
Generative Perception of Shape and Material from Differential Motion

Xinran Nicole Han
Harvard University
xinranhan@g.harvard.edu

Ko Nishino
Kyoto University
kon@i.kyoto-u.ac.jp

Todd Zickler
Harvard University
zickler@seas.harvard.edu

Abstract

Perceiving the shape and material of an object from a single image is inherently ambiguous, especially when lighting is unknown and unconstrained. Despite this, humans can often disentangle shape and material, and when they are uncertain, they often move their head slightly or rotate the object to help resolve the ambiguities. Inspired by this behavior, we introduce a novel conditional denoising-diffusion model that generates samples of shape-and-material maps from a short video of an object undergoing differential motions. Our parameter-efficient architecture allows training directly in pixel-space, and it generates many disentangled attributes of an object simultaneously. Trained on a modest number of synthetic object-motion videos with supervision on shape and material, the model exhibits compelling emergent behavior: For static observations, it produces diverse, multimodal predictions of plausible shape-and-material maps that capture the inherent ambiguities; and when objects move, the distributions converge to more accurate explanations. The model also produces high-quality shape-and-material estimates for less ambiguous, real-world objects. By moving beyond single-view to continuous motion observations, and by using generative perception to capture visual ambiguities, our work suggests ways to improve visual reasoning in physically-embodied systems.¹

1 Introduction

A single image of an object is inherently ambiguous. The image can be mimicked by a planar surface with paint, for example, or a curved mirror under suitable lighting. There are many possibilities, and they arise from the intertwined roles of shape and material in image formation. We argue that vision systems should reason about shape and material jointly, and that they should embrace and model their ambiguities, so that additional cues like motion or touch can be triggered for disambiguation. This will help embodied AI systems deal with very practical problems, such as distinguishing a wet floor from a glossy one, or distinguishing a painted visual cliff from an actual cliff.

Humans naturally resolve much of the uncertainty through motion. Imagine moving a shiny textured object under natural lighting: glossy highlights sweep across its surface, immediately revealing its material and geometric structure. This interplay of material, surface, and motion provides far more insight than a single static observation. Just as young children explore their environment by manipulating objects, computational vision systems can benefit from observing differential object motion—even across a few video frames. These observations motivate a perception framework that (i) explicitly models shape-material ambiguity rather than collapsing to a single estimate, and (ii) leverages object motion as a natural cue for reducing ambiguity when motion is available.

In contrast to these principles, recent learning-based approaches in computer vision typically focus on inferring either shape or material, but not both, and they are often optimized to predict a single

¹Project page: <https://xrhan.github.io/diffmotion/>

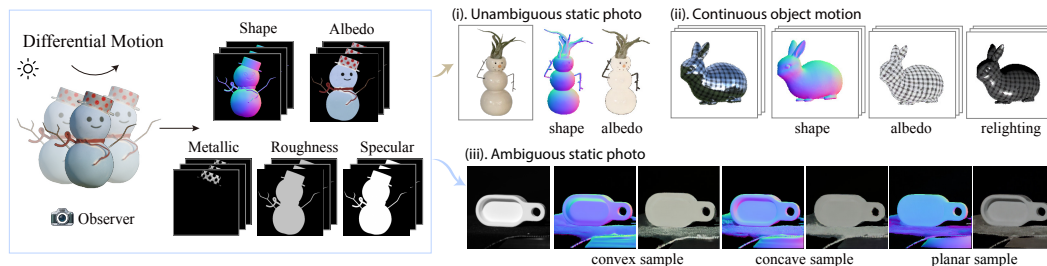


Figure 1: *Left*: We train our generative perception model to *jointly* infer shape and materials using synthetic three-frame videos of objects undergoing differential motions. *Right*: Our model (i) generalizes to captured real-world photos; (ii) leverages continuous observations to disentangle complex objects, and (iii) shows an emergent ability to provide multiple hypotheses for ambiguous static images, such as convex, concave and planar ‘postcard’ explanations.

“best” output even for tasks that are fundamentally ambiguous. Examples include the estimation of monocular shape (e.g., depth [33, 59] or normals [43, 60, 8]) and intrinsic texture or color [11]. In addition to focusing on a single attribute, such models tend to ignore the ambiguity inherent in the visual world, treating uncertainty as noise to be suppressed rather than something to be modeled.

