PROBING MATERIAL CUES IN VIDEO FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

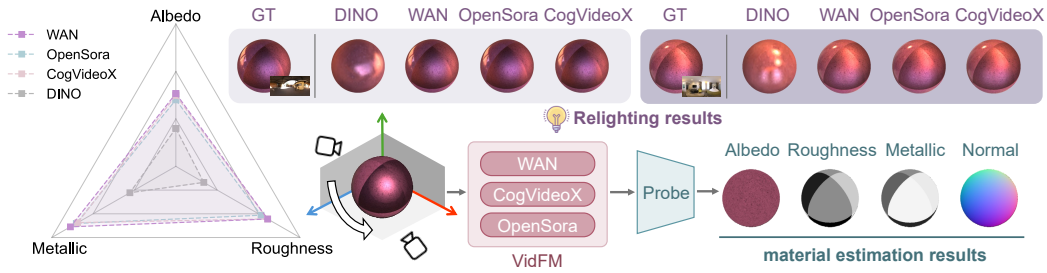


Figure 1: We probe material-relevant information in frozen pretrained video representations and characterize their complementarity with image representations for material-property estimation.

ABSTRACT

General-purpose video foundation models (VidFMs) generate realistic videos that capture how surface appearance evolves across viewpoints and illumination conditions, suggesting that their internal representations may encode intrinsic material information even without explicit material supervision. Motivated by this observation, we investigate the material information embedded in pretrained VidFMs through a lightweight probing framework. We keep the foundation models frozen and measure how well material properties can be recovered from their intermediate representations. Through systematic experiments across synthetic and real-world scenes, together with feature-source ablations, we find that VidFM features provide material cues beyond those available from strong image representations. In controlled settings requiring fine-grained material discrimination, video features show a pronounced advantage over DINOv2 features, while the two representations provide complementary information in more complex scenes. Further analyses show that material recoverability benefits from broader viewpoint coverage and is influenced by object-level semantics. Qualitative synthetic-to-real results further suggest that these material cues remain useful beyond the supervised training domain. Together, these findings provide evidence that large-scale video generative models learn reusable representations containing material-relevant information and characterize the conditions under which these representations offer advantages over image features.

1 INTRODUCTION

Recovering material properties from 2D visual observations is a long-standing problem in computer vision, with applications in realistic relighting, AR/VR, digital content creation, and scene understanding. Material reconstruction is inherently ambiguous because observed appearance is jointly determined by material, geometry, illumination, and viewpoint, making these factors difficult to disentangle under uncontrolled conditions. Existing image-based material reconstruction methods have made substantial progress. Per-scene inverse rendering approaches (Ye et al., 2024; Liang et al., 2024; Gao et al., 2024; Kouros et al., 2026; Jiang et al., 2024) jointly optimize material attributes, scene geometry, and illumination, but are typically computationally intensive and limited in cross-scene generalization. Feed-forward methods (Wu et al., 2025; Kocsis et al., 2025) directly predict material

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

properties from images, but depend strongly on supervised material annotations, which are scarce in the real world and often replaced by synthetic training data. These approaches primarily focus on solving material reconstruction itself, while providing limited insight into whether material-relevant information is already encoded in large-scale pretrained visual representations.

Videos offer a promising source of such representations. Compared with densely annotated material datasets, video data can be collected at substantially larger scale (Miech et al., 2019; Abu-El-Haija et al., 2016). More importantly, videos capture appearance variations induced by camera motion, illumination changes, and view-dependent reflectance. While intrinsic material properties remain largely stable, specular highlights, shading, reflections, and other appearance effects vary across observations. Such consistency and variation provide informative cues for material discrimination, suggesting that representations learned from video may encode material-relevant information beyond that available from individual images.

General-purpose video foundation models (VidFMs) provide a natural setting for investigating this possibility. Trained on massive video collections, they generate temporally coherent content in which surface appearance evolves across viewpoints and illumination conditions. This capability suggests that, despite being pretrained without explicit material annotations, their internal representations may contain cues correlated with intrinsic material properties. Recent studies have analyzed VidFMs from the perspectives of semantics (Zhu et al., 2024; Wang et al., 2024a), motion (Wang et al., 2024b), and 3D geometry (Huang et al., 2025), while their material-related representations remain much less characterized.

In this work, we systematically investigate the material information embedded in pretrained VidFMs, as illustrated in Figure 1. We characterize material awareness through the recoverability of albedo, roughness, metallicity, and surface normals from frozen representations using lightweight probing heads. The VidFM backbones remain fixed throughout training, allowing us to examine material information already present in pretrained representations without task-specific fine-tuning. We evaluate three representative general-purpose VidFMs—CogVideoX, OpenSora, and WAN—under a common probing setup and compare their material recoverability with strong image-based representations. We additionally include DiffusionRenderer as a task-adapted video-model baseline: it builds on video pretraining but has already undergone material-supervised adaptation. Throughout this paper, claims about material cues learned without explicit material supervision refer to the three general-purpose VidFM backbones, not to DiffusionRenderer or the supervised probes. We further study video and image features independently and in combination, allowing us to separate material cues encoded in video representations themselves from complementary information provided by image features. Rather than developing another end-to-end material reconstruction pipeline, our goal is to reveal and characterize material-relevant information already present in pretrained video representations.

Our experiments reveal substantial material-relevant information in pretrained VidFM representations. In controlled settings requiring fine-grained material discrimination, video features exhibit a pronounced advantage over DINOv2 features, even without image-feature fusion. In more complex scenes, the two representations become complementary, with their combination improving material recovery over either source alone. Further analyses show that broader viewpoint coverage improves material recoverability, particularly for metallicity, while object–material compatibility reveals an additional influence of semantic priors. Qualitative synthetic-to-real results further suggest that these material cues remain useful beyond the supervised training domain. Together, these results provide evidence that pretrained video generative models encode reusable material-relevant information, while characterizing both its advantages over DINOv2 features and the conditions under which these advantages emerge.

2 Related Work

2.1 Foundation models

Recent years have witnessed remarkable progress in image and video generation, with large-scale image and video foundation models trained on massive real-world image and video datasets demonstrating strong pixel-level generation capabilities (Oquab et al., 2023; Yang et al., 2024; Peng et al., 2025; Wan et al., 2025; Team, 2024; Kong et al., 2024). To further evaluate how well foundation models can understand and reason about the physical world, some researchers have begun benchmarking

these models on 3D vision tasks, such as depth estimation (El Banani et al., 2024; Zuo et al., 2025), multi-view object recognition (Bonnen et al., 2024), semantic consistency (Lilova et al., 2025; Zhang et al., 2023), and spatial reasoning (Zuo et al., 2025; Yang et al., 2025b;a; Ray et al., 2024) through visual question answering (Zuo et al., 2025; Chow et al., 2025) with performance often compared against human perception and cognition (Zuo et al., 2025; Bonnen et al., 2024; Fu et al., 2024). Complementarily, probe-based methods offer a more fine-grained perspective by directly examining the internal representations of foundation models, assessing whether they implicitly encode 3D structure (El Banani et al., 2024; Chen et al., 2025; Huang et al., 2025), spatial consistency (Man et al., 2024; Chen et al., 2025; Huang et al., 2025), and semantic correspondence (Man et al., 2024), independent of the model's language output ability. However, these studies primarily focus on the geometric and semantic awareness of foundation models, overlooking one critical aspect of physical world understanding: material properties. In the real world, objects are not only defined by their 3D structure and semantics, but also by their intrinsic physical properties, such as albedo, roughness, metallicity, and surface normal, which control how light interacts with surfaces. Whether large-scale video foundation models encode information related to intrinsic material properties remains an open and underexplored question.

2.2 Material estimation methods

Some existing works use per-scene optimization-based methods (Zhang et al., 2021; Boss et al., 2021; Ye et al., 2024; Liang et al., 2024; Gao et al., 2024; Li et al., 2025; Jiang et al., 2024) to estimate materials from multi-view images. However, these methods require scene-specific training data and typically take tens of minutes to optimize a single scene, making them computationally expensive and impractical at scale. Furthermore, the resulting models lack generalizability, as they require retraining from scratch. In contrast, directly estimating materials from images, without requiring per-scene optimization, offers a more generalizable and efficient alternative. Several feed-forward methods (Zeng et al., 2024b; Chen et al., 2024; Wu et al., 2025; Kocsis et al., 2025; Liang et al., 2025a; Munkberg et al., 2025) have demonstrated promising performance in directly predicting materials from images. However, these approaches rely on supervised training with ground-truth PBR material annotations (Zhu et al., 2022; Roberts et al., 2021; Li et al., 2023), which are extremely scarce compared to the billion-scale real-world data (Abu-El-Haija et al., 2016; Miech et al., 2019; Bain et al., 2021) used to train image and video generation models. Moreover, for real-world captured videos, ground-truth PBR annotations are nearly impossible to obtain. This data bottleneck fundamentally limits the generalizability of existing feed-forward material estimation models. Our focus is on general-purpose video foundation models pretrained without explicit material annotations. DiffusionRenderer also builds on video pretraining, but adapts its backbone to inverse rendering with material supervision; we therefore treat it as a task-adapted video-model baseline rather than evidence of material recoverability before material-specific adaptation. In this work, we show that general-purpose VidFM representations encode recoverable PBR material information, and demonstrate that video features offer stronger domain generalization than existing feed-forward models and the image-based features (Oquab et al., 2023) they rely on.

3 Method

We study the *material awareness* of video foundation models (VidFMs), namely, the extent to which material properties can be recovered from frozen visual representations using a lightweight probe.

As shown in Figure 2, we extract frozen video and image features, align and optionally fuse them, and decode material maps with a supervised probe. Only the projector and probe are trained. We compare individual feature sources and their combination using this architecture.

3.1 Feature Extraction

Diffusion Feature Extraction. Let $\mathbf{V} \in \mathbb{R}^{T_v \times 3 \times H_v \times W_v}$ denote an input video clip with frames $\{\mathbf{v}_t\}_{t=0}^{T_v-1}$, where T_v is the number of frames, H_v and W_v are the video height and width, and $\mathbf{v}_t \in \mathbb{R}^{3 \times H_v \times W_v}$.

We follow a DIFT-style (Tang et al., 2023) extraction procedure. Specifically, we first use the VidFM image encoder to encode the input frames into latent representations. We then add noise to

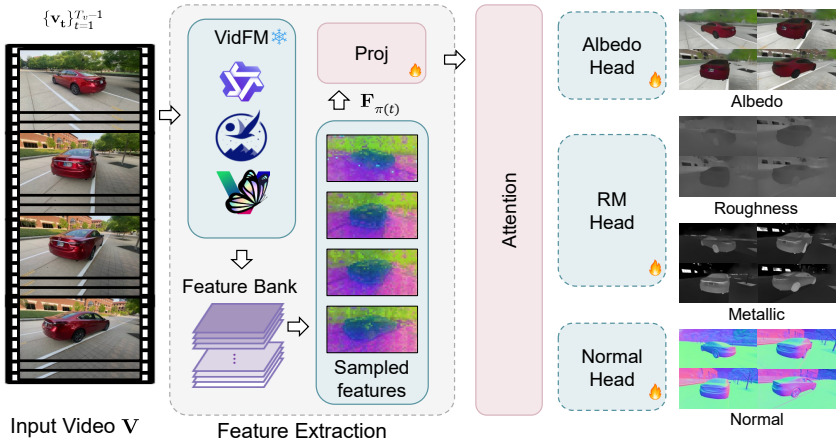


Figure 2: **The overview of the probe pipeline.** We extract various video features and image-based features from the input video, and keep a feature bank with fused video and image features. Four frames are sampled from the sequence and fed into a lightweight probe with three readout heads to decode albedo, roughness-metallic (RM), and normal maps for each frame.

