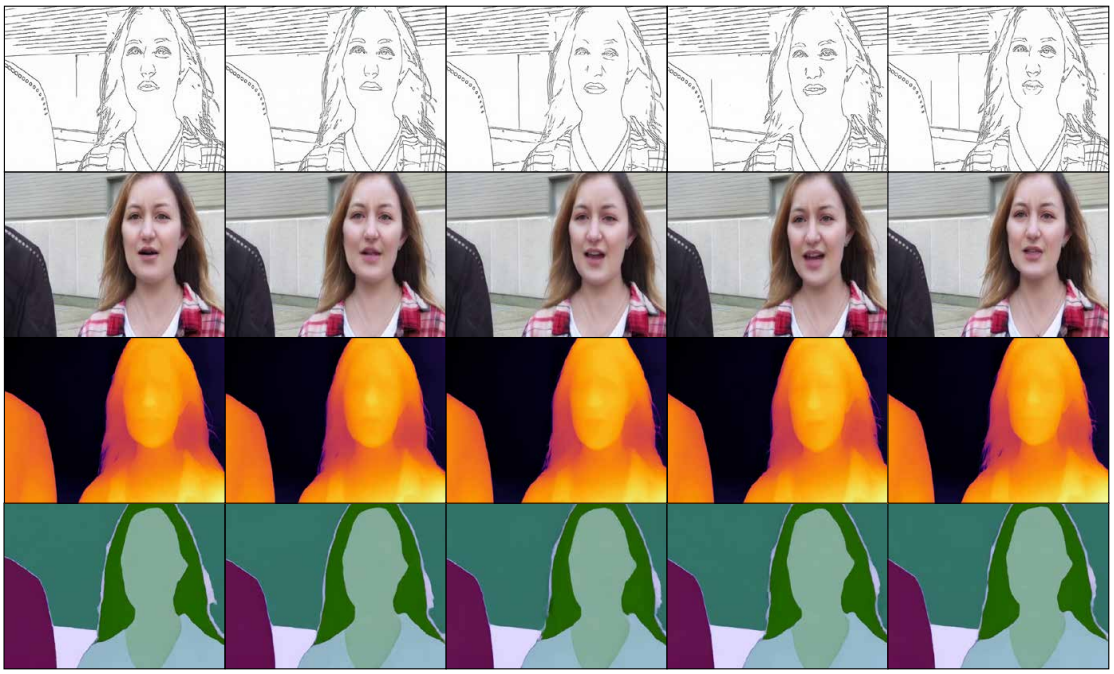


Figure 6: **Qualitative results of depth- and canny-conditioned video generation.** (a) shows video generation results conditioned on depth maps; (b) presents results conditioned on canny maps.

Figure 7: **Video2Video Style Control.** Given a reference video (a), *OmniVDiff* first estimates the corresponding depth (b) and uses it as a bridge to control the scene structure, enabling the generation of videos with diverse scene styles through text-based control (c,d,e).

Figure 8: **Scene reconstruction.** Given a reference video (a), our model estimates the corresponding depth (b). The depth video can be reprojected into the 3D world and rendered from novel viewpoints.

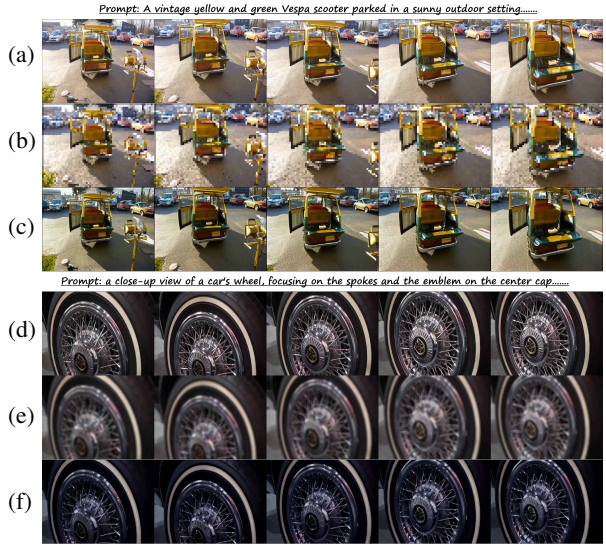


Figure 9: **Adaptation to new applications**. (1) Video super-resolution: (a) a reference video; (b) a low-resolution video from (a) as model input; (c) *OmniVDiff* generates a corresponding high-resolution video. (2) Video deblurring: Similarly, *OmniVDiff* can be fine-tuned for the video deblurring task, producing a sharp video (f) from a blurred input (e). The reference video is shown in (d).

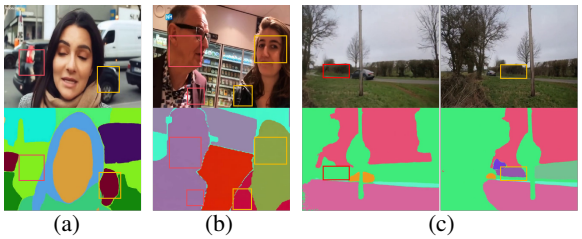


Figure 10: **Issues in training data (segmentation)**. (a) **Inconsistency**: the vehicle region is incorrectly segmented into multiple classes due to occlusion by a person; (b) **Ambiguity**: the segmentation granularity is not consistent; (c) **Flickering**: the segmentation labels of the same object (e.g., a tree) vary drastically across frames, leading to temporal inconsistency.