Instead, we advocate a *generative perception* approach, where a vision system generates diverse samples of shapes and materials that explain the input observation(s). When the input is a single image, the output samples should reflect the full range of plausible interpretations, expressing the fundamental ambiguities of the scene. Enyo and Nishino [18], Kocsis et al. [36], and Han et al. [23] have previously explored this idea for single images and particular attributes, demonstrating generative perception, respectively, of illumination and reflectance; materials; and shape. Here, we argue that generative perception should not be limited to a single view, or to a single object attribute. It should incorporate moving observations when available, and it should jointly disentangle all of shape, texture, and reflectance.

To realize this goal, we introduce a conditional video diffusion model for generative perception that jointly infers shape and material from both static and moving objects. Our denoising diffusion architecture, called **U-ViT3D-Mixer**, operates directly in 256×256 pixel space and uses spatio-temporal attention and cross-channel mixing to simultaneously reason about surface normals, diffuse texture and reflectance. At coarse scales it employs global attention to capture object-scale constraints, and at fine scales it employs parameter-efficient neighborhood attention to capture local space-time constraints on shape and material.

We train our model from scratch on short synthetic videos of moving and static objects, and we find that it generalizes to captured images and videos while exhibiting several desirable perceptual capabilities: (i) it achieves competitive results on existing static-image shape and material benchmarks; (ii) it exhibits an emergent ability to generate plausible multimodal samples when the input is ambiguous, such as the classic convex/concave ambiguity or the trivial postcard solution (see Figure 1); and crucially, (iii) it makes effective use of differential object motion to resolve perceptual ambiguities, improving prediction accuracy when additional motion frames are available.

Generative perception is a framework for understanding how vision systems can operate under inherent ambiguity. Our work advances this framework by extending from static images to motion videos, and by estimating shape and material together. We hope this opens new directions for cognitive modeling, and for embodied AI systems that aim to emulate human perception.

2 Related Work

We study a novel problem: generative disentanglement of object shape and material from short videos. This has not been directly addressed in prior work, but it relates to several research threads.

Intrinsic image decomposition and inverse rendering. Intrinsic image decomposition aims to separate an image into texture (albedo), diffuse shading, and sometimes additional reflectance properties such as metallic-ness and roughness, all without inferring shape [7, 36, 11]. Inverse

rendering extends this goal by recovering the full scene properties—including shape, materials, and illumination—that are necessary for physically accurate re-rendering [41, 4, 42, 47, 61]. Our work differs from these by estimating shape and materials without explicit illumination, thereby focusing on all of the scene attributes that are intrinsic to an object.

Recently, Zeng et al. [61] introduced RGB-X, which finetunes latent diffusion models for single-image prediction of surface normals, material properties, and illumination, while also supporting text-conditioned synthesis. The model differs from ours by outputting each attribute independently via a prompt selector, without modeling their interactions. Similarly, DiffusionRenderer [40] finetunes Stable Video Diffusion [9] for physical attributes prediction given video observations. The model conditions on a domain embedding (e.g. "normals", "roughness") to infer one target attribute at a time. In contrast, our model integrates shape and material reasoning within a single backbone, allowing their representations to inform each other.

Diffusion models for conditional prediction. Generative modeling with diffusion models has gained increasing recent attention [28, 54, 16, 49]. Many recent works have leveraged priors from pre-trained latent diffusion models for dense prediction tasks such as estimating depth [33, 26], normals [43, 19, 46, 60, 8], or materials [36, 61]. These methods often suppress output diversity by averaging multiple samples [33, 36] or regularizing variance [60], overlooking the inherent ambiguity in these tasks and leading to over-smoothed predictions that fail to capture multimodal outcomes. In contrast, recent *generative perception* models embrace ambiguity by using diffusion models to produce diverse samples of illumination and reflectance [18] or of shape [23] from a single image. Here we extend this idea to consider motion and to infer shape and materials together.