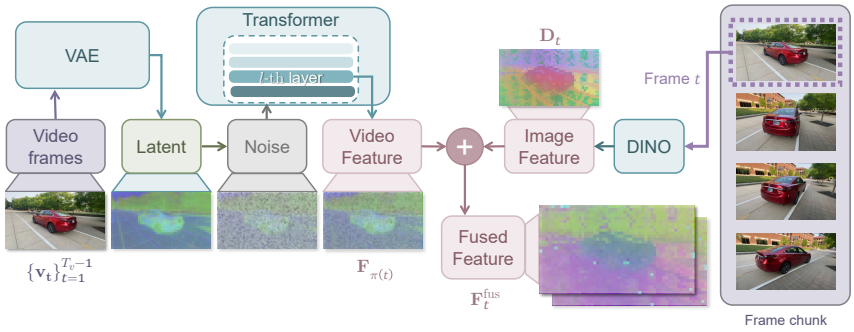


Figure 3: Feature extraction pipeline. Given an input video, we extract temporal features from a video foundation model alongside image-level features extracted from DINO. The two features are fused and fed into a lightweight probe for material estimation.

the latents at denoising timestep τ and feed the noised latents into the VidFM denoising network. Let $\{\mathbf{F}_c\}_{c=0}^{C_V-1}$ denote the resulting sequence of video features, where each $\mathbf{F}_c \in \mathbb{R}^{n_v \times H_f \times W_f}$ has spatial resolution $H_f \times W_f$, n_v is the feature dimension, and C_V is the temporal resolution. Because the video encoder reduces the temporal resolution of the input sequence, we generally have $C_V < T_v$, so the original T_v frames are grouped into C_V frame chunks and multiple input frames may correspond to the same chunk-level feature. We therefore introduce a frame-to-feature index map $\pi : \{0, \dots, T_v - 1\} \to \{0, \dots, C_V - 1\}$ and denote by $\mathbf{F}_{\pi(t)} \in \mathbb{R}^{n_v \times H_f \times W_f}$ the video feature assigned to frame $\mathbf{v}_t$. The hidden activation from the l-th layer of the denoising network provides the feature content of each $\mathbf{F}_c$.

Feature Fusion. We independently extract a DINOv2 feature map $\mathbf{D}_t \in \mathbb{R}^{n_d \times H_f \times W_f}$ for each frame, where n_d is the feature dimension. Feature extraction is identical during training and inference.

For each frame $\mathbf{v}_t$, we concatenate the VidFM feature $\mathbf{F}_{\pi(t)} \in \mathbb{R}^{n_v \times H_f \times W_f}$ and the DINO feature $\mathbf{D}_t \in \mathbb{R}^{n_d \times H_f \times W_f}$ along the channel dimension to obtain a fused feature map $\mathbf{F}_t^{\text{fus}}$:

$$\mathbf{F}_t^{\text{fus}} = \left[\mathbf{F}_{\pi(t)} \,\|\, \mathbf{D}_t\right] \in \mathbb{R}^{(n_v+n_d) \times H_f \times W_f}. \tag{1}$$

The extraction and alignment procedure is illustrated in Figure 3.

3.2 Material Probe

Architecture. For each input video $\mathbf{V}$, we sample four frames: the first frame $\mathbf{v}_0$ serves as a shared reference, and three additional frames are sampled as $\{\mathbf{v}_{t_{s_1}}, \mathbf{v}_{t_{s_2}}, \mathbf{v}_{t_{s_3}}\}$. For each sampled frame, we

project the fused feature map to a shared 1024-dimensional feature space,

$$\mathbf{P}_{t_{s_j}} = \mathrm{Proj}(\mathbf{F}^{\text{fus}}_{t_{s_j}}) \in \mathbb{R}^{1024\times H_f\times W_f}, \quad j = 0, 1, 2, 3, \tag{2}$$

where $\mathrm{Proj}(\cdot)$ is a learned projector consisting of a LayerNorm followed by a one-layer MLP without an activation function. We then convert $\{\mathbf{P}_{t_{s_j}}\}_{j=0}^{3}$ into per-frame tokens and apply four alternating-attention blocks. Each block contains a frame attention module that mixes tokens within a frame and a global attention module that mixes tokens across sampled frames, following the high-level design of VGGT. Finally, three DPT-style readout heads decode the material-related outputs for each sampled frame $\mathbf{v}_{t_{s_j}}$: an albedo map $\mathbf{A}_{t_{s_j}} \in \mathbb{R}^{3\times H_v\times W_v}$, a roughness-metallic map $\mathbf{RM}_{t_{s_j}} \in \mathbb{R}^{2\times H_v\times W_v}$, and a normal map $\mathbf{N}_{t_{s_j}} \in \mathbb{R}^{3\times H_v\times W_v}$.

Training. During training, we keep all foundation-model backbones frozen and optimize only the probe-side parameters, including the projector and the material-awareness probe. For each training video, we sample the four frames described above, gather their fused features, and supervise the resulting predictions with a multi-task objective similar to VGGT:

$$\mathcal{L} = \mathcal{L}_{\text{A}} + \mathcal{L}_{\text{RM}} + \mathcal{L}_{\text{N}}. \tag{3}$$

Here, $\mathcal{L}_{\text{A}}$, $\mathcal{L}_{\text{RM}}$, and $\mathcal{L}_{\text{N}}$ are all L1 losses between the predicted outputs and the corresponding ground-truth targets. Unless otherwise stated, this objective encourages the probe to recover material-related quantities from frozen visual representations, and thereby measures the material awareness encoded in those representations.

Inference and downstream model. Inference follows the same frame sampling and decoding procedure with all parameters fixed. For downstream validation, we also integrate VidFM features into a deeper MVInverse-based system, **VidFM-MatHead**, with an auxiliary cross-view consistency loss. Its architecture and regularization are described in Section 9 of the supplementary material.

4 Experiments

4.1 Data construction

We conduct experiments across three datasets: one real-world dataset **DL3DV** (Ling et al., 2024), our Room-scale synthetic dataset **RoomTour** rendered with Blender (Blender Online Community, 2018), and a synthetic **Ball** dataset.

DL3DV is a large-scale real-world dataset containing diverse indoor and outdoor scenes with challenging reflections and lighting conditions. Since no ground-truth PBR annotations are available, we apply MVInverse (Wu et al., 2025) to generate pseudo ground-truth material maps for each image. For each scene, we randomly select a starting frame and sample 81 consecutive frames, resulting in 5,400 sequences for training and 600 sequences for testing.

RoomTour is a photorealistic synthetic video dataset rendered from indoor Blender scenes sourced from BlenderKit (BlenderKit s.r.o., 2026) and Infinigen Room (Raistrick et al., 2024), designed to simulate real-world indoor environments while providing accurate ground-truth PBR maps for evaluation. Camera trajectories simulate handheld video capture along room-scale touring paths, and each frame is rendered with physically-based lighting in Blender, yielding ground-truth RGB images paired with ground-truth material maps. The dataset comprises 26209 videos across 2270 Blender scenes, with 129 sequences reserved for testing and no scene overlap between training and test splits.

Ball is a synthetic dataset where each ball sample is textured with an albedo map from MatSynth (Vecchio & Deschaintre, 2024) or Textures Dataset (Huang et al., 2020). Each ball is partitioned into 8 equal regions, with each region assigned roughness and metallic values independently sampled at random from (0,1). RGB images are rendered using physically-based rendering[1] under environment lighting sampled from PolyHaven (Poly Haven, 2021). The dataset comprises 10k training and 1k test sequences.

Evaluation scope. RoomTour and Ball provide ground-truth PBR maps and therefore form the primary basis for our quantitative conclusions about material recovery. DL3DV provides supporting

[1]Details of the rendering equation are provided in the Supplementary Materials

evidence on real-world imagery, but its numerical results measure agreement with pseudo material annotations rather than definitive physical ground truth.

Compared Representations and Metrics We evaluate three general-purpose pretrained VidFMs: CogVideoX-2B (Yang et al., 2024), Open-Sora2.0 (Peng et al., 2025), and WAN2.1-14B (Wan et al., 2025). We additionally evaluate DiffusionRenderer (Liang et al., 2025a) as a task-adapted video-model baseline whose backbone has received material supervision. We use DINOv2 (Oquab et al., 2023) as the primary image-feature baseline. DINOv2 features are extracted independently from each frame, while VidFM features are extracted following Section 3. All feature-extraction backbones remain frozen during our probe training; this does not imply the absence of material supervision in their earlier training. The resulting representations are decoded using the probing architecture described in Section 3. Here, frame-wise DINOv2 extraction refers to the backbone feature-extraction stage, rather than a change to the supervised prediction task. We also compare with RGB-X (Zeng et al., 2024b) as an additional image-based pipeline but specifically for material estimation.

The OpenSora, CogVideoX, and WAN variants in the main comparison use the fused VidFM–DINOv2 representation defined in Section 3. The DiffusionRenderer baseline likewise fuses its video features with DINOv2. We separately evaluate DINOv2-only, WAN-only, and fused features to assess material recoverability with each feature source.

We further adapt MVInverse (Wu et al., 2025) into MVInverse*, a pure image-feature baseline that concatenates pretrained ResNet (He et al., 2016) and DINOv2 features and feeds them into the same probe structure. This provides an additional image-feature baseline beyond DINOv2 alone.

We use PSNR, SSIM, and LPIPS to evaluate albedo, roughness, and metallic maps. We also evaluate surface normals as an auxiliary geometric output. Following (Wu et al., 2025), normals are evaluated using mean angular error and the percentage of pixels with angular errors below 11.25° and 30°.

The evaluated backbones differ in architecture, pretraining data, model scale, and learning objective. We therefore interpret the experiments as comparisons of material recoverability from the resulting pretrained representations, rather than as a causal isolation of temporal pretraining from all other model differences. Throughout this section, recoverability refers to prediction using a supervised nonlinear probe on frozen features; it does not establish that the backbone explicitly disentangles material from other appearance factors.

4.2 Material Information in Frozen VidFM Representations

We first examine whether material-related information can be recovered from the frozen video representation itself. We then broaden the comparison to multiple VidFMs and datasets.

Direct Feature-Source Comparison Table 1 provides direct evidence that material properties can be recovered from frozen WAN features without DINOv2 fusion. On Ball, WAN-only features recover spatially varying roughness and metallicity substantially more accurately than DINOv2-only features. The fused representation performs similarly, showing that the pronounced material-discrimination advantage in this controlled setting does not depend on image-feature fusion.

Table 1: Feature-source comparison using the same probe architecture. The Ball results directly compare material recoverability from DINOv2 and WAN representations, while the RoomTour results additionally illustrate their complementarity under the evaluated setup.

Dataset	Feature source	PSNR↑		
		Albedo	Roughness	Metallic
Ball	DINOv2 only	19.67	10.82	11.15
	WAN only	**27.76**	**30.64**	29.44
	DINOv2 + WAN	27.36	30.47	**29.56**
RoomTour	DINOv2 only	20.33	20.37	25.28
	WAN only	20.49	18.40	23.37
	DINOv2 + WAN	**22.26**	**21.04**	**26.10**

This direct video-only conclusion is specific to WAN, for which the feature-source ablation is available. Results for the other VidFMs below evaluate their contribution in the fused setting and are not used to claim that every evaluated VidFM independently outperforms DINOv2.

RoomTour presents a different pattern. WAN-only features do not uniformly outperform DINOv2-only features, whereas their combination improves all three reported material properties. These

Table 2: Quantitative comparison of VidFMs and image-based baselines across the DL3DV, RoomTour, and Ball datasets. WAN demonstrates the strongest overall material awareness on both real-world and synthetic scenes. Rows shaded in blue indicate methods based on pure image features, while rows shaded in green indicate methods that use video features. Bold and underlined entries denote the best and second-best results per metric, respectively.

