V-RGBX: Video Editing with Accurate Controls over Intrinsic Properties

Supplementary Material

Contents

A. Overview

In this appendix, we provide additional details and results that are not included in the main paper due to the space limit. The attached video includes intuitive and interesting qualitative results of V-RGBX.

B. Workflow & Implementation Details

B.1. Video editing workflow explanation

Intrinsic decomposition and keyframe editing. As shown in Fig. S1, we first decompose the input RGB video into intrinsic channels, including albedo, irradiance, normal, and material. These intrinsic channels form a physically structured representation that separates appearance, illumination, and geometry, enabling more reliable and controllable video editing. Selected keyframes are edited with a text-driven image editing tool NanoBanana and then decomposed again to obtain their edited intrinsics, ensuring that user-intended modifications (e.g., material changes or relighting) are reflected in the intrinsic domain.

Intrinsic conditioning sampling. To propagate the edits beyond the keyframes, we employ an intrinsic conditioning sampler that aggregates both the original per-frame intrinsics and the edited intrinsic channels. The sampler constructs an interleaved intrinsic sequence V'_X, inserting edited intrinsic cues at the keyframe positions while preserving unmodified channels for all other frames. This provides a unified intrinsic sequence that encodes both preserved and edited content in a temporally aligned manner.

Forward rendering of edited content. The interleaved intrinsic video is then passed through our forward renderer R, which synthesizes the final edited RGB video. The edited keyframes provide both edited intrinsic cues and reference appearance keyframes, and conditioning on intrinsics leverages their structured, disentangled nature to support faithful and controllable propagation of edits. We show more qualitative results in Sec. D.4 and D.5.

B.2. RGB→X→RGB cycle workflow

We adopt an RGB→X→RGB cycle setup to assess how well the intrinsic representation retains the information needed for accurate reconstruction and for supporting reliable edit propagation. This evaluation setting provides a clear and comprehensive way to examine how appearance, geometry-related cues, and illumination are preserved when passing through each stage of our framework.

As illustrated in Fig. S2, an input RGB video is first decomposed by our inverse renderer into its intrinsic channels. The predicted intrinsic sequence is temporally consistent, and the intrinsic output of the first frame additionally serves as a keyframe to anchor the forward synthesis. These intrinsic channels are then fed into our forward renderer to reconstruct the video. By comparing the reconstructed sequence with the original input, as reported in Table 3, we evaluate how well the intrinsic space maintains pixel-level fidelity, structural detail, and temporal continuity with baseline methods. This cycle analysis also indicates the stability of intrinsic-based edits when propagated across frames. We show more qualitative results in Sec D.3.

B.3. Inference details

During inference of forward rendering, classifier-free guidance is applied to the reference branch while keeping V'_X as a shared condition:

$$\epsilon_{\text{CFG}} = \epsilon_\theta(z_t, \varnothing, V'_X) + s[\epsilon_\theta(z_t, v_{\text{ref}}, V'_X) - \epsilon_\theta(z_t, \varnothing, V'_X)], \tag{S1}$$

where the two terms denote predictions without/with the reference input, respectively. Following the notation in the main text, the reference is defined as $v_{\text{ref}} = \{v'_{i_1}, \ldots, v'_{i_k}\}$. This modulates reference-driven appearance while preserving the structural/physical priors encoded in V'_X. In our implementation, the guidance scale is set as $s = 1.5$.