Shape from differential motion. Prior mathematical analyses have used differential equations to study the reconstruction of shape from small motions under idealized conditions, such as directional lighting and Lambertian or mirror-like materials [44, 5, 57]. Belhumeur et al. [6] show that for Lambertian shapes with unknown lighting there are continuous stretches and tilts (the generalized bas-relief ambiguity) that cannot be resolved by small motions. When lighting and reflectance (BRDF) are both unknown, Chandraker et al. [12] show that three observations of differential motions around three distinct axes are needed to correctly reconstruct the shape along a characteristic surface curve. Thus, from a mathematical standpoint, perceptual ambiguity persists even when observing differential motion, which highlights the need for generative models that can represent the distribution of possibilities. A learning-based approach like ours also has the benefit of applying to a broader range of lights and materials that would be very difficult to characterize analytically.

3 Methodology

We consider a rigid object undergoing small motion with respect to a stationary viewer and lighting environment. The observer is represented by an orthographic camera along the z -axis in world coordinates. Given a sequence of temporally continuous RGB image observations $\mathbf{c} \in \mathbb{R}^{F \times H \times W \times 3}$ with F frames, we aim to infer time-varying (F -frame) *shape-and-material* fields $\mathbf{X} = \{\mathbf{n}, \mathbf{M} = \{\rho_d, r, \gamma, \rho_s\}\}$ from $\mathbf{c}$. We use bold symbols to denote variables for multiples frames. Figure 1 depicts our overall approach.

3.1 Shape and Material Representations

We represent shape by the viewer-centric surface normal field $\mathbf{n} \in \mathbb{R}^{F \times H \times W \times 3}$. The normal vector at the backprojection of each image point (u, v) in a single frame has unit length and is denoted by $n(u, v)$. We represent material attributes $\mathbf{M}$ using the point-wise surface albedo (i.e., diffuse color/texture) $\rho_d(u, v)$, roughness $r(u, v)$, metallic-ness $\gamma(u, v)$ and specularity $\rho_s(u, v)$. Each parameter is a time and space varying map with the same dimension $F \times H \times W$ that has either three channels (ρ_d) or one channel (r, γ, ρ_s) and values in the range $[0, 1]$. These parameters are chosen from a subset of those of an analytical reflectance model—the Disney principled BSDF model [10]. Together they determine the bidirectional reflectance distribution function (BRDF)

$$f_r(\omega_i, \omega_o; \{\rho_d, \rho_s, r, \gamma\}) = (1 - \gamma) \frac{\rho_d}{\pi} (f_{\text{diff}}(\omega_i, \omega_o) + f_{\text{retro}}(\omega_i, \omega_o; r)) + f_{\text{spec}}(\omega_i, \omega_o; \rho_d, \rho_s, r, \gamma). \quad (1)$$

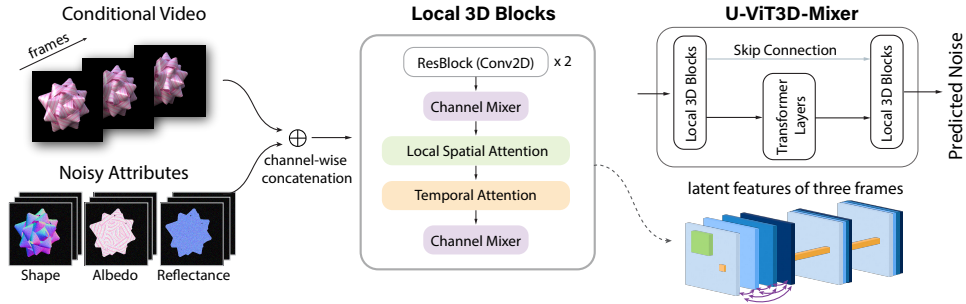


Figure 2: Our parameter-efficient denoising network, U-ViT3D-Mixer, takes in a channel-wise concatenation of conditional video frames and noisy shape-and-material frames. At high spatial resolutions, it uses efficient local 3D blocks (middle) with decoupled spatial, temporal, and channel-wise interactions. At lower spatial resolutions, it uses global transformer layers with full 3D attention.