Method	DL3DV				RoomTour				Ball		
	Albedo PSNR↑	Roughness PSNR↑	Metallic PSNR↑	Normal mean↓	Albedo PSNR↑	Roughness PSNR↑	Metallic PSNR↑	Normal mean↓	Albedo PSNR↑	Roughness PSNR↑	Metallic PSNR↑
DINO (Oquab et al., 2023)	20.08	30.06	24.04	9.630	20.33	20.37	25.28	6.56	19.67	10.82	11.15
MVInverse* (Wu et al., 2025)	20.59	<u>30.23</u>	<u>24.35</u>	<u>9.251</u>	21.03	<u>20.52</u>	<u>25.96</u>	<u>6.16</u>	23.86	10.99	11.08
RGB↔X (Zeng et al., 2024a)	17.94	28.56	21.94	12.57	20.18	19.68	24.61	6.83	23.65	27.47	25.17
Diffusion-Renderer (Liang et al., 2025b)	19.57	28.69	22.92	11.15	20.57	19.16	24.10	7.40	28.12	30.17	30.91
OpenSora (Peng et al., 2025)	20.10	29.05	23.36	10.55	21.00	19.53	24.14	7.06	**28.17**	<u>30.98</u>	**30.27**
CogVideoX (Yang et al., 2024)	<u>20.89</u>	29.66	24.05	9.816	<u>21.86</u>	20.03	24.77	6.41	<u>28.09</u>	**31.36**	<u>29.58</u>
WAN (Wan et al., 2025)	**21.65**	**30.29**	**24.80**	**8.784**	**22.26**	**21.04**	**26.10**	**5.90**	27.36	30.47	29.56

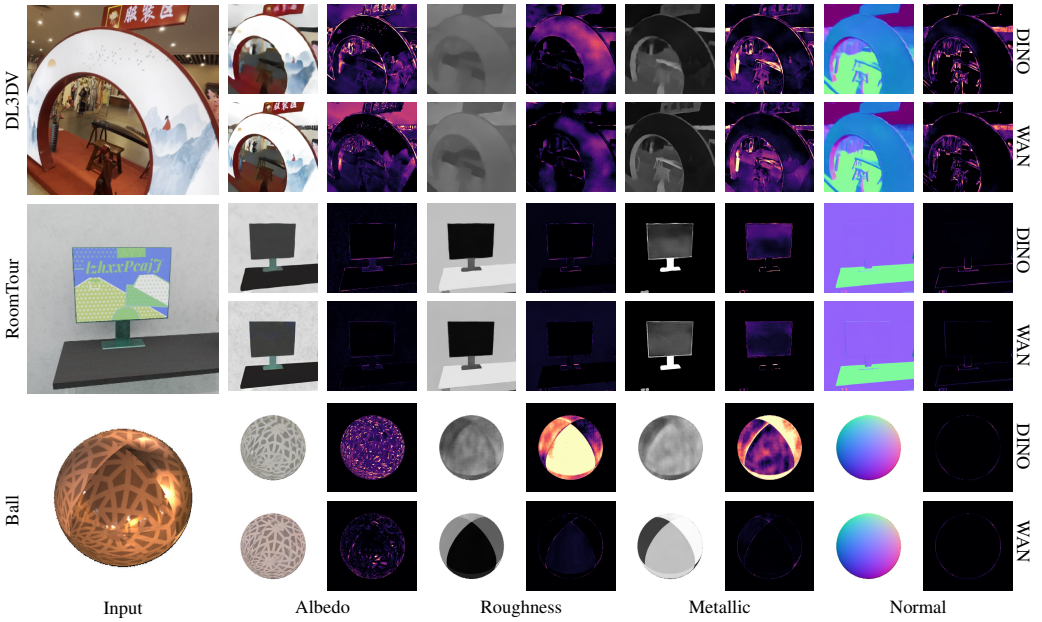


Figure 4: Qualitative comparison of WAN and DINO on the DL3DV, RoomTour, and Ball datasets. Error maps are visualized in a heat map where brighter regions indicate larger errors. Video-based models tend to yield more accurate material estimation, particularly for roughness and metallic.

results support complementarity under the evaluated setup, without establishing a specific division of spatial, semantic, or reflectance information between the two feature sources.

Results across VidFMs and Datasets We evaluate multiple VidFMs to examine whether the gains observed with WAN extend to other video backbones and how consistently these gains appear across datasets. Table 2 broadens the evaluation to three general-purpose VidFMs and an additional task-adapted video-model baseline across three datasets. The OpenSora, CogVideoX, and WAN entries use the fused representations described above and are referred to as *VidFM-augmented* variants; DiffusionRenderer is reported separately as a task-adapted feature-fusion baseline. On Ball, all three VidFM-augmented variants provide large gains over the DINOv2 baseline in roughness and metallicity recovery, demonstrating that the observation is not limited to a single video backbone in the fused setting.

The DINOv2–WAN combination achieves the best reported scores on DL3DV and RoomTour, with DL3DV scores measuring agreement with pseudo annotations. The qualitative examples in Figure 4 show corresponding improvements in material maps, particularly for roughness and metallicity. On these more complex datasets, however, the magnitude of the improvement varies across models and material properties. We therefore use the feature-source comparison in Table 1, rather than the fused main table, as the primary evidence of material recoverability without image-feature fusion.

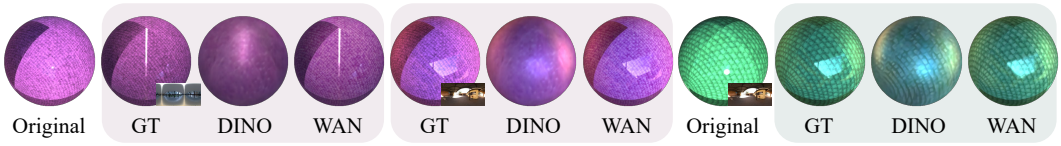


Figure 5: Qualitative relighting comparisons on the Ball dataset. Accurate material estimation tends to enable more physically plausible relighting results, whereas DINO appears to struggle in this setting, likely due to its limited ability to recover spatially varying material properties, which can lead to less realistic relighting artifacts.

The comparison with MVInverse* provides an additional control. Although MVInverse* enriches DINOv2 with pretrained ResNet features, the VidFM-augmented variants remain substantially stronger on Ball. Under the current probing setup, this shows that adding ResNet features to DINOv2 does not reproduce the gains obtained with the evaluated VidFM features. The image-based RGB↔X entry also achieves substantially higher roughness and metallicity scores than DINOv2 on Ball. Thus, the large gains over DINOv2 should not be interpreted as an advantage over image-based representations in general, or as evidence that temporal modeling alone accounts for the differences.

RGB↔X and DiffusionRenderer provide comparisons with models previously adapted to material estimation. WAN fusion achieves better scores on DL3DV and RoomTour, whereas DiffusionRenderer fusion achieves the highest metallicity PSNR on Ball. Thus, neither prior material supervision nor its absence ensures uniformly better performance in these comparisons. Because the models also differ in architecture and training data, these results do not isolate the effect of material supervision. DiffusionRenderer's results are not used to support our claim about material cues present before material-supervised backbone adaptation.

Across the evaluated VidFMs, gains in controlled material discrimination extend beyond WAN in the fused setting, whereas improvements on complex scenes depend on backbone and material property.

Relighting as a Functional Check Ball also permits a controlled relighting evaluation because its ground-truth PBR parameters and rendering process are available. As shown in Figure 5, the nearly uniform roughness and metallicity predictions obtained from DINOv2 fail to reproduce the spatially varying reflectance behavior under novel illumination. The VidFM-based predictions preserve these material regions more accurately and produce relighting results closer to the ground truth.

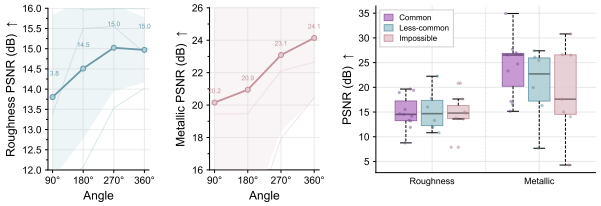
We use this experiment as a functional check on the recovered material maps: the improvements affect the rendered appearance under changed illumination, rather than only the per-pixel material metrics. Relighting is therefore treated as supporting validation of material recoverability, rather than as a separate application claim.

4.3 Factors Associated with Material Recoverability

We next examine the observation and semantic factors associated with material recovery. These experiments characterize trends in the evaluated prediction pipeline; they are not intended to causally attribute the observed effects to a single component of VidFM pretraining.

Material Composition We partition part of the DL3DV test set according to the metallic values in the pseudo ground-truth maps. Pixels with metallic values above 30 are treated as metallic and those below 30 as non-metallic. We select a *metallic-dominant split* of 20 sequences in which more than 70% of pixels are metallic, and a *non-metallic-dominant split* of 20 sequences in which more than 60% of pixels are non-metallic.

As shown in Table 3, the gains of the VidFM-augmented representation over DINOv2 are generally larger on the metallic-dominant split. This is a descriptive comparison of agreement with pseudo annotations on two scene subsets. Differences in scene content and material distributions prevent attributing the larger gains specifically to view-dependent appearance. The split is defined by metallicity and does not distinguish glossy from diffuse surfaces.



(a) Viewpoint coverage Effect. (b) Object–material compatibility groups.

Figure 6: Observation and semantic factors associated with material recoverability.

Maps	Metrics	Glossy Split			Diffuse Split		
		DINO	WAN	Diff	DINO	WAN	Diff
Albedo	PSNR ↑	19.59	21.49	+1.900	21.40	22.08	+0.679
	SSIM ↑	0.457	0.578	+0.121	0.460	0.527	+0.067
	LPIPS ↓	0.464	0.293	-0.171	0.364	0.249	-0.115
Roughness	PSNR ↑	32.11	32.80	+0.686	29.51	29.52	+0.003
	SSIM ↑	0.535	0.540	+0.005	0.569	0.571	+0.002
	LPIPS ↓	0.208	0.178	-0.030	0.195	0.162	-0.033
Metallic	PSNR ↑	23.56	24.43	+0.876	26.33	26.91	+0.585
	SSIM ↑	0.701	0.730	+0.029	0.791	0.805	+0.014
	LPIPS ↓	0.360	0.257	-0.102	0.235	0.178	-0.058
Normal	mean ↓	11.30	10.35	-0.948	8.137	7.540	-0.597
	11.25°↑	66.64	69.24	+2.600	80.18	81.79	+1.610
	30°↑	91.37	93.36	+1.990	95.94	97.16	+1.220

Table 3: Quantitative comparison of the metallic and non-metallic splits.

Viewpoint Coverage We examine viewpoint diversity using video sequences of three RoomTour objects captured along circular trajectories with uniform rotation. As shown in Figure 6a, increasing angular coverage improves roughness and metallicity recovery for the evaluated predictor on the three tested objects, with the larger change observed for metallicity. This trend is consistent with broader viewpoint coverage exposing additional view-dependent appearance variation. Since this experiment does not include a matched DINOv2 coverage curve, we use it to characterize the tested VidFM-based predictor rather than to claim that VidFMs necessarily exploit viewpoint diversity better than all image representations.

Object–Material Compatibility We additionally study whether predictions are sensitive to common object–material associations. For sofa, table, and chair instances from RoomTour, we replace the original material with seven alternatives spanning three levels of compatibility: materials that could reasonably occur on the object, materials that are uncommon but plausible, and materials that are semantically incompatible. Figure 6b shows that performance progressively decreases for less compatible materials. This association is consistent with an influence of object–material priors, although substitutions also change appearance and may differ in frequency or estimation difficulty. The experiment does not separate these explanations or localize the effect to video features, image features, or the supervised probe.

4.4 Synthetic-to-Real Transfer

We train the full model on RoomTour and apply it to real-world YouTube (AlexLuxDecor, 2026) videos without target-domain supervision. Figure 11 in the supplementary material compares the original MVInverse, MVInverse fine-tuned on RoomTour, and our VidFM-MatHead: incorporating WAN features reduces wrongly predicted metallic on surfaces that should not be metallic, also reflected in the relighting images, where the sofa appears cotton rather than leather. Figure 12 shows the relighting check of the predicted materials. These observations suggest utility beyond the supervised synthetic domain, but do not establish general target-domain accuracy, as ground-truth PBR maps are unavailable.

Downstream validation and additional ablations. VidFM-MatHead obtains the highest PSNR in the full-system comparison in the supplementary materials (Table 4), serving as downstream validation rather than an isolated test of VidFM features. Additional DL3DV ablations identify middle layers as best among the three tested, and show no uniform gain with model size (Table 5). Details are in Section 10 of the supplementary material.

5 Conclusion

We investigate material recoverability from frozen VidFM representations. WAN-only features outperform DINOv2 on controlled, spatially varying Ball materials, while their combination performs better than either alone on RoomTour. All three general-purpose VidFMs show gains in the fused Ball setting. Viewpoint coverage and object–material compatibility characterize conditions for recovery in the evaluated predictor; relighting and qualitative synthetic-to-real examples provide functional validation. These findings support reusable material-relevant information in pretrained video representations, without isolating temporal pretraining as the cause. We will release our code and datasets to support further study of material representations in video foundation models.

REFERENCES

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

AlexLuxDecor. AlexLuxDecor YouTube Channel, 2026. URL https://www.youtube.com/@alexluxdecor. Accessed: 2026-09-26.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1728–1738, 2021.

Blender Online Community. Blender. http://www.blender.org, 2018.

BlenderKit s.r.o. BlenderKit: a community-driven 3d asset library integrated into blender 3d. https://www.blenderkit.com, 2026. Accessed: 2026-05-05.