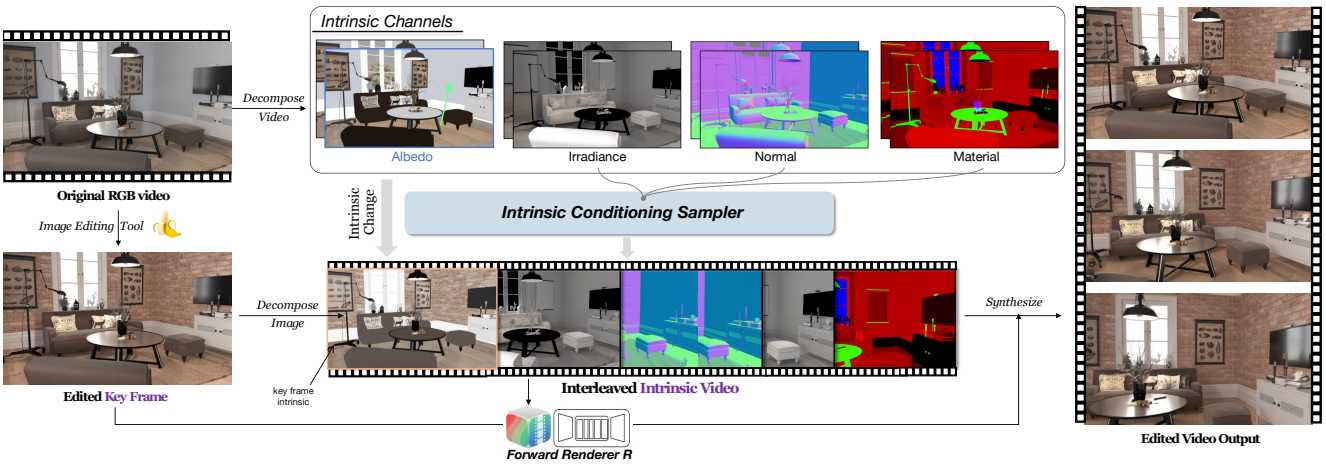

Figure S1. **Intrinsic-aware video editing workflow of V-RGBX.** Given an input video and edited keyframes, we decompose them into intrinsic channels, and the intrinsic conditioning sampler uses these representations to produce an intrinsic video. The forward renderer then synthesizes the final edited sequence using both the intrinsic video and the appearance cues provided by the edited keyframes.

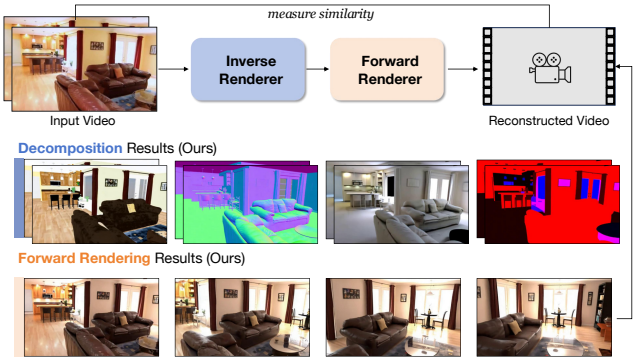

Figure S2. **Overview of RGB→X→RGB cycle workflow.** An input RGB video is first decomposed into intrinsic components by our inverse renderer, then reconstructed by the forward renderer using the predicted intrinsic sequence and a first-frame keyframe. The decomposition and forward-rendering results illustrate the quality of our intrinsic predictions and the rendered video.

Figure S3. **Visual ablations on the Temporal-aware Intrinsic Embedding (TIE) module and the reference condition.** Columns show the base model, adding TIE, adding both TIE and the reference condition (ours), and the ground truth. The intrinsic maps (top) and reconstructed RGB frames (bottom) illustrate that TIE reduces temporal and modality inconsistencies, while the reference condition further improves reflections and color fidelity.

C. Additional Experiments

In this section, we present additional ablation studies focusing on the two modules that most strongly affect the propagation behavior of our Forward Renderer: the Temporal-aware Intrinsic Embedding (TIE) module and the Reference Condition. For each ablation, we remove the corresponding module and retrain the model from scratch under the same training iterations and hyperparameters as V-RGBX. Both quantitative and qualitative analyses are provided.

C.1. The effectiveness of TIE module

As shown in Tab. S2, comparing the first and second rows reveals that removing the TIE module—and thus relying solely on interleaved intrinsic conditioning—leads to consistent drops across all evaluation metrics. The qualitative results in Fig. S3 further show noticeable temporal flickering and color inconsistencies. We observe that when intrin-

Table S1. **User study on Evermotion and in-the-wild videos.** Participants evaluate results based on Edit Alignment, Temporal Stability, Content Preservation, and Intrinsic Disentanglement. Scores are aggregated using Average User Ranking (AUR).

Dataset	Method	User Study				
		Edit Align.↑	Temp. Stab.↑	Cont. Pres.↑	Intri. Disen.↑	Overall↑
Evermotion	anyV2V	1.6392	1.8174	2.0652	1.9783	1.8751
	VACE	1.8391	1.6739	0.9783	1.2391	1.4326
	Ours	**2.5217**	**2.5087**	**2.9565**	**2.7826**	**2.6923**
In-the-wild	anyV2V	1.5286	1.4285	**2.3714**	2.4286	1.9393
	VACE	2.1563	2.1859	1.3857	1.0102	1.6845
	Ours	**2.3151**	**2.3856**	2.2429	**2.5612**	**2.3762**

Table S2. **Quantitative ablations of the TIE module and the reference condition.** Adding TIE consistently improves reconstruction quality and temporal stability across all metrics, and further incorporating the reference condition yields the best overall performance, with noticeable gains in PSNR, LPIPS, FID, FVD, and smoothness.

Method	TIE	Key Reference	PSNR↑	LPIPS↓	SSIM↑	FID↓	FVD↓	Smooth.↑
Base (no added modules)	×	×	20.96	0.2372	0.7818	37.81	532.21	0.9769
+ TIE	✓	×	21.47	0.2149	**0.7994**	35.39	405.79	0.9802
+ TIE + Reference (ours)	✓	✓	**22.42**	**0.1930**	0.7952	**29.83**	**367.89**	**0.9805**

Table S3. **Quantitative evaluation of intrinsic disentanglement**, split by edit type. ΔX measures the change on intrinsic channels that are not involved in the edit.