In an idealized setting without ambient occlusion, this BRDF determines the reflected radiance at each surface point by the physically-based rendering equation [32]

$$L_r(u, v, \omega_o) = \int_{\omega} f_r(\omega_i, \omega_o) L_i(\omega_i) (\omega_i \cdot n(u, v)) d\omega_i, \quad (2)$$

where ω_i and ω_o are the incident and outgoing directions of light in the local coordinate frame of the hemisphere centered at normal $n(u, v)$, and L_i is the incident directional illumination. Unlike inverse rendering methods (e.g., [42]), we do not infer lighting and solely focus on recovering normals n and material properties M . We provide the full details of the rendering formulation in the appendix.

3.2 Generative Perception from Differential Motion

We design a conditional video diffusion model that can exploit the temporal continuity of differential motion and that supports simultaneous stochastic sampling of shape, texture, and reflectance. Our model also processes still images by treating them as static videos.

3.2.1 Joint Inference with a Diffusion Model

Our model has the unique feature of predicting shape and materials *jointly* using a single backbone. This design choice is motivated by our preliminary probing experiment (see Appendix A.2), where a model trained to predict albedo can be probed using DPT-style probes as in [17] to produce reasonable estimates of surface normals. This suggests strong cross-attribute interaction and motivates a unified architecture. Our approach is very different from existing models that require fine-tuning multiple pre-trained diffusion models for shape or materials [37], or that use a text prompt to switch between one output attribute at a time [61, 40].

We use a conditional diffusion probabilistic model [28] (DDPM) to generate shape-and-material attributes $\mathbf{x} \sim q(\mathbf{X}|\mathbf{c})$ by iterative denoising. To train the model we define a ‘forward process’ that adds Gaussian noise to a clean shape-and-material map $\mathbf{x}$:

$$\mathbf{z}_t = \alpha_t \mathbf{x} + \sigma_t \epsilon, \quad \text{where } \epsilon \sim \mathcal{N}(0, \mathbf{I}). \quad (3)$$

We use a continuous noise schedule where $\mathbf{z}_0$ is close to clean data and $\mathbf{z}_1$ is close to Gaussian noise, and we use a variance-preserving [28] noise schedule which implies $\alpha_t^2 + \sigma_t^2 = 1$.

We use a neural network with weights θ to approximate $q(\mathbf{X}|\mathbf{c})$ by a tractable distribution $p_\theta(\mathbf{X}|\mathbf{c})$. The network learns to estimate the noise in intermediate timesteps $t \in (0, 1)$ for all frames. During training, each conditioning frame $\mathbf{c}^f \in \mathbf{c}$ is associated with the same noise level σ_t and with different, randomly sampled Gaussian noise ϵ^f . We train the network by minimizing the prediction error averaged over all frames:

$$L(\theta) := \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \mathbb{E}_{t \sim \mathcal{U}(0,1), \epsilon^f \sim \mathcal{N}(0, \mathbf{I})} [\|\epsilon^f - \epsilon_\theta(\mathbf{z}_t^f, t; \mathbf{c}^f)\|_2^2]. \quad (4)$$

During sampling, given the conditional video frames $\mathbf{c}$, we perform the reverse diffusion process using DDIM [52, 20] beginning with a sample $\mathbf{z}_1$ from the standard Gaussian distribution. Given sample $\mathbf{z}_t$ at intermediate timestep t , we predict the clean sample $\hat{\mathbf{x}}$ with the noise estimate $\hat{\epsilon}_\theta$ from the denoiser neural network and project it to a lower noise level $s < t$ using

$$\mathbf{z}_s = \alpha_s \hat{\mathbf{x}} + \sigma_s \hat{\epsilon}_\theta(\mathbf{z}_t, t; \mathbf{c}), \quad \text{where } \hat{\mathbf{x}} = (\mathbf{z}_t - \sigma_t \hat{\epsilon}_\theta) / \alpha_t. \quad (5)$$

As $s \rightarrow 0$, we obtain the shape-and-material prediction $\hat{\mathbf{x}}^f$ for $f \in \mathcal{F}$. The sampling process of DDIM can be made deterministic and thus the randomness only comes from the initial noise $\mathbf{z}_1$.

3.2.2 Spatio-temporal and Multi-attribute Interactions

Unlike prior work that fine-tunes large pretrained diffusion models for estimating either shape [33, 43, 19, 46, 60] or materials [61, 7], we train our model to estimate both *from scratch* using a modest set of synthetic videos. This choice is motivated by findings from Han et al. [23], which demonstrate that pretrained models with strong priors can lack the flexibility to provide multimodal output distributions in the presence of ambiguity. Our training strategy, along with an improved noise schedule [35, 30] and our parameter-efficient architecture, allows us to investigate whether multimodal, generative perception can emerge naturally from learning, without being affected by the potential biases in large pretrained models.

