

UnityVideo: Unified Multi-Modal Multi-Task Learning for Enhancing World-Aware Video Generation

Appendix

The appendix contains the following sections:

- [More Analysis of Model Design](#)
- [More Experiments and Analysis](#)
- [Details of OpenUni and UniBench](#)
- [More Visuals and Applications](#)

A. More Analysis of Model Design

A.1. Modal Interaction Analysis

To further investigate the cross-modal interactions within our unified framework, we visualize the evolution of self-attention maps throughout the training process. We partition the attention map into four distinct regions based on modality interactions: self-modality regions comprising (RGB, RGB) and (optical flow, optical flow), and cross-modality regions consisting of (RGB, optical flow) and (optical flow, RGB), where Flow represents various auxiliary modality features. As illustrated in Figure 1, our analysis reveals three key findings. First, as joint training progresses, the interaction between RGB and auxiliary modalities becomes progressively more pronounced (A), indicating deepening cross-modal feature exchange. Second, the visualization results demonstrate that the model learns increasingly rich geometric representations with improved text-following capabilities (B), validating the effectiveness of our unified training paradigm in enhancing both visual understanding and conditional generation quality. This empirical evidence confirms that our unified framework not only enables technical integration of multiple modalities but also facilitates meaningful feature-level interactions that contribute to improved world modeling capabilities.

A.2. Modality-Specific Output Layers

While our modality switcher and in-context learner effectively differentiate between modalities, we observed occasional modality confusion as the number of modalities scales. For instance, when instructed to generate segmentation masks, the model infrequently produces skeleton outputs instead. This confusion stems from all modalities sharing a common output layer, which can conflate distinct modality-specific features at the final projection stage.

To address this limitation, we introduce modality-specific output layers (adaptive layer) while maintaining a unified input layer (share layer) for cross-modal information sharing. Each modality receives its own dedicated output projection layer, initialized independently, while the

Table 1. Comparison of different layer strategies.

	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Averaged
Baseline	96.51	96.06	98.73	99.30	97.650
Share Layer	98.31	97.54	99.35	99.54	98.685
Adaptive Layer	98.26	97.49	99.44	99.61	98.700

Table 2. Comparison with standalone T2V. Joint generation achieves better performance, with unified modality showing further improvements.

	Subject Consistency	Background Consistency	Imaging Quality	Overall Consistency	Averaged
Baseline	96.51	96.06	64.99	23.17	70.1825
T2V	96.51	97.23	66.52	23.44	70.9250
<i>Depth Modality</i>					
JointGen (Depth)	98.13 (+1.62)	97.29 (+0.06)	69.09 (+2.57)	23.48 (+0.04)	71.998 (+1.073)
JointGen (Unified)	98.01 (+1.50)	97.24 (+0.01)	69.18 (+2.66)	23.75 (+0.31)	72.045 (+1.120)
<i>Optical Flow Modality</i>					
JointGen (Optical Flow)	97.82 (+1.31)	97.14 (-0.09)	67.34 (+0.82)	23.70 (+0.26)	71.500 (+0.575)
JointGen (Unified)	97.97 (+1.46)	97.19 (-0.04)	69.36 (+2.84)	23.74 (+0.30)	72.065 (+1.140)
<i>Densepose Modality</i>					
JointGen (Densepose)	98.08 (+1.57)	97.38 (+0.15)	67.05 (+0.53)	23.49 (+0.05)	71.500 (+0.575)
JointGen (Unified)	98.03 (+1.52)	97.30 (+0.07)	70.20 (+3.68)	23.53 (+0.09)	72.265 (+1.340)

input processing remains shared to preserve inter-modal knowledge transfer. This architectural refinement ensures clear modality boundaries during generation without sacrificing the benefits of unified representation learning.

As shown in Table 1, this lightweight design effectively eliminates modality confusion during scaled training while maintaining comparable performance across metrics. The modality-specific output layers provide improved flexibility and achieve balanced performance across diverse evaluation criteria, validating this architectural choice for scalable multi-modal generation.

B. More Experiments and Analysis

B.1. Compare with T2V

While results in main paper demonstrates promising gains from joint generation over the baseline, we further investigate whether joint generation provides advantages over standard supervised fine-tuning (SFT) for text-to-video generation. We conduct extensive ablation studies across different modalities, training models with identical data and steps to ensure fair comparison of their text-to-video capabilities.