Method	Albedo Editing				Irradiance Editing			
	ΔNormal Ang.Err.↓	ΔRough. RMSE↓	ΔMetal. RMSE↓	ΔIrrad. si-PSNR↑	ΔAlbedo si-PSNR↑	ΔNormal Ang.Err.↓	ΔRough. RMSE↓	ΔMetal. RMSE↓
AnyV2V	22.17°	0.1297	0.2905	20.30	17.99	23.12°	0.1442	0.2269
VACE	25.65°	0.1713	0.3461	17.09	15.36	31.27°	0.1715	0.2502
Ours	**18.22°**	**0.1154**	**0.2547**	**22.81**	**19.46**	**16.06°**	**0.1317**	**0.2112**

Table S4. **Ablation on intrinsic sampling strategies.** We compare our sampling strategy with a fixed modality ordering in three settings: Full Intrinsics, Drop Channel, and 1st-Frame Guided.

Sampling Strategy	Full Intrinsics		Drop Channel		1st-Frame Guided	
	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
Fixed Ordering	20.91	0.202	10.86	0.629	11.07	0.601
Ours	**22.42**	**0.193**	**20.83**	**0.235**	**21.65**	**0.206**

sic channels are interleaved on a per-frame basis without explicit type disambiguation, the model may confuse modality identities across frames. Such confusion can couple signals across different channels in the color space, causing visually incorrect or unstable predictions.

After introducing the TIE module, all metrics improve, and the generated videos (Fig. S3, second column) exhibit much better temporal stability and appearance consistency. This confirms that TIE provides an effective mechanism for disentangling the intrinsic channels over time, reducing cross-channel conflicts and producing more reliable visual outcomes.

C.2. The effectiveness of reference condition

As shown in Tab. S2, adding the reference condition leads to further improvements over using TIE alone. Note that this experiment evaluates whether the model is trained with reference supervision, which differs from the evaluation in the main paper where we study whether reference images are provided at inference time. Here, the ablation aims to understand the contribution of the reference module itself.

The qualitative comparisons also show consistent gains: in the first two rows of Fig. S3, reflections become clearer and more coherent, while in the third and fourth rows, object colors and tones align more closely with the ground truth. These observations suggest that the reference condition offers important complementary cues that help correct biases in the X→RGB mapping and significantly improve reconstruction fidelity.

C.3. User-centric study of intrinsic editing results

We further conduct a user-centric study to evaluate the perceptual quality and faithfulness of intrinsic editing results. We collect 25 indoor scenes from Evermotion and 25 real-world editing cases, and invite 16 participants to rank the results produced by different methods. The evaluation considers four aspects: Edit Alignment, Temporal Stability, Content Preservation, and Intrinsic Disentanglement. The final scores are aggregated using Average User Ranking (AUR). The results are summarized in Tab. S1.

Our method consistently achieves the best overall performance on both datasets, indicating that it more faithfully propagates the keyframe editing intent while maintaining stronger temporal stability and better intrinsic disentanglement compared to existing baselines.

C.4. More evaluation of intrinsic disentanglement

To quantitatively verify intrinsic disentanglement, we measure whether editing one intrinsic channel introduces unintended changes to other channels. Using the editing cases described in Sec. C.3, we estimate per-frame intrinsic maps using RGB↔X and compute ΔX on modalities that are not involved in the edit (i.e., the change of each intrinsic channel before and after editing, measured by similarity metrics such as PSNR or RMSE). As shown in Table S3, our method consistently produces smaller changes on unchanged intrinsic channels, indicating better disentanglement and reduced attribute leakage during editing.

C.5. The effectiveness of sampling strategy

We further evaluate the effectiveness of the intrinsic sampling strategy introduced in the main paper (Sec. 3.3). We compare our proposed sampling with a fixed modality ordering during conditioning. As shown in Table S4, fixed ordering achieves comparable performance for standard X→RGB reconstruction tasks. However, it degrades

noticeably under more challenging settings such as missing modalities (Drop Channel) and keyframe intrinsic propagation (1st-Frame Guided). These results suggest that our sampling strategy helps resolve conflicts among intrinsic modalities during conditioning, improving robustness under incomplete intrinsic inputs.

D. Additional Qualitative Results

D.1. Additional RGB→X results

In the main paper (Sec. 4.2), we have already reported quantitative results for the inverse-rendering task (RGB→X). Here, we provide additional qualitative results in Fig. S4 and Fig. S5, along with representative comparisons against baseline methods.