Tyler Bonnen, Stephanie Fu, Yutong Bai, Thomas O'Connell, Yoni Friedman, Nancy Kanwisher, Joshua B Tenenbaum, and Alexei A Efros. Evaluating multiview object consistency in humans and image models. *Advances in Neural Information Processing Systems*, 37:43533–43548, 2024.

Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12684–12694, 2021.

Brent Burley and Walt Disney Animation Studios. Physically-based shading at disney. In *Acm siggraph*, volume 2012, pp. 1–7. vol. 2012, 2012.

Xi Chen, Sida Peng, Dongchen Yang, Yuan Liu, Bowen Pan, Chengfei Lv, and Xiaowei Zhou. Intrinsicanything: Learning diffusion priors for inverse rendering under unknown illumination. In *European Conference on Computer Vision*, pp. 450–467. Springer, 2024.

Yue Chen, Xingyu Chen, Anpei Chen, Gerard Pons-Moll, and Yuliang Xiu. Feat2gs: Probing visual foundation models with gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6348–6361, 2025.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025.

Robert L Cook and Kenneth E. Torrance. A reflectance model for computer graphics. *ACM Transactions on Graphics (ToG)*, 1(1):7–24, 1982.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21795–21806, 2024.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Jian Gao, Chun Gu, Youtian Lin, Zhihao Li, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. Relightable 3d gaussians: Realistic point cloud relighting with brdf decomposition and ray tracing. In *European Conference on Computer Vision*, pp. 73–89. Springer, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Yibin Huang, Congying Qiu, Xiaonan Wang, Shijun Wang, and Kui Yuan. A compact convolutional neural network for surface defect inspection. *Sensors*, 20(7):1974, 2020.

Zixuan Huang, Xiang Li, Zhaoyang Lv, and James M Rehg. How much 3d do video foundation models encode? *arXiv preprint arXiv:2512.19949*, 2025.

Yingwenqi Jiang, Jiadong Tu, Yuan Liu, Xifeng Gao, Xiaoxiao Long, Wenping Wang, and Yuexin Ma. Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5322–5332, 2024.

Brian Karis and Epic Games. Real shading in unreal engine 4. *Proc. Physically Based Shading Theory Practice*, 4(3):1, 2013.

Peter Kocsis, Lukas Höllein, and Matthias Nießner. Intrinsix: High-quality pbr generation using image priors. *arXiv preprint arXiv:2504.01008*, 2025.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Georgios Kouros, Minye Wu, and Tinne Tuytelaars. Spec-gloss surfels and normal-diffuse priors for relightable glossy objects. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4417–4426, 2026.

Jingzhi Li, Zongwei Wu, Eduard Zamfir, and Radu Timofte. Recap: Better gaussian relighting with cross-environment captures. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21307–21316, 2025.

Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3205–3215, 2023.

Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Chih-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, et al. Diffusion renderer: Neural inverse and forward rendering with video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26069–26080, 2025a.

Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Chih-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, et al. Diffusionrenderer: Neural inverse and forward rendering with video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26069–26080. IEEE, 2025b.

Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, and Kui Jia. Gs-ir: 3d gaussian splatting for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21644–21653, 2024.

Valentina Lilova, Toyesh Chakravorty, Julian I Bibo, Emma Boccaletti, Brandon Li, Lívia Baxová, Cees GM Snoek, and Mohammadreza Salehi. Evaluating foundation models' 3d understanding through multi-view correspondence analysis. *arXiv preprint arXiv:2512.11574*, 2025.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert, Liang-Yan Gui, and Yu-Xiong Wang. Lexicon3d: Probing visual foundation models for complex 3d scene understanding. *Advances in Neural Information Processing Systems*, 37:76819–76847, 2024.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2630–2640, 2019.

Jacob Munkberg, Zian Wang, Ruofan Liang, Tianchang Shen, and Jon Hasselgren. Videomat: Extracting pbr materials from video diffusion models. In *Computer Graphics Forum*, volume 44, pp. e70180. Wiley Online Library, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, et al. Open-sora 2.0: Training a commercial-level video generation model in $200k. *arXiv preprint arXiv:2503.09642*, 2025.

Poly Haven. Poly haven, the public 3d asset library. `https://polyhaven.com`, 2021. Accessed: 2026-05-05.

Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21783–21794, June 2024.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, et al. Sat: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10912–10922, 2021.

Christophe Schlick. An inexpensive brdf model for physically-based rendering. In *Computer graphics forum*, volume 13, pp. 233–246. Wiley Online Library, 1994.

Bruce Smith. Geometrical shadowing of a random rough surface. *IEEE transactions on antennas and propagation*, 15(5):668–671, 1967.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in neural information processing systems*, 36: 1363–1389, 2023.

Genmo Team. Mochi 1. `https://github.com/genmoai/models`, 2024.

TS Trowbridge and Karl P Reitz. Average irregularity representation of a rough surface for ray reflection. *Journal of the optical society of America*, 65(5):531–536, 1975.

Giuseppe Vecchio and Valentin Deschaintre. Matsynth: A modern pbr materials dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22109–22118, 2024.

Bruce Walter, Stephen R Marschner, Hongsong Li, and Kenneth E Torrance. Microfacet models for refraction through rough surfaces. *Rendering techniques*, 2007:18th, 2007.

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

Qian Wang, Abdelrahman Eldesokey, Mohit Mendiratta, Fangneng Zhan, Adam Kortylewski, Christian Theobalt, and Peter Wonka. Zero-shot video semantic segmentation based on pre-trained diffusion models. *arXiv preprint arXiv:2405.16947*, 2024a.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024b.

Xiangzuo Wu, Chengwei Ren, Jun Zhou, Xiu Li, and Yuan Liu. Mvinverse: Feed-forward multi-view inverse rendering in seconds. *arXiv preprint arXiv:2512.21003*, 2025.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5987–5995. IEEE, 2017.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025a.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis L Brown II, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. In *The Fourteenth International Conference on Learning Representations*, 2025b.

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.

Keyang Ye, Qiming Hou, and Kun Zhou. 3d gaussian splatting with deferred reflection. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–10, 2024.

Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. Rgb↔x: Image decomposition and synthesis using material-and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 conference papers*, pp. 1–11, 2024a.

Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. Rgb↔x: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, SIGGRAPH '24, New York, NY, USA, 2024b. Association for Computing Machinery. ISBN 9798400705250. doi: 10.1145/3641519.3657445. URL https://doi.org/10.1145/3641519.3657445.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36:45533–45547, 2023.

Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5453–5462, 2021.

Jingsen Zhu, Fujun Luan, Yuchi Huo, Zihao Lin, Zhihua Zhong, Dianbing Xi, Rui Wang, Hujun Bao, Jiaxiang Zheng, and Rui Tang. Learning-based inverse rendering of complex indoor scenes with differentiable monte carlo raytracing. In *Siggraph asia 2022 conference papers*, pp. 1–8, 2022.

Zixin Zhu, Xuelu Feng, Dongdong Chen, Junsong Yuan, Chunming Qiao, and Gang Hua. Exploring pre-trained text-to-video diffusion models for referring video object segmentation. In *European Conference on Computer Vision*, pp. 452–469. Springer, 2024.

Yiming Zuo, Karhan Kayan, Maggie Wang, Kevin Jeon, Jia Deng, and Thomas L Griffiths. Towards foundation models for 3d vision: How close are we? In *2025 International Conference on 3D Vision (3DV)*, pp. 1285–1296. IEEE, 2025.

6 APPENDIX

7 DATASET GENERATION DETAILS

7.1 BALL DATASET GENERATION

To generate the rendering images given generated material maps, we adopt a physically-based rendering model simplified from Disney Principled BRDF (Burley & Studios, 2012) that uses materials parameterized by albedo $\mathbf{A}$, roughness $\mathbf{R}$, metallic $\mathbf{M}$, and Normal map $\mathbf{N}$ following (Li et al., 2025) when generating the ball dataset. The shading function is computed via split-sum approximation (Karis & Games, 2013) and can be written as

$$L = E_a(\mathbf{N})(1 - \mathbf{M})\mathbf{A} + E_r(\mathbf{r}, \mathbf{R}) \left[F_0\beta_1(\mathbf{N} \cdot \mathbf{v}, \mathbf{R}) + \beta_2(\mathbf{N} \cdot \mathbf{v}, \mathbf{R})\right] \tag{4}$$

Where $E_a(\mathbf{N})$ is the ambient irradiance sampled from the pre-filtered environment map along the normal direction $\mathbf{N}$, and $E_r(\mathbf{r}, \mathbf{R})$ is the specular radiance sampled from the environment map along the reflection direction $\mathbf{r} = \text{reflect}(-\mathbf{v}, \mathbf{N})$ with roughness $\mathbf{R}$ controlling the mip level, $F_0 = (1 - \mathbf{M}) \cdot 0.04 + \mathbf{A} \cdot \mathbf{M}$ is the Fresnel reflectance at normal incidence, and β_1, β_2 are precomputed BRDF scale and bias terms from the FG lookup table (Karis & Games, 2013). We show some representative images from the Ball dataset we constructed in figure 7.

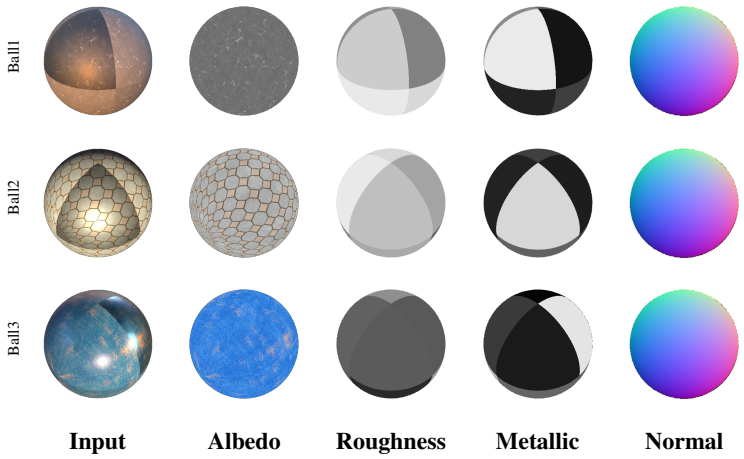


Figure 7: Visualizations of representative the Ball scenes showing RGB images alongside their corresponding ground-truth PBR maps.

7.2 ROOMTOUR DATASET GENERATION

RoomTour dataset is built from two Blender sources, BlenderKit (BlenderKit s.r.o., 2026) and Infinigen Room (Raistrick et al., 2024). We show some representative images from scenes built from BlenderKit in Figure 8 and scenes built from Infinigen Room in Figure 9.

To automatically generate Blender scenes, camera trajectories, and the corresponding synthetic RGB and material images from Infinigen, we design an automated camera path generation method and increase the scene and material diversity within its pipeline.

Camera path generation. Algorithm 1 summarizes our geometry-guided camera planning. The trajectory combines center-based visits, local camera translations for multi-view close-ups, and two-height center sweeps as shown in Figure 10. Pose validation checks room membership, obstacle clearance, visibility, and framing. Scene-name filters exclude room structures, windows, and glass objects from close-up targets. Rendering views are sampled from the resulting trajectory.

Scene diversity. We assign higher sampling weights to less frequently selected asset generators and increase the inclusion probabilities of optional furniture and decorations. These adjustments promote more diverse combinations of shelves, lamps, mirrors, and tables while respecting room-specific



Figure 8: Visualizations of representative BlenderKit scenes showing RGB images alongside their corresponding ground-truth PBR maps. We select scenes that exhibit photo-realistic rendering quality, providing accurate ground-truth material maps for fair quantitative evaluation.

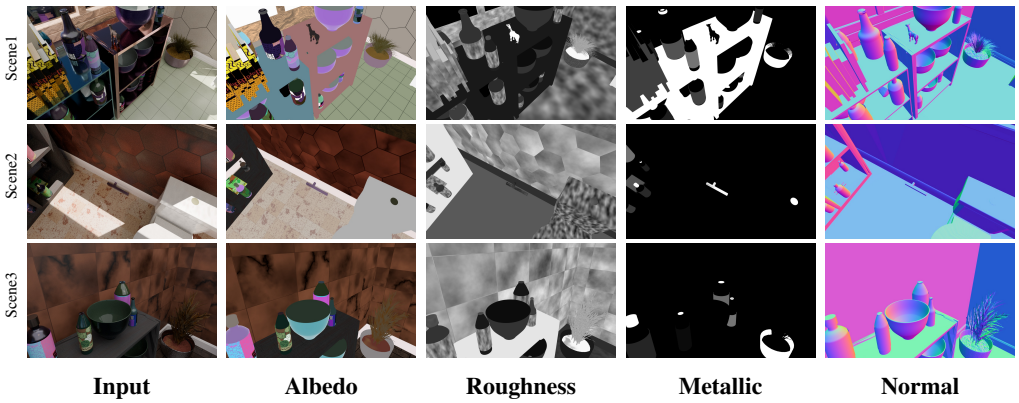


Figure 9: Visualizations of representative Infinigen scenes showing RGB images alongside their corresponding ground-truth PBR maps. We select scenes that exhibit photo-realistic rendering quality, providing accurate ground-truth material maps for fair quantitative evaluation.

layout constraints. To facilitate camera access, we also favor placements of dining and coffee tables near room boundaries. The existing count limits for common object categories remain unchanged.