As shown in Figure 2, our model takes in a set of input video frames $\mathbf{c}^f$ and concatenates it channel-wise with a set of noisy per-frame shape-and-material maps $\mathbf{x}_t^f = \{n_t^f, M_t^f\}$, $f \in \mathcal{F}$. This yields 12 channels per frame as the input to the denoising network. We introduce the following network components that act across spatial, temporal, and attribute dimensions to achieve joint shape-and-material perception from observations of differential motion.

Spatio-temporal interaction. Scaling pixel-space diffusion models to operate at higher resolutions (e.g., 256×256) introduces significant computational challenges. For instance, the diffusion UNet [28] and DiT [48] operate with full self-attention in each block, which becomes infeasible at higher resolutions due to a quadratic growth in token count.

To address this, we adopt a hybrid design inspired by U-ViT [29, 30] and Hourglass Diffusion Transformer [14]. At high spatial resolutions, we use ResNet-style convolution blocks [27] followed by separate local spatial attention (via shift-invariant neighborhood attention [25]) and temporal attention layers. At coarser resolutions (e.g., 32×32, 16×16), we apply transformer blocks [48] with global spatio-temporal attention and use 3D Rotary Positional Embeddings [55].

This spatial-temporal interaction at higher resolutions is particularly suited to exploiting the local space-time cues for shape, material and motion that are known to exist mathematically [12]. Our ablations show that it notably improves performance for both shape and material estimation.

Multi-attribute mixer. While convolution and local attention modules extract strong spatio-temporal features, they do not promote interaction across channels representing different physical attributes. To address this, we incorporate channel mixer blocks inspired by MLP-Mixer [56] after early convolutional and local attention layers. These blocks enable explicit inter-attribute communication, accelerating convergence and improving predictions.

Since our model performs multi-attribute denoising across shape, albedo and reflectance, we apply attribute-specific loss weighting coefficients $\lambda_{\text{shape}}, \lambda_{\text{albedo}}, \lambda_{\text{reflectance}}$. We find that upweighting shape and albedo ($\lambda_{\text{shape}} = \lambda_{\text{albedo}} = 0.4$) relative to metallic-ness, specular and roughness parameters ($\lambda_{\text{reflectance}} = 0.2$) leads to faster convergence and better overall performance.

In summary, our architecture extends existing pixel-space image diffusion architectures [50, 29, 29, 30, 14] to a video-conditioned, multi-attribute setting, enabling high-resolution generation with rich spatial, temporal, and cross-attribute interactions.

4 Experiments

To the best of our knowledge, no existing dataset provides high-quality annotations of physical object attributes under object motion. We therefore construct a synthetic dataset for training using the Mitsuba3 renderer [31] with custom integrators to extract ground-truth shape and material. We source objects from Adobe 3D Assets [1], obtaining 1100 artist-designed models. To enhance diversity,

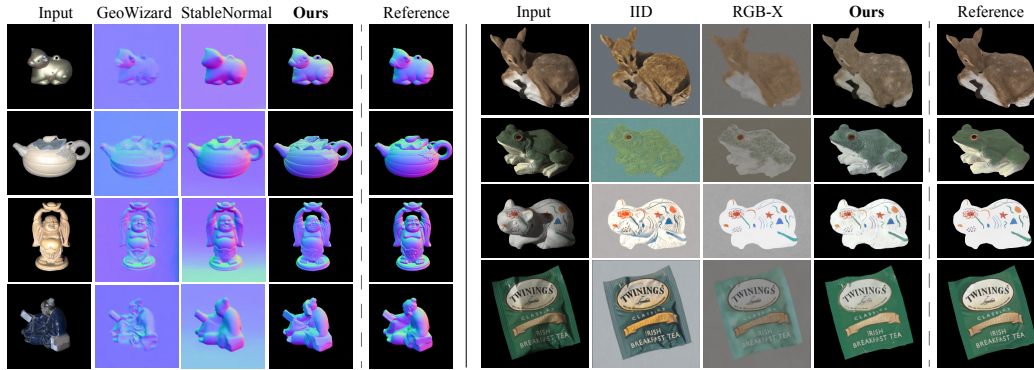


Figure 3: Qualitative comparisons of estimated shape and albedo/texture from a static scene. For fair comparison of scale-invariant albedos, we visualize the scaled albedo from each model closest to the ground truth in the masked object region. Our model achieves comparable shape predictions with existing baselines that specializes at shape prediction, and it achieves better quality in albedo estimation. Our model also produces greater spatial detail.