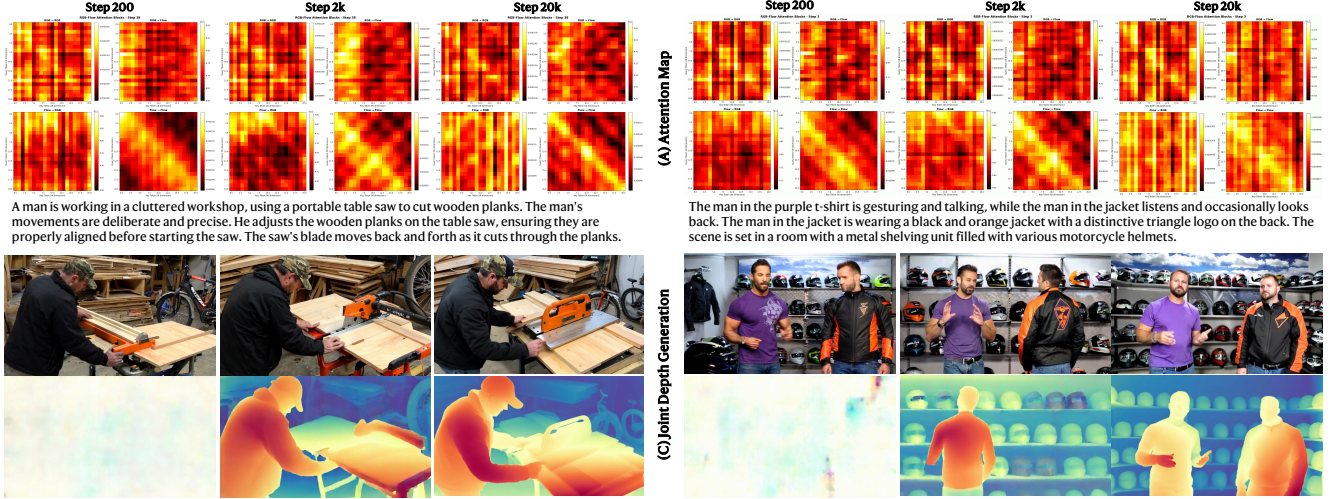


Figure 1. **Evolution of attention patterns in UnityVideo.** Analysis of attention maps shows that interactions between RGB and auxiliary modalities strengthen progressively across layers. Meanwhile, the model’s text-following behavior and spatial reasoning capabilities also improve, reflecting more coherent cross-modal integration.

As shown in Table 2, all modality configurations with joint generation achieve significant improvements over both the baseline and T2V-only training. Each auxiliary modality contributes distinct supervisory signals that enhance the model’s visual understanding, confirming the complementary nature of different modalities. Moreover, unified multi-modal training outperforms single-modality joint training by achieving better balance across evaluation dimensions, with substantial gains in overall performance (Averaged column). These results validate that diverse modality supervision collectively strengthens video generation through mutual reinforcement rather than simply additive improvements.

B.2. Scalability with Increasing Modalities

To demonstrate UnityVideo’s ability to continuously improve with expanded modality training, we evaluate performance scaling on both joint generation and controllable generation tasks. As shown in Table 3, UnityVideo achieves consistent performance gains across all metrics as the number of modalities increases. Specifically, we compare models trained with three modalities (depth, optical flow, and DensePose) against those trained with five modalities (additionally incorporating skeleton and segmentation).

The results reveal monotonic improvements across all evaluation criteria, confirming that our framework effectively leverages additional modality supervision without suffering from negative interference. This strong scalability suggests that UnityVideo’s architecture can accommodate further expansion in both model parameters and modality diversity, potentially enabling emergent world perception capabilities as the framework scales. The consistent gains validate our unified training paradigm as a promis-

Table 3. Analysis of the benefits brought by extended modal training for joint generation and control generation.

	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness
Baseline	96.51	96.06	98.73	99.30
<i>Joint Generation</i>				
Depth	96.53	95.58	98.45	99.28
Three Modalities	98.01	97.24	99.10	99.44
Five Modalities	98.31	97.54	99.35	99.54
<i>Control Generation</i>				
Depth	97.78	96.79	98.80	99.30
Three Modalities	97.83	96.86	98.87	99.33
Five Modalities	97.87	97.32	99.57	99.39

ing foundation for developing increasingly comprehensive video world models through continued modality integration.