Material diversity. To reduce the predominance of wood in furniture, we increase the probability of assigning metal to suitable components, such as bed frames, shelves, and wall-mounted supports. Wood, plastic, and other material options remain available, and we preserve the appearance variations produced by the existing procedural shaders. This sampling bias aims to increase the representation of reflective surfaces and their view-dependent appearance variations.

8 Implementation Details

Probe model. All probes are trained on images of resolution 294×518, using AdamW with a learning rate of 1×10^{-4} and a linear warmup over 10 epochs. Probes are trained from scratch for all datasets. All experiments are conducted on 8 NVIDIA H200 GPUs (141GB memory each). Training the WAN2.1-14B probe on DL3DV takes approximately 172 GPU hours, and inference over 600 test sequences completes in around 20 minutes. The total compute for all reported experiments, including ablations, amounts to approximately 2000 GPU hours. Additional data processing and feature extraction prior to method development required approximately 200 GPU hours beyond the reported results. For inference, extracting features from an 81-frame sequence takes approximately 7–8 seconds and requires 43.5GB of GPU memory; running the probe on each sequence takes approximately 22.4 seconds and requires 16GB of GPU memory.

Algorithm 1 Geometry-guided camera trajectory generation

Input: Scene geometry and camera field of view
Output: Camera trajectory

```
Select a furnished room and candidate objects.
Choose a center viewpoint and group targets by direction.
for each direction in angular order do
    Observe the current direction from the center.
    for each selected target in this direction do
        Propose a base close-up pose using object size and field of view.
        if the base pose passes validation then
            Append the base pose.
            Propose left, right, up, and down translations facing the target.
            Validate the translated poses and append accepted views.
        end if
    end for
    Return to the center before advancing.
end for
Append center sweeps at the original and a raised height.
Check connecting segments for obstructions and consider detours.
Set transition durations from position and viewing-direction changes.
Interpolate keyframes to form the camera trajectory.
```

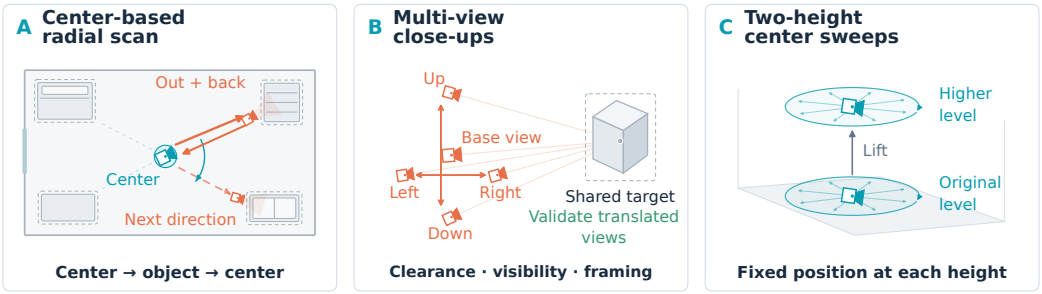


Figure 10: Camera motions: center-based visits (A), multi-view close-ups (B), and two-height center sweeps (C).

Full model. For the full model, we follow the training protocol of MVInverse (Wu et al., 2025). Input images are resized to a width of 518 with a 9:16 aspect ratio (294×518). Each training sample contains a variable number of views drawn from the same sequence, between 2 and 12, and each GPU processes at most 12 images per iteration. The DINOv2 and ResNeXt-101 encoders are frozen. The VidFM projection layer, the 36-block alternating-attention decoder and the prediction heads are trained from scratch with AdamW. The learning rate is warmed up linearly to 5×10^{-5} over the first 5% of training and then decays to 1×10^{-8} with a cosine schedule. Gradients are clipped to a norm of 1.0. The model is trained for 200 epochs of 1,000 iterations each. We supervise albedo, roughness and metallic with an L1 loss plus a gradient-matching term, and surface normals with a negative cosine loss. After 5k iterations, we add a cross-view consistency loss with weight 0.3, which compares predictions at pixel correspondences between views. Training runs on 8 NVIDIA GH200 GPUs (96GB memory each) and takes about 50 hours ($\sim$400 GPU hours). At inference, the full model processes an 81-frame 280×518 sequence in a single forward pass in 3.2 s on one NVIDIA GH200 GPU (25 FPS, 35 GB peak memory), excluding VidFM feature extraction. Adding the VidFM features increases the latency by less than 1%.

9 Full-Model Architecture

Model Structure. For downstream evaluation and synthetic-to-real transfer, we plug VidFM features into the full inverse-rendering architecture of MVInverse (Wu et al., 2025). Compared with the probe (four pairs of frame and global attention modules), MVInverse uses a much deeper alternating-attention decoder with 36 blocks that interleave frame-wise and global attention. We concatenate the

VidFM features with the DINOv2 patch tokens along the channel dimension and project them to the decoder width with a LayerNorm and a linear layer. In the original MVInverse, only the albedo and shading heads fuse multi-scale ResNeXt (Xie et al., 2017) features. We extend this to the roughness and metallic DPT heads as well, so that all material heads recover high-frequency detail, and we drop the shading head. The DINOv2 and ResNeXt encoders stay frozen, and only the projection layer, the decoder, and the prediction heads are trained. We call our full model **VidFM-MatHead**.

Cross-view consistency. Nevertheless, without an explicit cross-view constraint, the predicted material maps may still be inconsistent across views and geometrically misaligned. We further add an output-level multi-view consistency regularization to the full model. For each training sequence, we use depth maps predicted by DepthAnything (Lin et al., 2025) together with camera poses and intrinsics to compute correspondences between valid frame pairs. Given a pixel $\mathbf{u}$ in frame i, we analytically warp it to frame j using the estimated depth and camera parameters:

$$\mathbf{u}_{i\to j} = \pi\big(\mathbf{K}_j \mathbf{T}_j^{-1} \mathbf{T}_i \left(d_i(\mathbf{u}) \mathbf{K}_i^{-1} \bar{\mathbf{u}}\right)\big), \tag{5}$$

where $\bar{\mathbf{u}} = [u, v, 1]^\top$ is the homogeneous pixel coordinate, d_i, $\mathbf{K}_i$, and $\mathbf{T}_i$ denote the depth, intrinsic matrix, and camera-to-world pose of frame i, respectively. Let Ω_{ij} be the set of source pixels with valid warped locations inside the target image. Let $\mathbf{Y}$ denote one of the predicted output maps, i.e., $\mathbf{Y} \in \{\mathbf{A}, \mathbf{RM}, \mathbf{N}\}$. We penalize prediction differences across views:

$$\mathcal{L}_{\text{mv}}(\mathbf{Y}) = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \frac{1}{|\Omega_{ij}|} \sum_{\mathbf{u}\in\Omega_{ij}} \left\| \mathbf{Y}_i(\mathbf{u}) - \mathbf{Y}_j(\mathbf{u}_{i\to j}) \right\|_1, \tag{6}$$

where $\mathcal{P}$ denotes the set of valid frame pairs. Following the supervised objective in Eq. equation 3, the full-model training loss becomes

$$\mathcal{L} = \mathcal{L}_{\text{A}} + \mathcal{L}_{\text{RM}} + \mathcal{L}_{\text{N}} + \lambda_{\text{mv}} \big[\mathcal{L}_{\text{mv}}(\mathbf{A}) + \mathcal{L}_{\text{mv}}(\mathbf{RM}) + \mathcal{L}_{\text{mv}}(\mathbf{N})\big]. \tag{7}$$

10 Downstream Validation and Additional Ablations

10.1 Downstream Validation in a Feed-Forward Model

The preceding sections use frozen-feature probing to study representation content. We additionally evaluate whether these features can benefit a more expressive feed-forward material estimation system. We integrate VidFM features into the full architecture described in Section 9 and compare it with material-estimation baselines under their original checkpoints and after fine-tuning on our datasets. We also include DiffusionRenderer (Liang et al., 2025a), which uses a video-model backbone for inverse and forward rendering. Its original checkpoint already includes material-supervised adaptation; fine-tuning denotes additional training on our datasets. This comparison evaluates complete material-estimation systems, whereas Table 2 compares frozen features decoded by the probing architecture.

Table 4: Comparison of feed-forward material estimation methods on the Ball, DL3DV, and RoomTour datasets. We compare against baselines both under their original checkpoints and after fine-tuning on our dataset. O: original checkpoint; FT: fine-tuned on our dataset; Our model, VidFM-MatHead, consistently outperforms all baselines across all three datasets.

Methods	PSNR↑		
	Albedo	Roughness	Metallic
Ball			
Diffusion-Renderer (O)	11.60	9.40	8.49
Diffusion-Renderer (FT)	23.77	22.32	20.20
MVInverse (O)	15.04	11.34	9.24
MVInverse (FT)	23.64	30.80	26.64
VidFM-MatHead (Ours)	**26.65**	**34.88**	**32.39**
DL3DV			
Diffusion-Renderer (O)	14.03	12.71	12.83
Diffusion-Renderer (FT)	20.46	28.53	23.31
MVInverse (FT)	22.11	30.43	25.07
VidFM-MatHead (Ours)	**22.70**	**30.89**	**25.57**
RoomTour			
Diffusion-Renderer (O)	17.05	11.73	12.26
Diffusion-Renderer (FT)	20.98	18.58	19.46
MVInverse (O)	13.88	11.57	12.48
MVInverse (FT)	24.97	23.23	30.53
VidFM-MatHead (Ours)	**25.49**	**23.34**	**30.88**

As shown in Table 4, VidFM-MatHead achieves the highest reported PSNR values in this table across the three datasets. These results demonstrate the performance of the combined full system; the DL3DV results again measure agreement with pseudo annotations.

Because the full systems differ in several architectural and training components, this table is not used to isolate the contribution of VidFM features. The feature-source comparison in Table 1 provides the separate evidence for recoverability and complementarity in the probing setup.

10.2 Downstream Validation in Synthetic-to-Real Transfer

We evaluate synthetic-to-real transfer on 10 real-world YouTube videos (AlexLuxDecor, 2026), whose data distribution differs from that of DL3DV. We apply the full model trained on the synthetic RoomTour dataset directly to these videos and assess both material estimation and relighting results.

Material Estimation: Figure 11 compares our predictions with those of MVInverse using its original checkpoint (O) and after fine-tuning on our dataset (FT). Both MVInverse (O) and MVInverse (FT) incorrectly predict metallic values for non-metallic surfaces across these scenes. In rows 1 and 4, fabric surfaces on pillows and sofas are misclassified as metallic. In rows 2 and 3, spurious metallic regions appear on wooden tabletops, wooden chairs, and dark leather chairs, as well as in areas outside the windows. Similar errors occur on walls and tabletops in rows 5 and 6. VidFM-MatHead correctly identifies these non-metallic surfaces while suppressing errors caused by surface patterns and illumination. This improves material estimation throughout the scene, particularly the distinction between metallic and non-metallic surfaces.

Relighting: Figure 12 presents relighting results obtained from the predicted materials. We relight each frame with a single point light using deferred physically based shading. For each valid pixel, we unproject the depth estimated by DA3 (Lin et al., 2025) into a world-space point and compute shading using the albedo, metallic, roughness, and normal maps predicted by the method being evaluated.

Specular reflection follows the Cook–Torrance microfacet model (Cook & Torrance, 1982), with a GGX normal distribution (Walter et al., 2007; Trowbridge & Reitz, 1975), a Smith–Schlick geometry term (Smith, 1967), and Schlick's Fresnel approximation (Schlick, 1994). Diffuse reflection is Lambertian. We use a standard metallic workflow with inverse-square light attenuation and a small image-derived ambient term. The shading model accounts for direct illumination without shadows or interreflections. All methods share the same geometry, light trajectory, and light intensity, so differences in the relit images arise solely from the predicted maps.