Table 1: Quantitative evaluation on shape and normal estimation tasks.

Model	Shape Estimation (DiLiGent)		Albedo Estimation (MIT-Intrinsic)		
	Mean AE ($\downarrow$)	Median AE ($\downarrow$)	SSIM ($\uparrow$)	RMSE ($\downarrow$)	PSNR ($\uparrow$)
Marigold-E2E-FT [46]	17.74	<u>13.86</u>	-	-	-
StableNormal [60]	18.21	14.59	-	-	-
DSINE [3]	20.29	16.17	-	-	-
GeoWizard [19]	24.62	21.55	-	-	-
IID [36]	-	-	0.61	0.13	18.19
ColorfulShading [11]	-	-	<u>0.72</u>	<u>0.10</u>	20.56
Marigold-Albedo [34]	-	-	<u>0.72</u>	0.11	19.86
RGB-X [61]	34.99	32.10	0.71	0.11	<u>20.82</u>
Ours	<u>18.03</u>	12.31	0.80	0.08	23.14

each shape is rendered under two material settings: (i) its original artist-designed materials, and (ii) a randomly sampled texture from a set of 400 open-source assets [13, 15, 58] combined with spatially-uniform reflectance from randomly sampled metallic, roughness, and specular coefficients.

All scenes are lit with one of 200 hundred randomly-sampled HDR environment maps collected from [2, 21], with horizontal and vertical environment flips for augmentation. To simulate object motion, we place the object centroid in the image center, randomly sample a 3D rotation for initial pose, and apply a motion rotation about the vertical image axis with a random angle sampled from $\{[-8^\circ, -2^\circ], [0^\circ, 2^\circ], [2^\circ, 8^\circ]\}$ with probability $[0.4, 0.2, 0.4]$. This means that 20% of training samples are static scenes containing no motion. We generate 45 short video clips (5 frames each) for each object, and we split them into pairs of consecutive 3-frame clips ($F = 3$) for training. This results in a dataset of approximately 100K video-attribute pairs.

We train our model using the AdamW optimizer [45]. Training requires roughly five days using four H100 GPUs. We follow the sigmoid loss reweighting strategy [35, 30] and adopt a shifted-cosine continuous noise schedule [30, 53]. Complete hyperparameter settings for our architecture are provided in the appendix. During inference, we use DDIM sampling [52] with 50 steps, which takes about 2.7 seconds per input video on a single A100 GPU.

4.1 Static Object Experiments

Our model processes single images of static objects by using F image copies as input, so we can assess its shape and material performance using existing single-image datasets. We evaluate our model in two ways: (i) its similarity to ground truth for everyday objects that are less ambiguous, and (ii) its ability to generate plausible multimodal solutions when the input is more ambiguous.

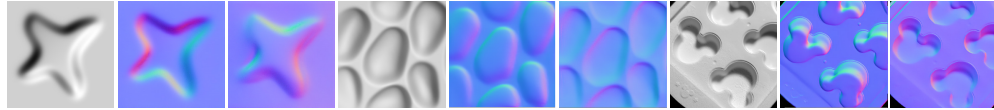


Figure 4: Our model exhibits multimodal shape perception on ambiguous visual stimuli presented in Han et al. [23] despite only being trained on everyday objects and without specific data augmentation.

Note that all comparison models in this section except RGB-X [61] are specialized for either shape or material estimation. Also, we cannot compare to the previous generative perception approaches of [18, 23] because they require either known shape or textureless diffuse objects.