Fig. S4 and Fig. S5 show input videos from both synthetic and real scenes together with the intrinsic predictions produced by V-RGBX, including albedo, normal, material, and irradiance channels. We further compare our method with RGBX and DiffusionRenderer in Fig. S6. Consistent with the quantitative findings, V-RGBX yields more stable albedo and normal reconstructions, while RGBX often exhibits temporal instability and color inconsistencies. DiffusionRenderer also shows some failure cases, such as collapsed normal maps and inaccurate color estimates. Moreover, our model demonstrates strong generalization ability, producing reliable intrinsic decompositions even under challenging real-world and outdoor lighting conditions.

D.2. Additional X→RGB results

As discussed in Sec. 4.3, we evaluate the X→RGB task, and Fig. S8 provides additional qualitative examples together with comparisons against the baseline methods. These examples show that V-RGBX handles complex lighting effects and geometric structures more reliably, producing RGB sequences with more stable shading, reflections, and temporal coherence. Overall, the supplemental results further illustrate the robustness of our approach when generating videos from intrinsic representations.

D.3. Additional RGB→X→RGB results

As discussed in the main paper (Sec. 4.3), we quantitatively evaluate the RGB→X→RGB cycle to assess whether the intrinsic representation preserves sufficient information for accurate reconstruction and reliable edit propagation. In Figs. S9 and S10, we provide additional qualitative comparisons on both synthetic and real-world videos, showing the reconstructed sequences produced by our approach and the baseline methods. Our method achieves more stable temporal behavior and better preserves scene appearance across the full cycle.

D.4. Keyframe editing results

As described in Sec. 4.5, we demonstrate the intrinsic-aware video editing capability of V−RGBX in the main paper. To provide a clearer view of how the edits are propagated through our intrinsic pipeline, we include the complete set of intermediate results in Fig. S11. Specifically, we visualize the input video frames, the edited keyframes produced by the NanoBanana tool, and the extracted intrinsic channels that jointly condition the generation process. These intermediate visualizations help illustrate how the edited albedo, normal, material, or irradiance attributes guide the final synthesis. Following the editing workflow shown in Fig. S1, V-RGBX takes the modified keyframes and intrinsic channels as conditioning signals and generates temporally consistent, intrinsically coherent outputs.

D.5. Real-world challenging cases

We provide additional demonstrations of our editing capability on diverse and challenging real-world scenarios. Please refer to the attached video for full results. Our evaluations cover real indoor scenes, self-captured videos, general object videos, and cases with complex lighting, showcasing robust intrinsic-aware editing performance across a wide range of real-world conditions.

As discussed in the main paper (Sec. 5), ground-truth intrinsic annotations for real-world videos are notoriously difficult to acquire, which motivates our reliance on synthetic supervision during training and inevitably introduces a domain gap when applied to real-world scenes. Despite this limitation, our model still shows encouraging transfer to real-world data, as evidenced by improved cycle-consistency performance on Re10K and consistent qualitative results across diverse in-the-wild examples (Figs. S5, S6, S7, and S10). Beyond editing applications, our framework may also serve as a useful data engine for producing intrinsic annotations or editing supervision for in-the-wild videos, helping scale intrinsic-aware video generation and editing to more diverse real-world data.

For dynamic scenes, although the training data mainly contains static subjects with moving cameras, we observe that the model can still handle moderately dynamic scenarios. We conjecture that camera motion in the training data may also encourage motion-aware intrinsic propagation across frames. However, cases involving strong object motion or rapid camera movement remain challenging. Incorporating explicit motion cues, such as camera trajectories or tracking features, is a promising direction for improving intrinsic propagation under dynamic conditions.