Using a relighting ball, we render the materials predicted by all three methods under lighting directions, with the results labeled *Relighting1* and *Relighting2*, respectively. Both MVInverse (O) and MVInverse (FT) predict nonzero metallic values for the sofa, producing strong metallic reflections on the seat and back cushions. VidFM-MatHead correctly predicts zero metallic values for the sofa, yielding weaker reflections consistent with the reflectance characteristics of fabric.

Across these real-world videos, the model using video-based features produces material estimates and relighting results that align more closely with human perception than those obtained using image-based DINO features. These results show that large-scale video pretraining across diverse synthetic and real-world domains yields representations that substantially reduce the domain gap in material estimation, enabling models trained solely on synthetic data to generalize effectively to real-world scenes.

10.3 Additional Ablations

The remaining ablations examine feature layer, model scale, and the auxiliary consistency regularization.

Table 5: Ablation on layer and model size.

	Methods	PSNR↑		
		Albedo	Roughness	Metallic
	DL3DV			
Layers	l-10	21.00	30.25	24.49
	l-20	21.65	30.29	24.80
	l-24	20.99	30.24	24.48
	DL3DV			
Model Size	WAN 1.3B	21.49	30.25	24.70
	WAN 14B	21.65	30.29	24.80
	CogVideoX-2B	20.89	29.66	24.05
	CogVideoX-5B	20.85	29.63	23.97

Feature layer. We ablate the VidFM feature layer l on DL3DV. Among the three evaluated layers, layer 20 obtains the highest reported PSNR for all three material maps against the pseudo annotations. This comparison identifies a useful extraction layer for the evaluated setup, without establishing why its features perform better.

Model scale. We compare WAN2.1-1.3B and CogVideoX-2B with their larger counterparts, WAN2.1-14B and CogVideoX-5B, under the same probing configurations. WAN2.1-14B obtains slightly higher reported scores than WAN2.1-1.3B, whereas CogVideoX-5B obtains slightly lower scores than CogVideoX-2B. These DL3DV results show no uniform improvement with parameter count across the evaluated pairs; they do not establish a general scaling trend.

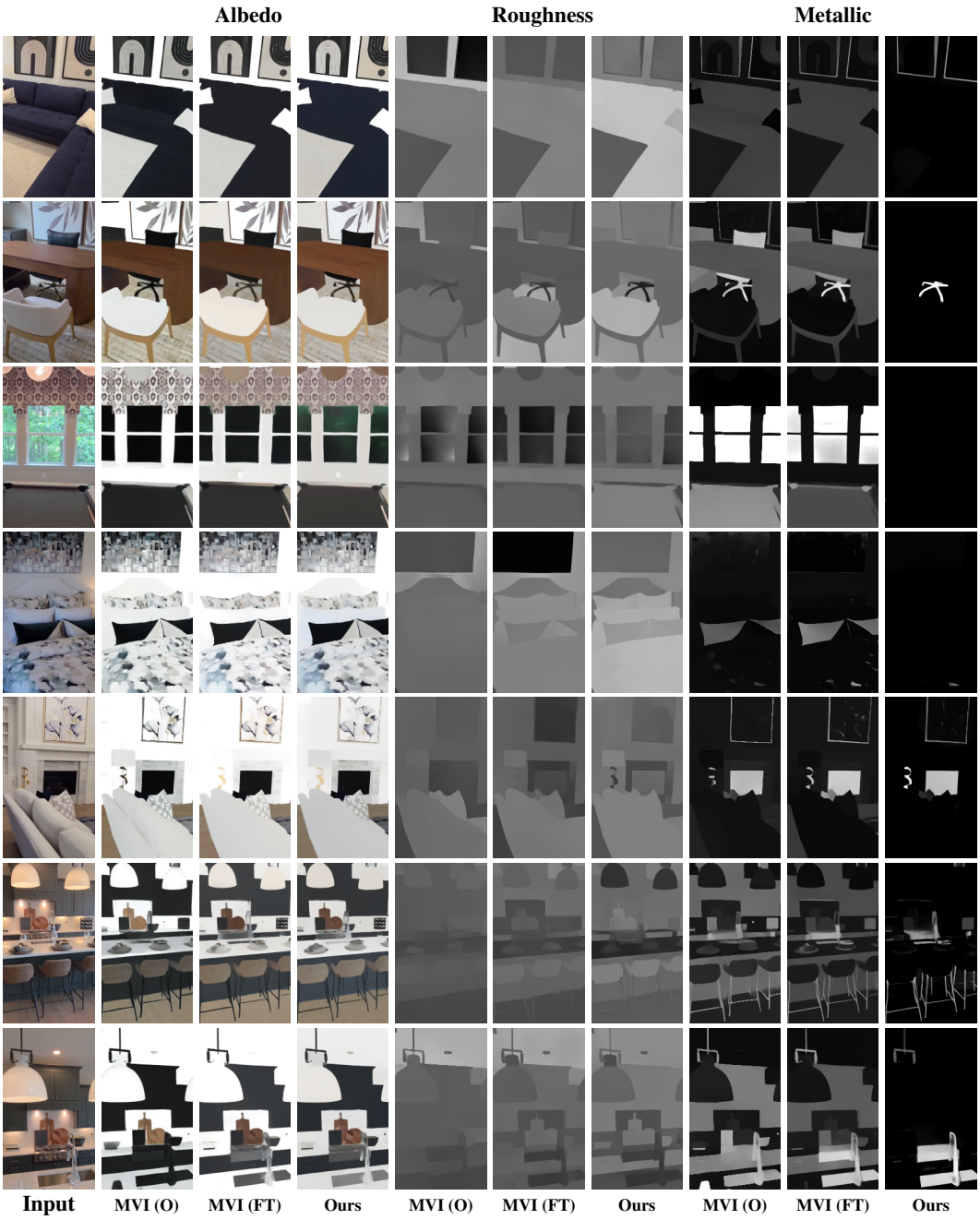


Figure 11: Domain adaptation results on real-world YouTube videos, comparing the original MVInverse, abbreviated MVI (O), MVInverse fine-tuned on our dataset, abbreviated MVI (FT), and our full model VidFM-MatHead. MVI (O) and MVI (FT) misclassify non-metallic materials such as fabric, wood, and leather as metallic, whereas our VidFM-MatHead generalizes well to these unseen, in-the-wild scenes and produces more accurate metallic material estimation.

11 More Results

11.1 More Results of VidFMs Comparisons with Probe Model

We show qualitative comparisons of the three general-purpose VidFM fusion variants on DL3DV in Figure 13, RoomTour in Figure 14, and Ball in Figure 15. These entries combine video features with



Figure 12: Relighting comparison under lighting direction, comparing the original MVInverse, abbreviated MVI (O), MVInverse fine-tuned on our dataset, abbreviated MVI (FT), and our full model VidFM-MatHead. MVI (O) and MVI (FT) predict nonzero metallic values for the sofa, resulting in unrealistic metallic reflections, while our VidFM-MatHead produces reflections consistent with fabric.

DINOv2 and are distinct from the task-adapted DiffusionRenderer baseline in Table 2. WAN fusion leads quantitatively on DL3DV and RoomTour, while the best Ball scores vary by material property and backbone.

All three general-purpose VidFM fusion variants produce clearer material boundaries and more accurate roughness and metallicity maps than DINOv2 alone on Ball. These comparisons show that the gains in the fused setting extend beyond WAN. Evidence of recoverability from video features alone comes from the separate WAN-only ablation in Table 1; the fused comparisons do not isolate temporal modeling as the cause of the gains.

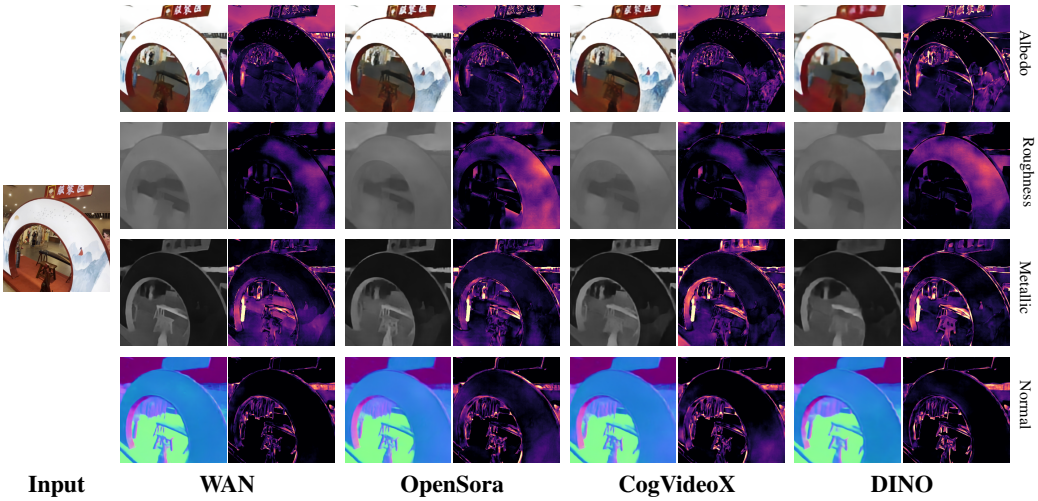


Figure 13: Qualitative comparison of material estimation on the DL3DV dataset across VidFMs and image-based baselines.

11.2 More Relighting Comparisons with Probe Model

Table 6: Quantitative relighting comparisons on the Ball dataset across 8 environment maps. VidFMs consistently produce strong relighting results, while DINO fails to capture spatially varying material properties.

Methods	PSNR↑								
	bridge	courtyard	fireplace	forest	interior	night	snow	sunset	Avg.
DINO (Oquab et al., 2023)	21.20	17.52	22.29	21.35	17.66	24.77	17.93	20.11	20.35
WAN (Wan et al., 2025)	29.40	26.26	31.14	29.65	25.87	32.87	25.47	28.23	28.61
CogVideoX (Yang et al., 2024)	**29.90**	26.44	31.31	**30.03**	26.25	33.17	25.70	28.50	28.91
OpenSora (Peng et al., 2025)	29.81	**26.57**	**31.39**	29.99	**26.30**	**33.21**	**25.75**	**28.56**	**28.95**

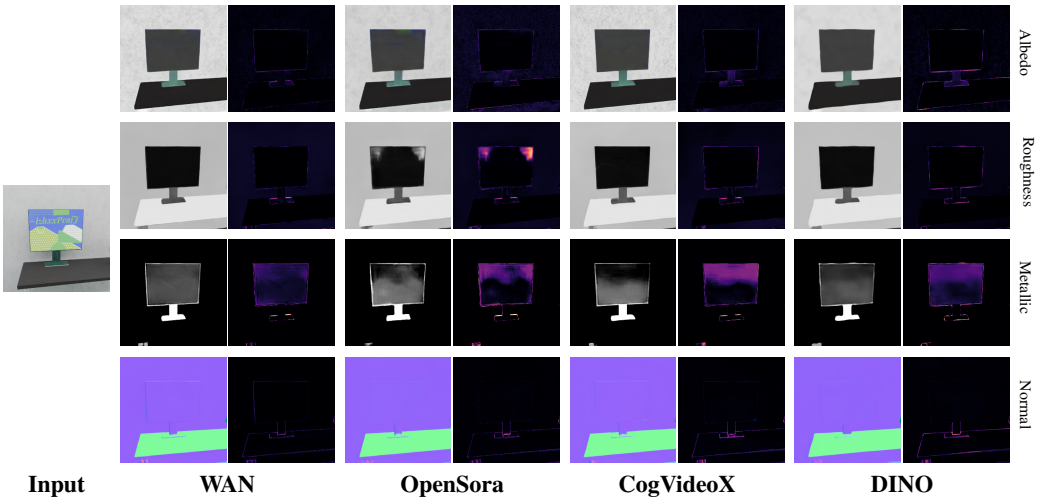


Figure 14: Qualitative comparison of material estimation on the RoomTour dataset across VidFMs and image-based baselines.