We evaluate shape accuracy using the DiliGENT dataset [51]. This dataset contains photos taken under point light sources, originally designed for the photometric stereo task. Notably, this lighting is quite different from the environment lighting in our training set. For each of the 10 objects, we randomly sample 5 images and compute the mean and median angular surface normal errors within the masked object region. We compare with state-of-the-art shape prediction methods: StableNormal [60], Marigold-E2E-FT [46], GeoWizard [19], and DSINE [3]. All except DSINE are fine-tuned from pre-trained image diffusion models. For stochastic methods (ours, StableNormal, GeoWizard), we report the average of the three lowest error values from among 10 predictions.

We assess albedo/texture estimation using the MIT Intrinsic Image Dataset [22] which provides object photos under laboratory lighting and ground-truth albedo maps. We compare to a deterministic method, ColorfulShading [11], and diffusion-based methods Intrinsic Image Diffusion (IID) [36], Marigold-Albedo [34] and RGB-X [61]. Following prior work [36, 11], we use scale-invariant SSIM, RMSE, and PSNR for evaluation. For the IID model, we follow their recommendation by using the mean of 10 samples. For RGB-X, Marigold-Albedo and our model, we report the average of the three lowest error values from 10 predictions.

Table 1 shows quantitative results, and Figure 3 shows qualitative comparisons with other diffusion-based methods. For shape estimation, our model performs competitively with existing methods that specializes in shape prediction. Note that our model estimates the material in addition to the shape and is trained solely on synthetic data. For albedo estimation, our model outperforms existing models by a large margin. We attribute this improvement to (i) the use of pixel-space diffusion, which helps preserve fine details, and (ii) our synthetic training dataset with diverse and challenging objects using randomized texture and reflectance.

Overall, our model matches or exceeds all of the specialized shape or material models on existing benchmarks. The only other model that estimates both is RGB-X, but we find that it often struggles with complex shapes and provides less competitive performance across diverse objects.

Perceptual ambiguity. Han et al. [23] introduced a set of test images designed to probe the ambiguity-awareness of computational models, particularly for the convex/concave ambiguity. When evaluated on these images and as shown in Figures 1(iii) and 4, our model produces multimodal predictions, similar to human perception. Interestingly, due to training on textured rather than homogeneous objects, our model also occasionally exhibits a ‘postcard’ interpretation, perceiving a flat surface with painted texture. We find the joint estimation of shape and material further enhances interpretability through disentanglement and can offer insight into the model’s perceptual hypotheses.

4.2 Moving Object Experiments

We compare three ablated variants of our model: (i) **U-ViT3D (base)** is a baseline with convolutional layers at high resolutions and transformer blocks in the bottleneck; (ii) **+ Local Spatio-Temporal Attention** replaces one ResNet block per resolution level with a local spatial attention and a temporal attention module; and (iii) **+ Multi-Attribute Mixer** further adds lightweight channel mixers to each local 3D block.

We conduct ablation studies on a held-out synthetic evaluation set containing objects undergoing differential motion. These are generated using the same pipeline as our training data but with distinct objects to ensure generalization. We evaluate performance under both moving and static observations. For each video, we draw 10 samples and report the mean of the top-3 based on a joint ranking of

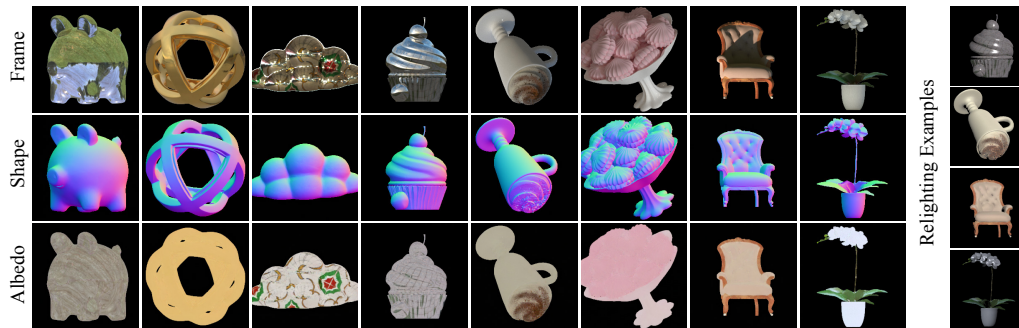


Figure 5: Shape and material estimates from our full model. For each three-frame test video, we show one representative frame. The first four columns use randomly sampled materials, and the last four use original materials. Note that some specular objects are quite challenging; we demonstrate in the supplementary videos how motion aids disambiguation. On the right we show relighting examples using the estimated shape, albedo, and reflectance parameters under directional lighting.