B.3. The influence of different modalities

As shown in main paper, incorporating additional modalities yields further improvements for the *JointGeneration* task compared with training on a single modality. To examine whether this benefit also extends to *ControlGeneration*, we conduct the ablation study summarized in Table 4. Here, *Only* denotes models trained on ControlGeneration using a single modality, while *Ours* refers to models trained jointly with three modalities. All training data and iteration budgets are kept strictly identical to ensure a fair comparison.

The results show that unified multimodal training consistently outperforms single-modality training on the ControlGeneration task. These findings demonstrate that *UnityVideo* effectively strengthens positive cross-modal interactions across tasks, enabling each modality to benefit from the shared training paradigm.”

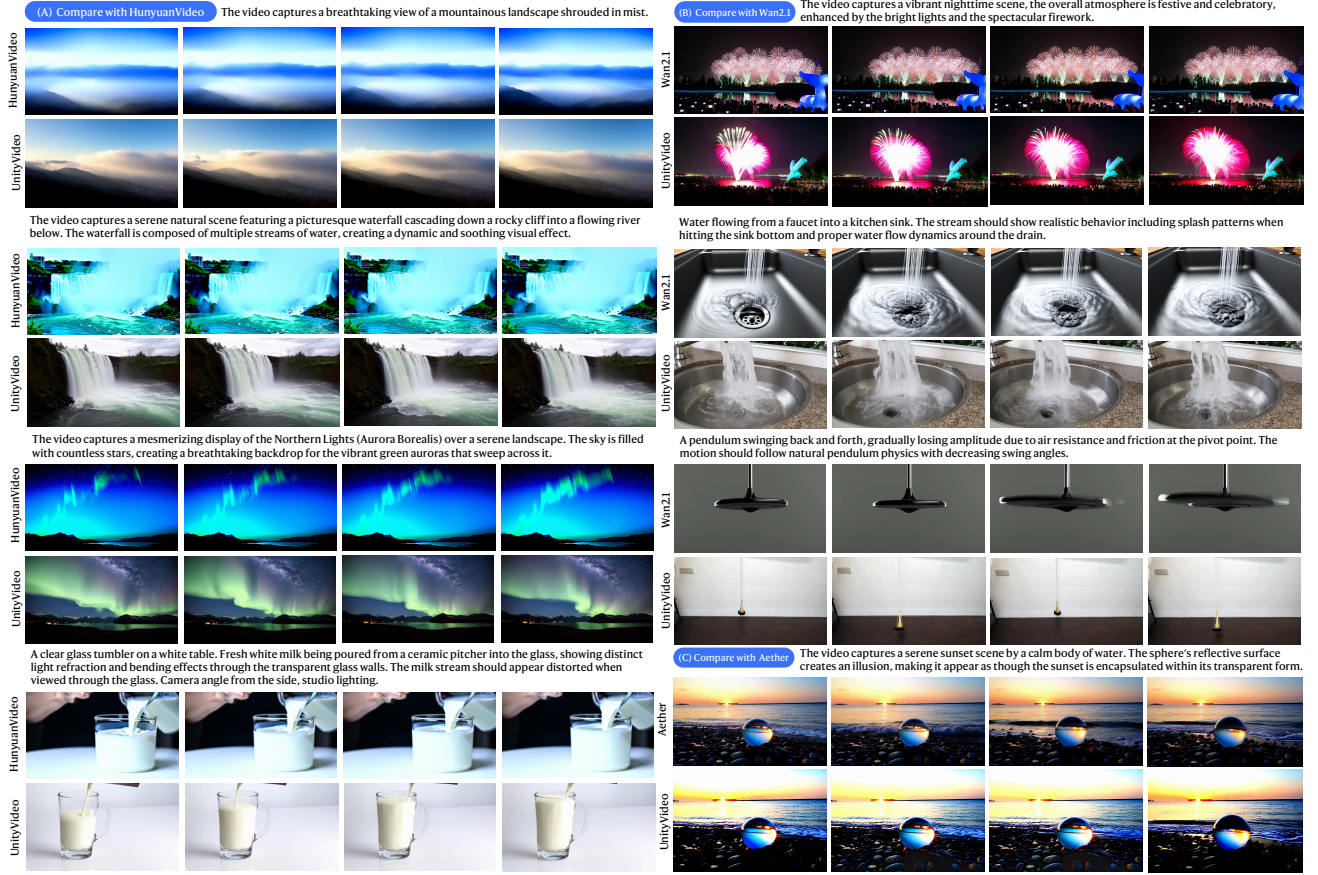


Figure 2. **Comparison of physical understanding.** UnityVideo demonstrates stronger physical reasoning and improved text alignment compared with current state-of-the-art video generation models.

Table 4. The gain of joint modal training compared with single modal on ControlGeneration tasks.

	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Averaged
Baseline	96.51	96.06	98.73	99.30	97.65
<i>Depth Modality</i>					
ControlGen (Depth)	97.78 (+1.27)	96.79 (+0.73)	98.80 (+0.07)	99.30 (+0.00)	98.1675 (+0.5175)
Unified (Depth)	97.83 (+1.32)	96.86 (+0.80)	98.87 (+0.14)	99.33 (+0.03)	98.2225 (+0.5725)
<i>Optical Flow Modality</i>					
ControlGen (Optical Flow)	97.40 (+0.89)	96.59 (+0.53)	98.67 (-0.06)	99.23 (-0.07)	97.9725 (+0.3225)