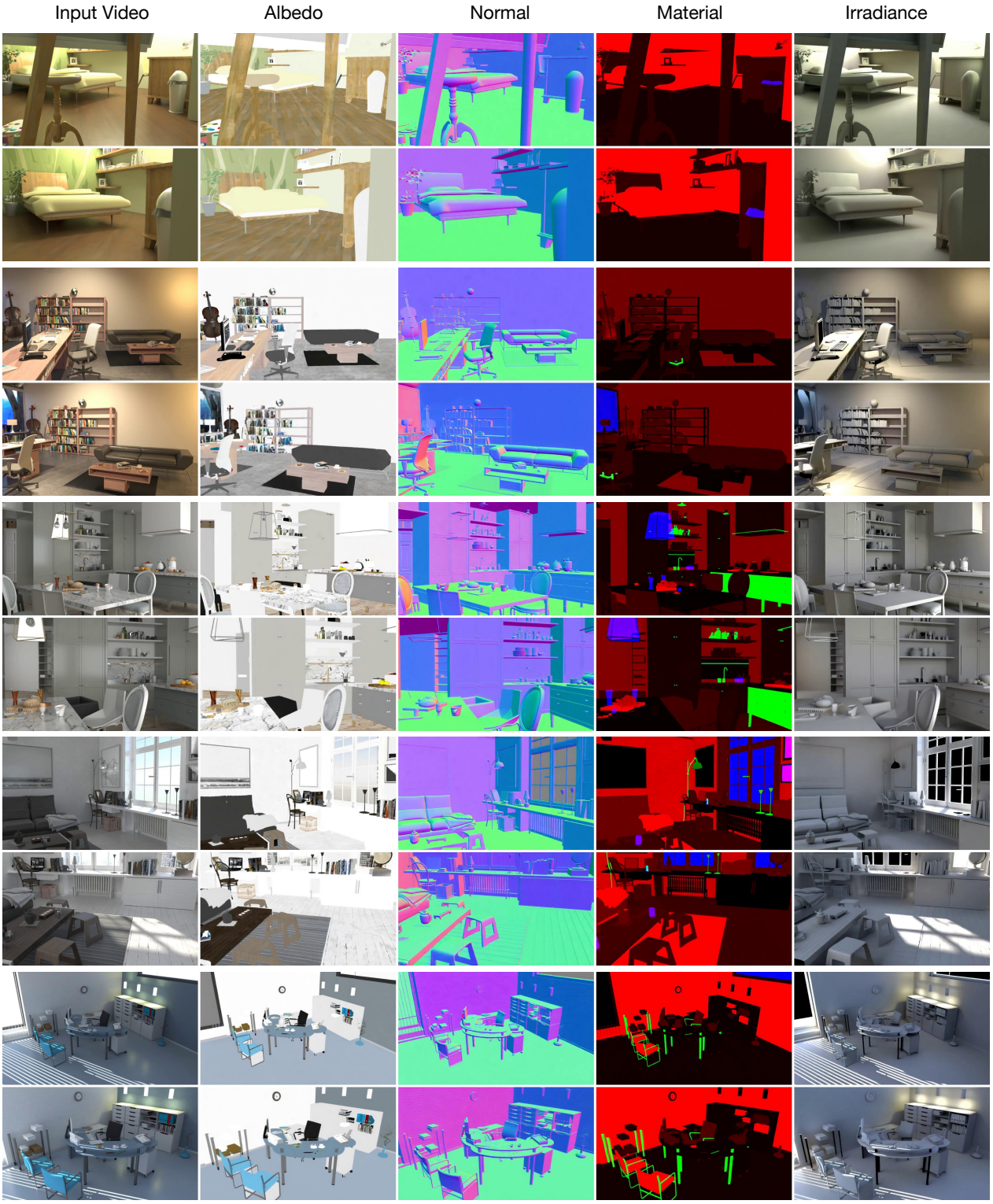

Figure S4. **RGB→X results on synthetic Evermotion scenes.** Given an input RGB video, V-RGBX decomposes it into albedo, normal, material, and irradiance channels. Each pair of rows shows two frames from the same video, and the second to fifth columns visualize the corresponding intrinsic channels, demonstrating spatially coherent and temporally stable decompositions across diverse indoor scenes.