One of the most compelling applications of accurate material estimation is object relighting, and relighting quality can serve as a useful proxy for how well the underlying material properties have been recovered. We therefore also examine the VidFMs' ability through relighting experiments, using the material maps estimated by the probe model. We relight 10 balls from the test sequences under 8 novel environment maps and report quantitative results in Table 6, with additional qualitative comparisons shown in Figure 16. The results suggest that accurate roughness and metallic estimation tends to be important for physically plausible relighting. Video-based models, which recover spatially varying material boundaries and per-region material values, tend to produce relighting results that better reflect the distinct appearance changes induced by different roughness and metallic configurations under novel illumination. In contrast, DINO's more spatially uniform roughness and metallic predictions appear to collapse the relighting results toward a single appearance, and often struggle to reproduce the visually distinct relighting effects that different material properties would be expected to exhibit. This suggests that precise material estimation may not merely be a quantitative improvement, but could also be an important factor for downstream applications such as relighting to produce perceptually meaningful results.

11.3 More comparisons on Glossy and Diffuse Splits with Probe Model

In Figure 17, we present additional qualitative comparisons between WAN and DINO across multiple views on the glossy and diffuse splits of the DL3DV dataset, corresponding to the quantitative results reported in Table 3 in the main paper. WAN consistently outperforms DINO on both splits, with more obvious improvements on the glossy split, as evidenced by the error maps. This suggests that glossy materials, whose appearance varies more dramatically with viewpoint, benefit more from the view-dependent temporal cues captured by video foundation models.

11.4 More Visual Comparisons with Full Model

In this section, in Figure 19, Figure 18 and Figure 20, we present qualitative comparisons for our full model structure, VidFM-MatHead, against several baselines: DiffusionRenderer with its original checkpoint, DiffusionRenderer fine-tuned on our dataset, MVInverse with its original checkpoint, and MVInverse fine-tuned on our dataset with our proposed consistency loss. These results correspond to and visualize the quantitative comparisons reported in Table 4. When equipped with a stronger probe structure, the purely image-based DINO features are able to recover a clear material boundary on the Ball dataset. However, a noticeable gap still remains compared to our full structure incorporating VidFM features. The other baselines, when evaluated directly with their original checkpoints, exhibit a clear domain gap on our dataset. Even after fine-tuning separately on each of our three datasets, DL3DV, RoomTour, and Ball, their performance still falls short of our method. This indicates that

our material estimation model, which incorporates VidFM features, can improve material estimation performance.

12 LIMITATIONS

We evaluate general-purpose VidFMs based on their publicly available pretrained checkpoints and include DiffusionRenderer as a material-supervised, task-adapted video-model baseline. Freezing these checkpoints during probe training does not equalize their prior supervision. The models we compare differ across multiple dimensions simultaneously, including architecture, training data, and optimization strategy, making it difficult to attribute observed performance differences to any single factor. Ideally, one would compare models that differ only in one variable at a time (e.g., data scale), but to our knowledge, no such controlled model series exists in the open-source community.

Furthermore, the scarcity of large-scale synthetic datasets with ground-truth PBR maps (albedo, roughness, and metallic) paired with corresponding RGB images poses a significant challenge for feed-forward material estimation. This data bottleneck directly impacts performance on the complex datasets, such as RoomTour, in our experiments. Existing datasets vary significantly in their material definitions and rendering pipelines, making joint training across datasets non-trivial. How to effectively train feed-forward material estimation models under such data constraints remains an open and interesting research direction.

13 BROADER IMPACTS

This work advances the understanding of material awareness in video foundation models, with potential benefits for applications in augmented reality, visual effects, and scene reconstruction. By open-sourcing our code and benchmark datasets, we also hope to lower the barrier for the research community to study physically-based material understanding at scale.

However, accurate material estimation and relighting capabilities could be misused to generate photorealistic synthetic media, potentially facilitating disinformation or Deepfake generation. To mitigate these risks, we advocate for the following measures: (1) responsible disclosure and gated release of high-fidelity relighting models to prevent immediate misuse; (2) continued investment in synthetic media detection methods that can identify manipulated lighting and material conditions; and (3) community-level monitoring mechanisms to track potential misuse of material estimation and relighting technologies as they become more accessible.

We note that our work is primarily a diagnostic and benchmarking study rather than a deployment-ready system, which limits the immediacy of potential harms. Nevertheless, we believe it is important to proactively acknowledge these risks as the underlying video foundation models continue to improve rapidly.

14 THE USE OF LARGE LANGUAGE MODELS (LLMS)

In this work, we used generative AI tools to assist with translation. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results. Providing critical ingredients for proving mathematical claims and assisting in the writing of proofs are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

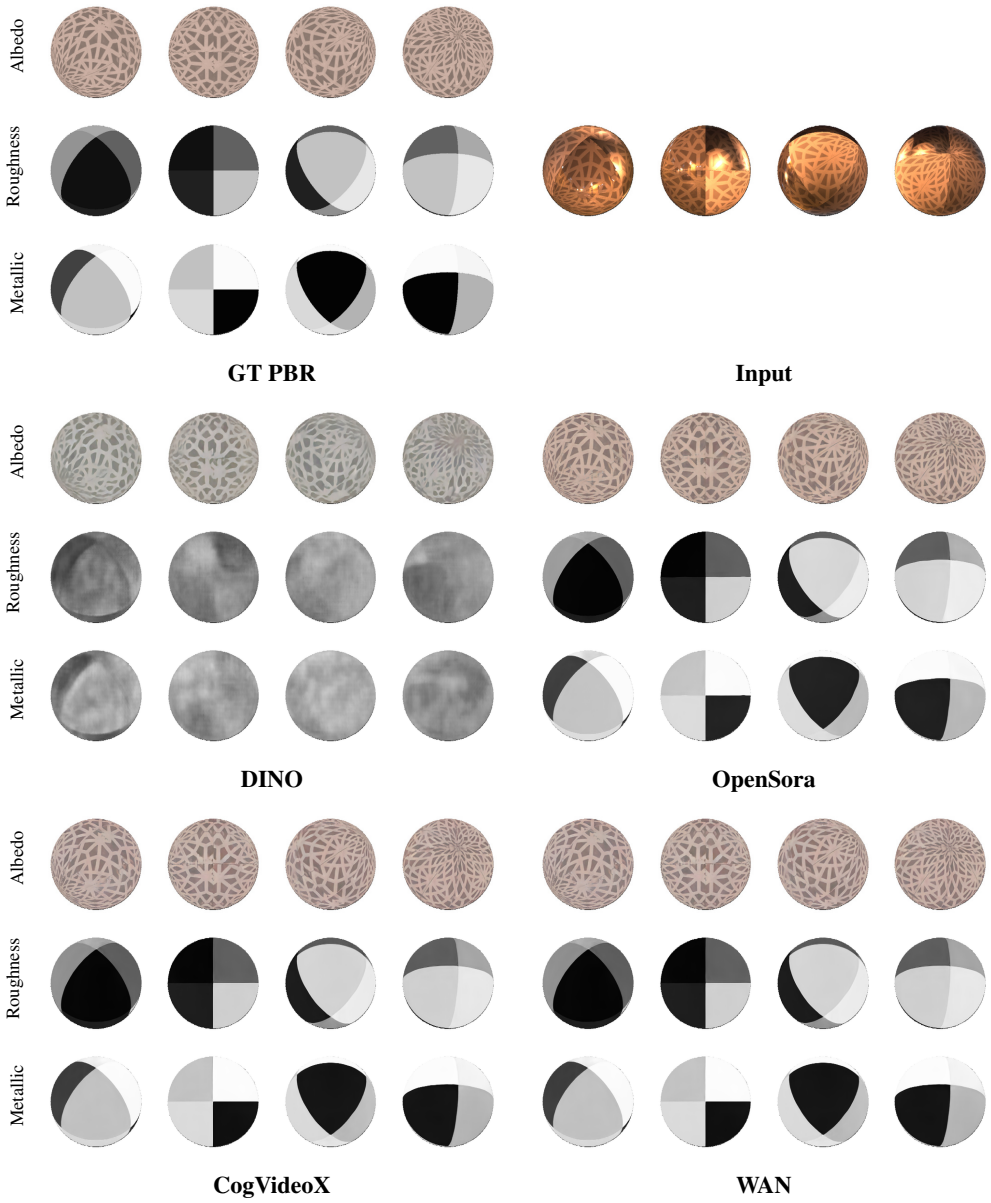


Figure 15: Qualitative comparison of material estimation on the Ball dataset. VidFMs tend to produce results that closely approximate the ground truth, whereas the image-based model DINO appears to struggle considerably in this setting.

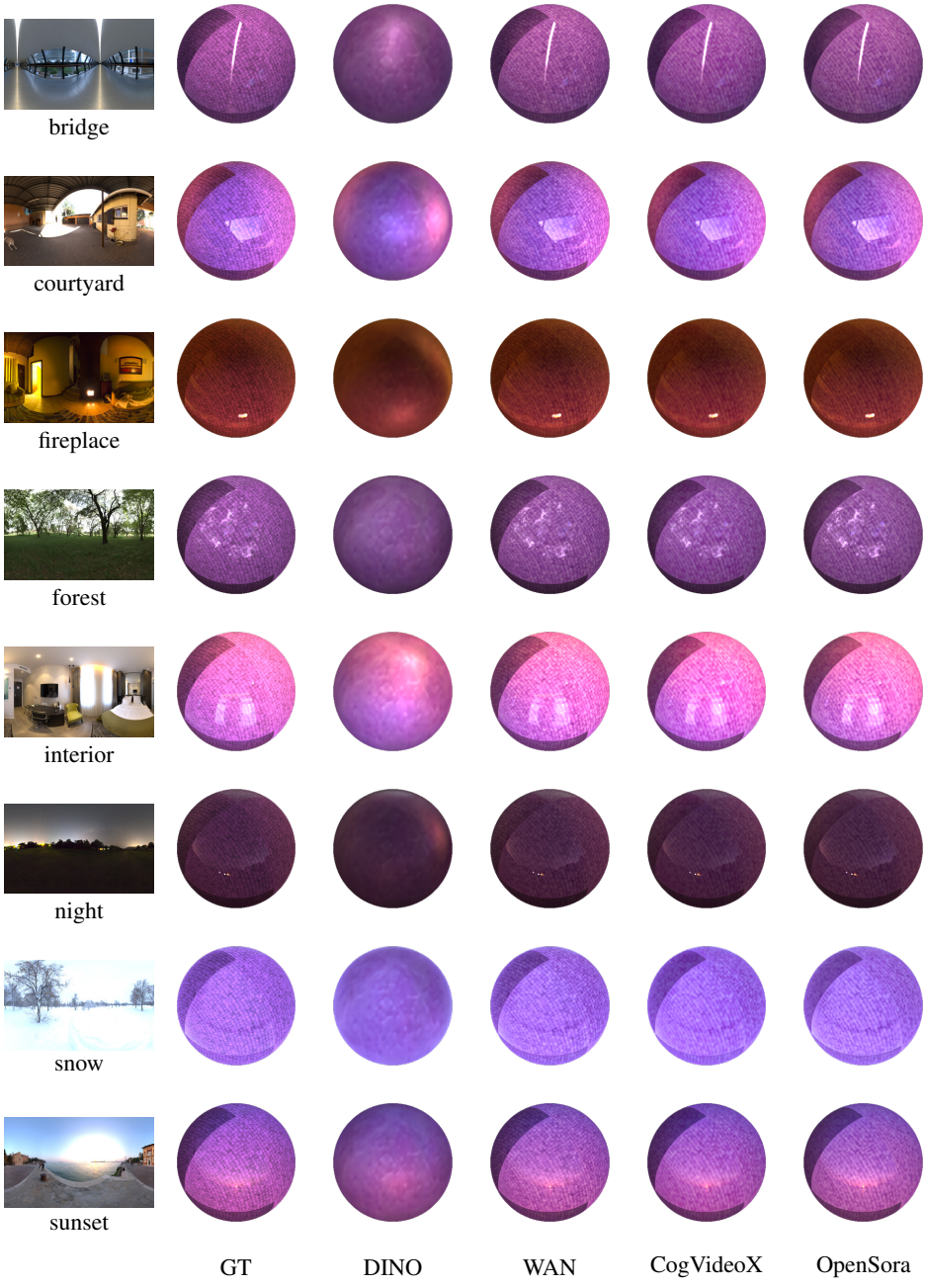


Figure 16: Qualitative relighting comparisons on the Ball dataset under 8 environment maps. VidFMs consistently outperform DINO, producing more physically plausible relighting results.