ControlGen (Unified)	97.47 (+0.96)	96.72 (+0.66)	98.83 (+0.10)	99.32 (+0.02)	98.0850 (+0.4350)
<i>Densepose Modality</i>					
ControlGen (Densepose)	97.01 (+0.50)	96.47 (+0.41)	98.58 (-0.15)	99.10 (+0.20)	97.790 (+0.5050)
ControlGen (Unified)	97.58 (+1.07)	96.79 (+0.73)	98.90 (+0.17)	99.35 (+0.05)	98.1550 (+0.5050)

B.4. World perception comparison

To further assess our model’s world understanding capabilities, we conduct comprehensive evaluations using physics-focused prompts that test fundamental physical principles. As shown in Figure 2, we evaluate models on scenarios in-

volving refraction, collision dynamics, and other physical phenomena that require accurate world modeling.

Our results demonstrate that UnityVideo exhibits superior understanding of physical laws compared to baseline methods. The model accurately captures light refraction through transparent media, realistic collision responses between objects, and physically plausible motion trajectories. These improvements stem from the complementary supervision provided by auxiliary modalities—depth enhances spatial reasoning, optical flow captures motion dynamics, and segmentation clarifies object boundaries—collectively enabling more accurate physical world modeling. This enhanced physical reasoning capability further validates the effectiveness of our unified multimodal training paradigm in developing world-aware video generation models.

C. Details of OpenUni and UniBench

C.1. OpenUni

The OpenUni dataset leverages diverse data sources and comprehensive modality extraction to create a large-scale multimodal training corpus. We employ multiple pretrained