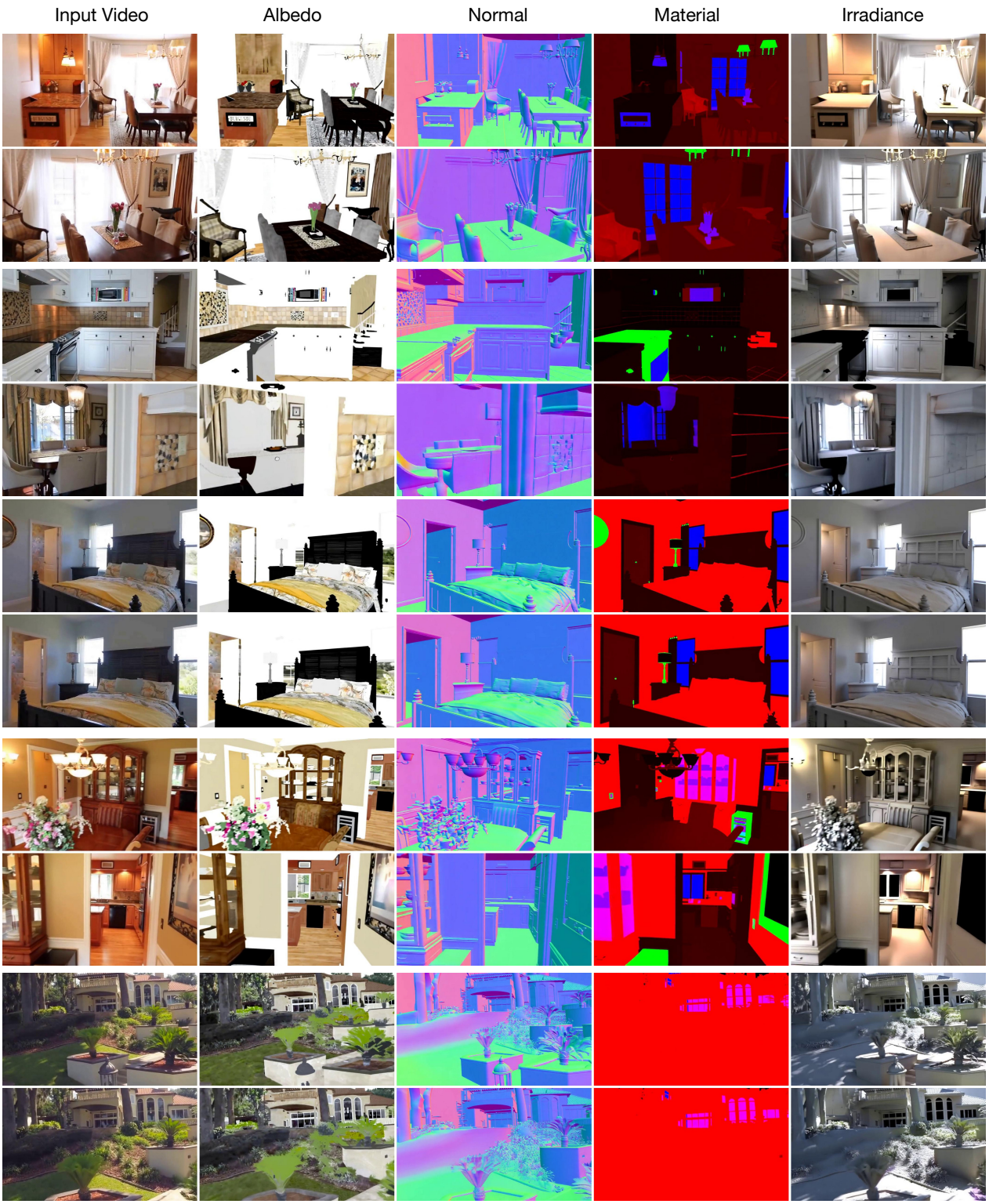

Figure S5. **RGB→X results on real-world RealEstate10K videos.** Given an input RGB video, V-RGBX decomposes it into albedo, normal, material, and irradiance channels. Each pair of rows shows two frames from the same video, and the second to fifth columns visualize the corresponding intrinsic channels, demonstrating coherent and temporally stable decompositions under challenging and unseen real-world conditions, while also showing reasonable generalization to outdoor scenes.