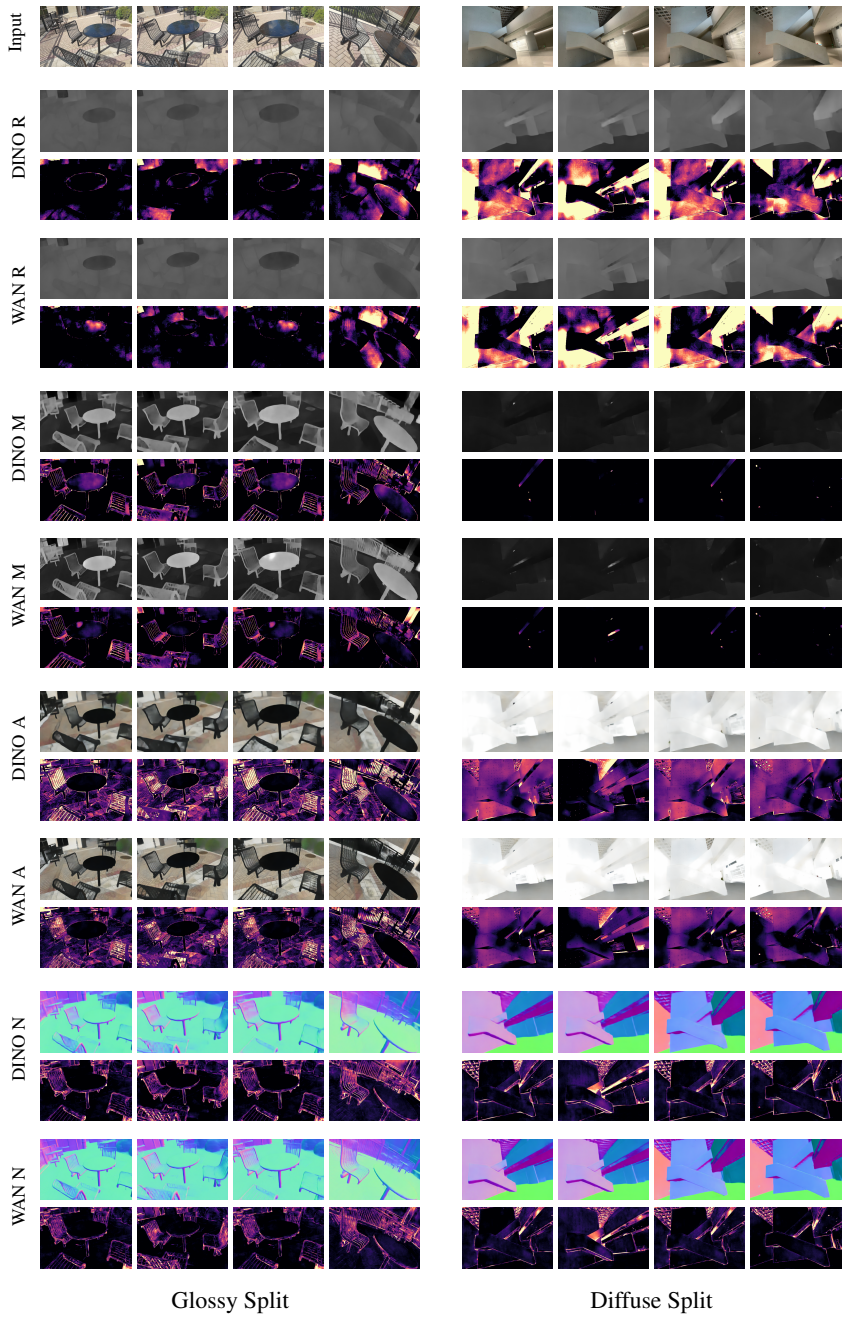


Figure 17: Qualitative comparisons on the metallic and diffuse splits. VidFM WAN yields more obvious improvements on the glossy split.

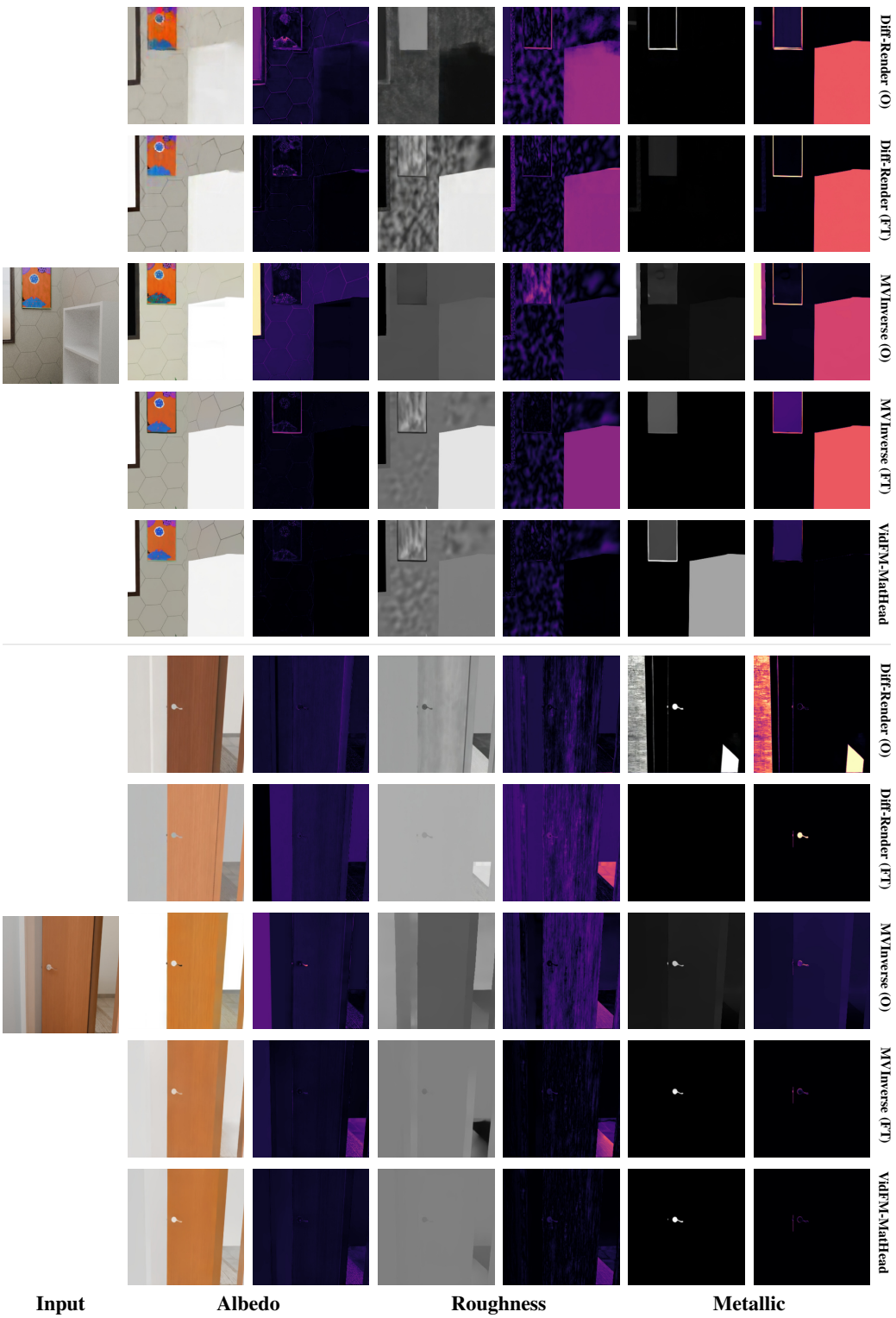


Figure 18: Qualitative comparison of material predictions across two scenes, alongside per-pixel difference maps against the ground truth. We compare the original Diffusion-Renderer (abbreviated Diff-Render, O), its version fine-tuned on our RoomTour dataset (FT), the original MVInverse (O), MVInverse fine-tuned on our dataset (FT), and our full model, VidFM-MatHead.Our method consistently yields more accurate albedo, roughness, and metallic estimates, as reflected by the lower error magnitude in the corresponding diff maps.

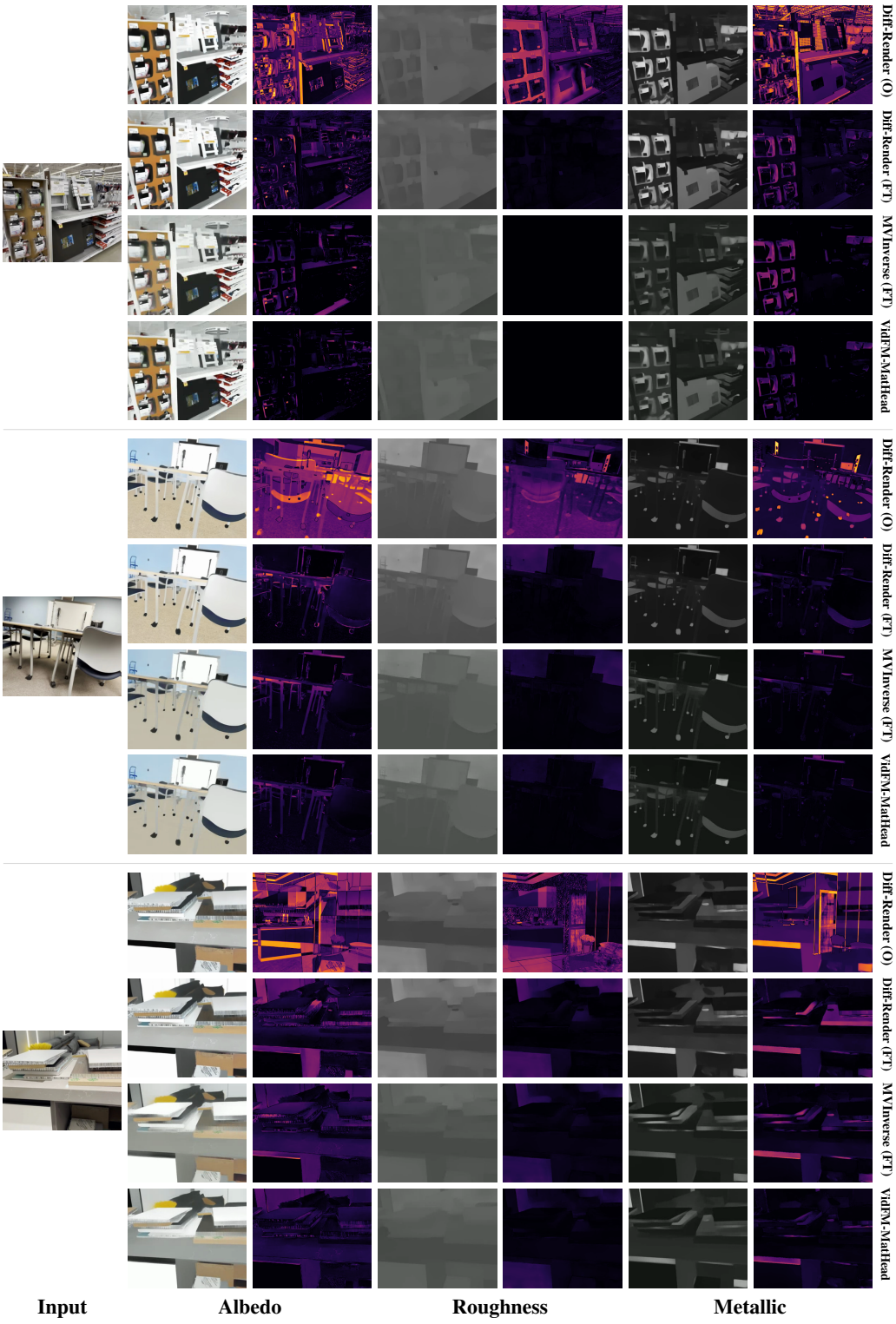


Figure 19: Qualitative comparison of material predictions across three selected DL3DV scenes. We compare the original Diffusion-Renderer (abbreviated Diff-Render, O), its version fine-tuned on our dataset (FT), the original MVInverse (O), MVInverse fine-tuned on our dataset (FT), and our full model, VidFM-MatHead. Each method is shown with albedo, roughness, and metallic predictions together with per-pixel difference maps against the ground truth.



Figure 20: Qualitative comparison of material predictions across two Ball scenes, alongside per-pixel difference maps against the ground truth. We compare the original Diffusion-Renderer (abbreviated Diff-Render, O), its version fine-tuned on our dataset (FT), the original MVInverse (O), MVInverse fine-tuned on our dataset (FT), and our full model, VidFM-MatHead. Our method consistently yields more accurate albedo, roughness, and metallic estimates, as reflected by the lower error magnitude in the corresponding diff maps.