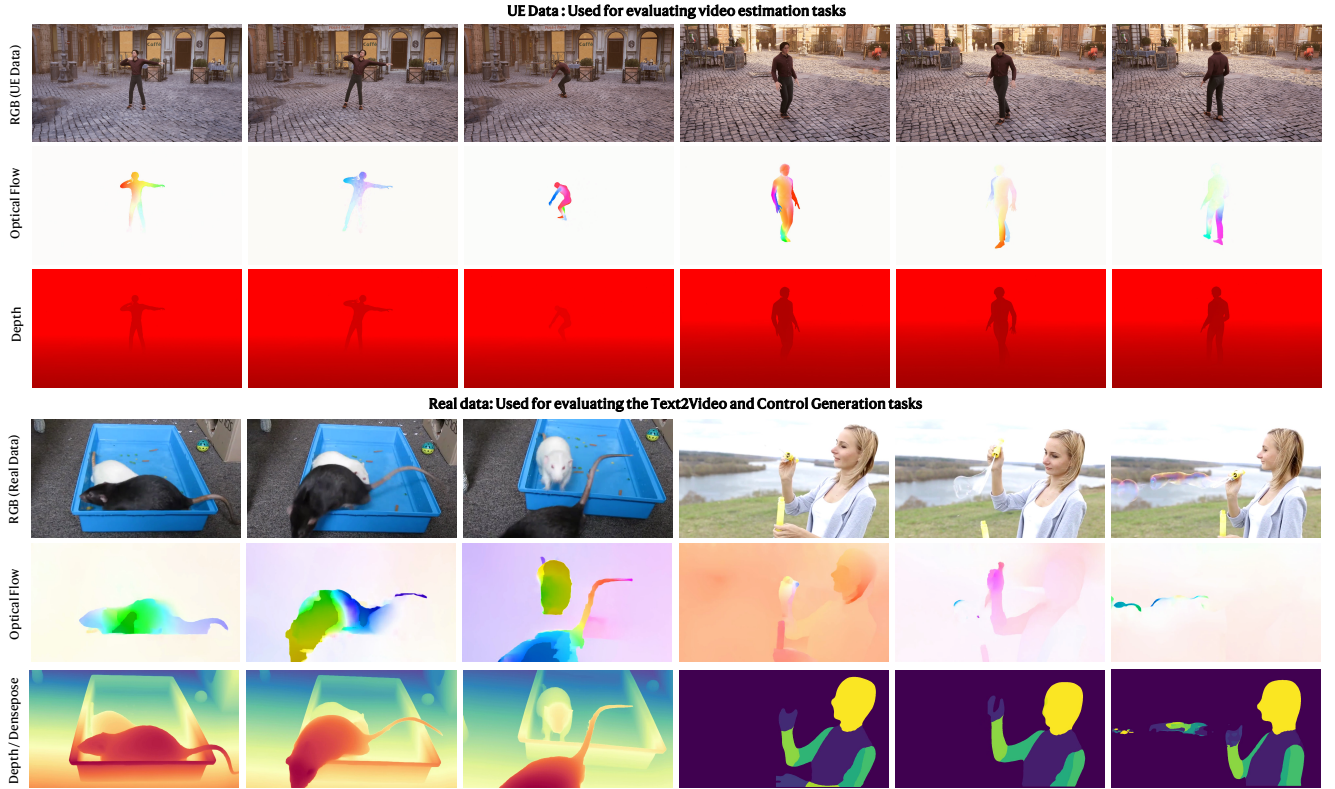


Figure 3. UniBench consists of two complementary components: (i) high-fidelity Unreal Engine depth data for evaluating depth estimation, and (ii) diverse real-world videos with rich multimodal annotations for assessing video generation quality.

models to extract modality-specific features and implement rigorous filtering pipelines to ensure data quality and usability.

Our data curation process follows strict quality criteria. We first filter source videos based on temporal, aesthetic, and resolution constraints: minimum duration of 5 seconds, aesthetic score exceeding 80/100, and spatial resolution above 512 pixels. Videos containing embedded text or subtitles are removed using OCR-based detection to prevent contamination of visual modalities. For each retained video, we extract corresponding modality annotations using specialized models—depth from Depth Anything V2, optical flow from RAFT, segmentation from SAM, skeleton from DWPose, and DensePose from Meta’s implementation. Automated quality metrics further filter low-quality modality extractions, ensuring reliable ground-truth annotations across all modalities.

Through this systematic pipeline, we obtain approximately 1.3M high-quality multimodal video pairs, each containing synchronized annotations across five modalities. This comprehensive dataset enables effective unified training while maintaining consistency and quality across diverse visual representations.

C.2. UniBench

To address the absence of standardized evaluation benchmarks for unified multimodal video tasks, we construct UniBench with two distinct evaluation categories tailored to different task requirements. For video estimation tasks requiring ground-truth annotations, we generate synthetic data using Unreal Engine to obtain pixel-accurate depth maps and optical flow. As shown in Figure 3, for controllable generation and text-to-video tasks requiring diverse modality conditions, we curate high-quality samples from our test split.

Specifically, we create 200 synthetic video sequences with precise ground-truth depth and optical flow using Unreal Engine’s rendering pipeline. These sequences feature significant camera and object motion to comprehensively evaluate depth estimation capabilities under challenging conditions. For generation tasks, we select 200 high-quality samples from the test subset, each containing complete annotations across all five modalities. This dual-track evaluation strategy enables rigorous assessment of both reconstruction accuracy and generation quality within our unified framework.