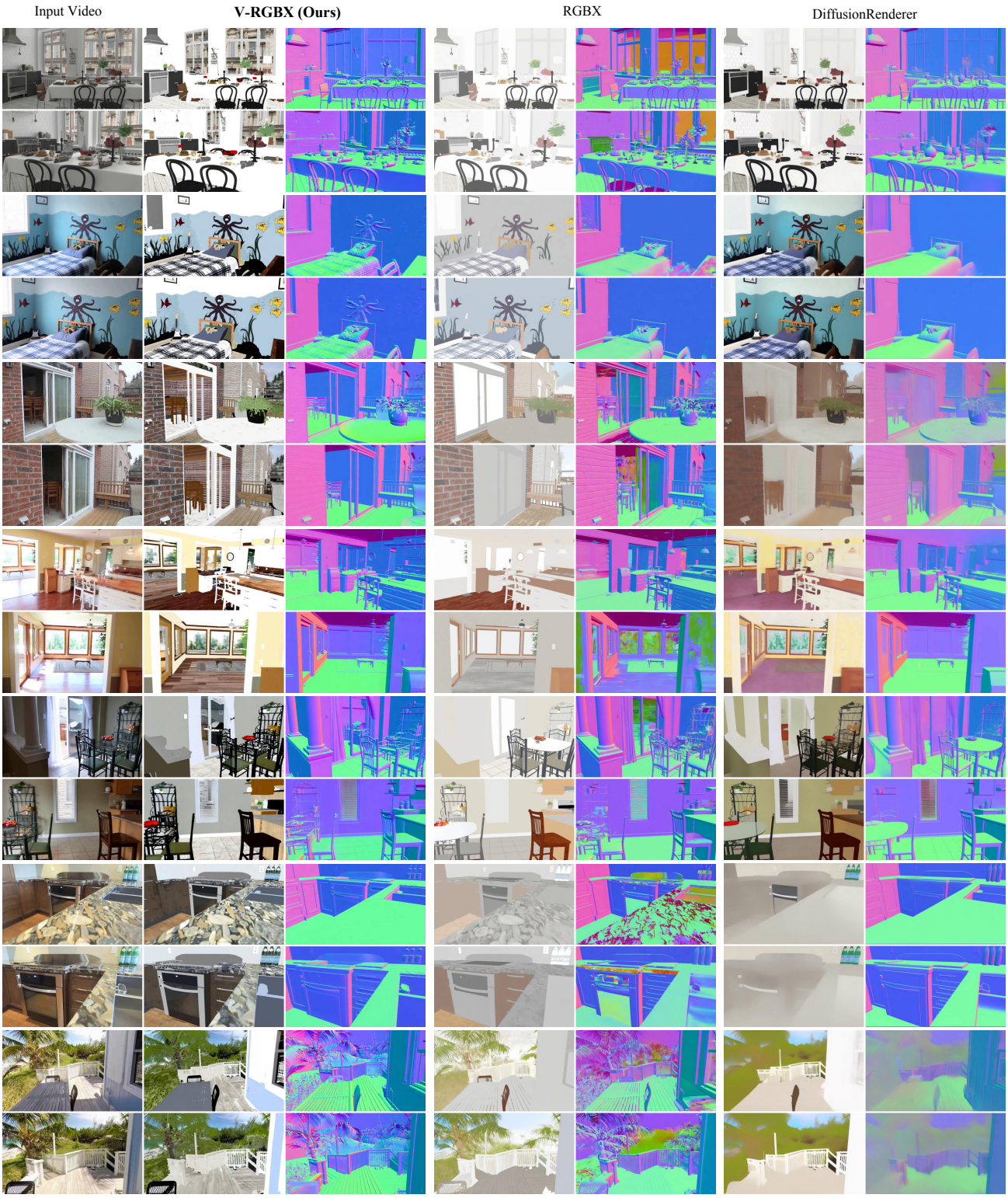

Figure S6. **Comparison of RGB→X decomposition results with baselines.** Each pair of rows shows two frames from the same input video (first column). For each method, the two columns visualize the predicted albedo and normal channels. Compared with RGBX and DiffusionRenderer, V-RGBX produces intrinsic decompositions with higher visual fidelity, more accurate albedo estimation, and more consistent normal predictions across frames.

Figure S7. **Comparison of irradiance decomposition with baselines.** The figure shows two different videos, with each pair of rows representing two frames from the same video. For each frame, the second and third columns show irradiance predictions from V-RGBX and RGBX. V-RGBX produces more accurate illumination and shadow modeling, resulting in clearer and more plausible irradiance maps.