D. More Visuals and Applications

Figure 4 and 5 showcases UnityVideo’s extensive generalization capabilities across three core tasks: controllable generation, video estimation, and joint generation. The model accepts arbitrary modality inputs for precise controllable generation while supporting flexible modality estimation for diverse subjects and scenarios.

Our framework demonstrates remarkable zero-shot generalization beyond its training distribution. While trained primarily on single-person data, UnityVideo successfully generalizes to multi-person scenarios for all modality estimations. Similarly, skeleton estimation capabilities trained on human subjects transfer effectively to animal motion capture without additional fine-tuning. The model also exhibits robust cross-domain transfer, accurately estimating depth and segmentation for out-of-distribution objects and scenes. These diverse examples collectively demonstrate that UnityVideo’s unified training paradigm not only achieves technical integration across modalities but also develops genuine world understanding that enables flexible generalization to novel contexts and subjects.

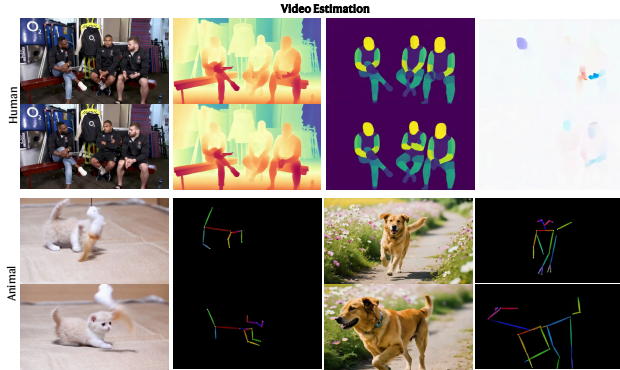


Figure 4. Representative outputs of UnityVideo on Video Estimation. The model consistently produces coherent RGB videos and aligned modalities—including densepose, optical flow, skeleton, and depth—demonstrating reliable cross-modal generation and estimation across diverse scenarios from human activities to animal motion.

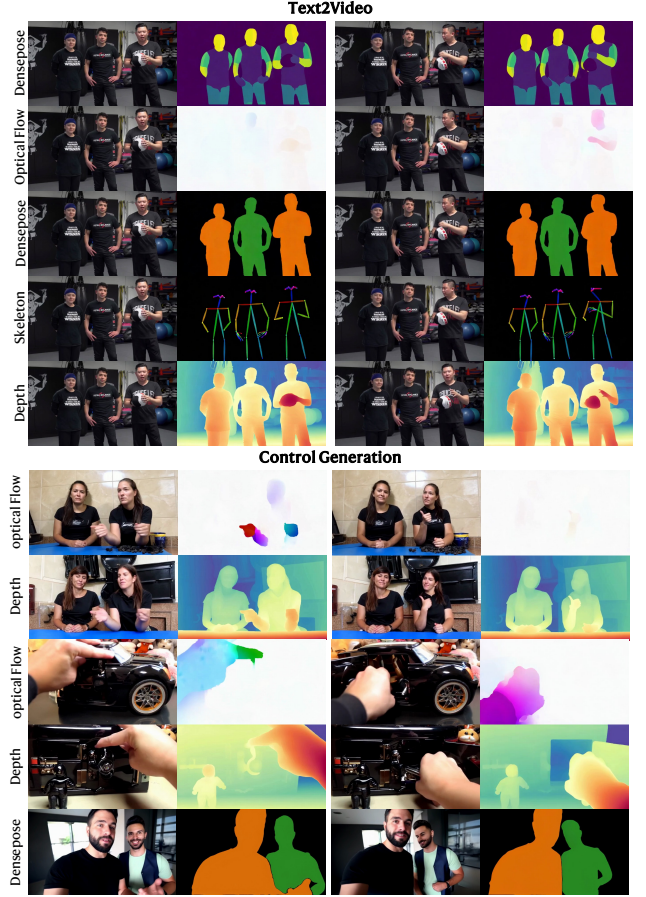


Figure 5. Representative outputs of UnityVideo on Text2Video and Control Generation. The model consistently produces coherent RGB videos and aligned modalities—including segmentation, densepose, optical flow, skeleton, and depth—demonstrating reliable cross-modal generation and estimation across various indoor and outdoor scenes with multiple subjects.