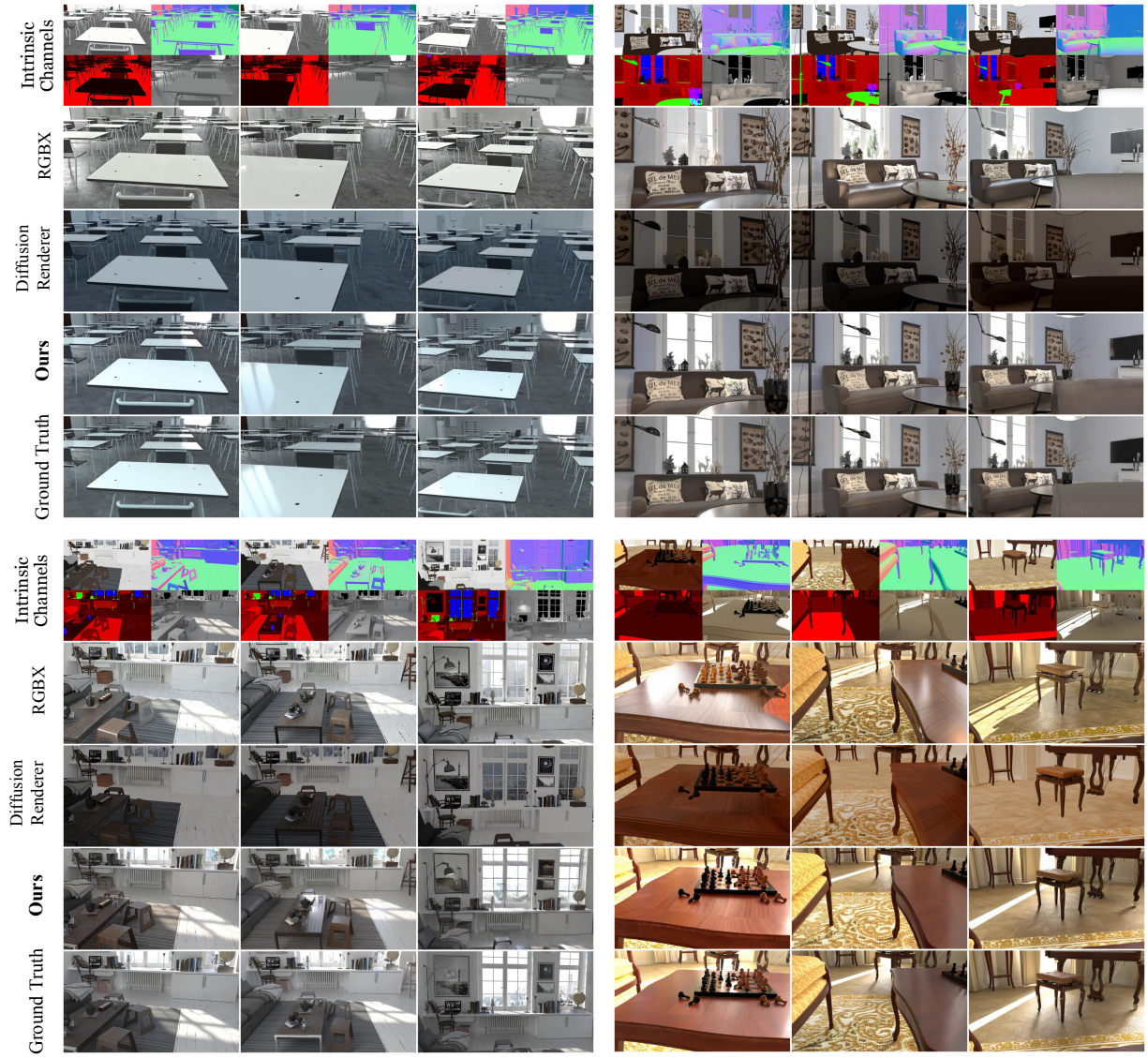

Figure S8. **More qualitative comparisons for the X→RGB task.** Each group of three columns shows three frames from the same video, while each row corresponds to a different method: intrinsic channels inputs, RGBX, DiffusionRenderer, our results, and the ground truth. The comparisons illustrate that our method performs better in scene appearance and temporal consistency across frames.

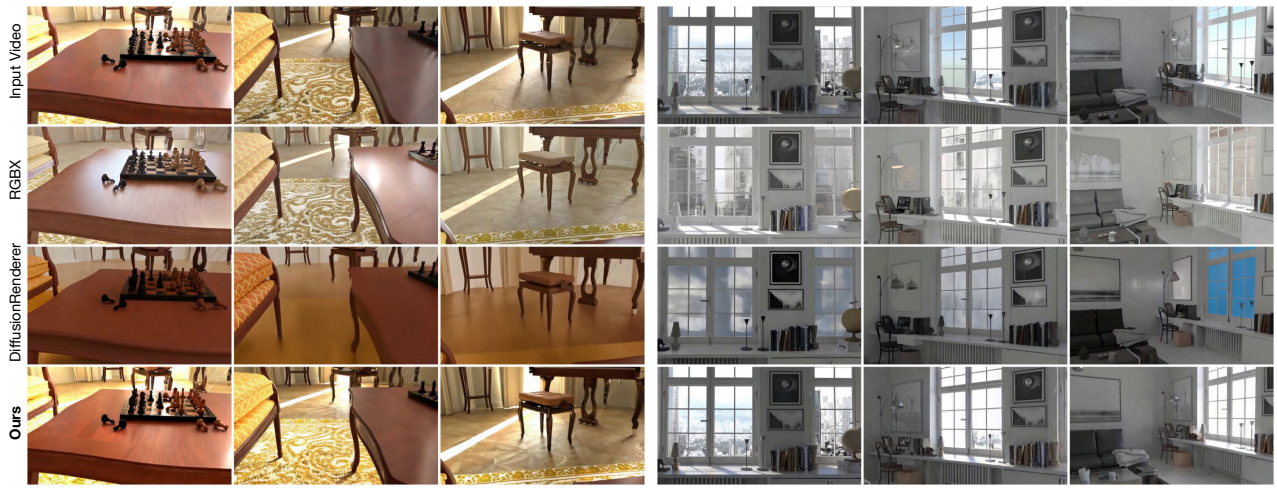

Figure S9. **RGB→X→RGB cycle results on the synthetic dataset.** Each row shows a different method (the first row is the input video as ground truth). Every three columns correspond to three frames from the same video. Our method produces reconstructions closest to the ground truth and better preserves scene appearance and structure throughout the sequence.

Figure S10. **RGB→X→RGB cycle results on the real-world dataset.** Each row shows a different method (the first row is the input video as ground truth). Every three columns correspond to frames from the same video. Our method gives a closer match to the ground truth.

Figure S11. **Intermediate results of the intrinsic-aware keyframe editing.** We visualize intermediate results used by V-RGBX across the following editing types: (1) solid color, (2) texture, (3) material, (4) normal, (5) light color, and (6) shadow editing. For each case, we show the original video frames, the edited keyframe produced by the NanoBanana tool, and the corresponding modified intrinsic channels (albedo, material, normal, or irradiance) that serve as conditioning inputs. These processes reveal how keyframe edits are translated into intrinsic-space modifications, which are then reliably propagated by V-RGBX to generate the final temporally consistent edited video.