Table 2: Ablation study on shape and albedo estimation using our differential-motion evaluation set. Reported metrics are computed over mean of the top 3 of 10 stochastic samples per video. Model variants are compared under both moving and static observation conditions.

Model Variant	# Params	With Motion		Without Motion	
		Mean AE ↓	SSIM ↑	Mean AE ↓	SSIM ↑
U-ViT3D (base)	95.7M	18.757	0.778	22.471	0.760
+ Local Spatio-Temporal Attention	97.4M	14.555	0.825	18.957	0.811
+ Multi-Attribute Mixer	100.3M	11.262	0.832	16.886	0.823

shape (Mean Angular Error (MAE)) and albedo (SSIM). This combined metric helps capture overall performance instead of focusing solely on shape or materials.

Table 2 shows that each module leads to notable performance gains with minimal added parameters. Specifically, we find that using spatial neighborhood attention is crucial for better shape estimation. The base model with only convolutions tends to produce over-smoothed results on complex inputs and often fails to reason about the shape of highly specular objects. Our model also shows better performance when it observes objects under motion instead of being static, highlighting its ability to leverage motion cues. Figure 5 visualizes predictions from the full model on motion inputs.

Perceptual ambiguity. To assess our model’s ability to adapt its hypotheses in the presence of motion, we apply it to the perceptual demonstrations from Hartung and Kersten [24]². These are videos of moving objects rendered with either (a) a uniform shiny material that reflects the environment, or (b) a spatially-varying matte material representing painted texture. The two scenarios are indistinguishable in the absence of motion, because they can produce individual frames that are identical.

We probe our model’s interpretations by visualizing its distribution of generated albedo samples, because the shiny interpretation implies a spatially-uniform albedo whereas the matte interpretation does not. The left of Figure 6 shows an embedding of 100 albedo predictions when seeing a shiny teapot that is either static or moving. The albedo samples are diverse in the static case, containing both shiny and matte interpretations, and they converge to the shiny explanation when there is motion.

The right of Figure 6 shows albedo samples for a single video frame that is common across two distinct videos of the same shape rendered as shiny or matte. The frame is ambiguous in isolation, so the interpretation is heavily influenced by temporal motion. We find that our model clearly distinguishes between the shiny and matte explanations, with samples forming two distinct clusters.

These experiments demonstrate that our model can effectively leverage differential motion to refine its material prediction, similar to human observers [24]. Note that our dataset does not contain any objects with painted textures that are purposefully designed to look like reflected environments. Nonetheless, it learns to include that possibility among its predictions for static objects, and it learns to either confirm or deny that possibility in the presence of motion.

²We encourage viewing of videos at <https://kerstenlab.psych.umn.edu/demos/motion-surface/>

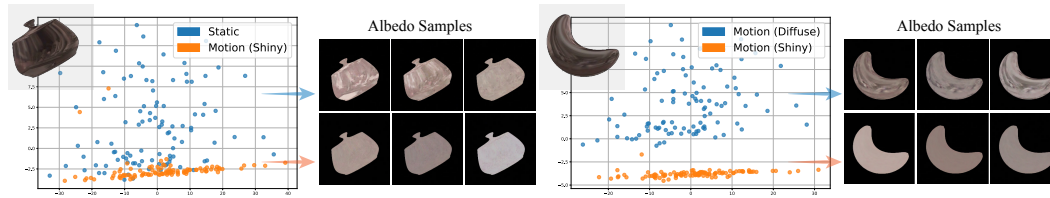


Figure 6: Motion disambiguation results for the Shiny or Matte Ambiguity [24]. Each plot shows a PCA embedding of 100 albedo samples corresponding to the frame (inset) that is common between two distinct input videos. *Left:* When a shiny teapot moves, the albedo samples converge to a *subset* of the ones produced for the static teapot. *Right:* The albedo samples for a matte-rendered motion video are clearly *separated* from those of a shiny-rendered video. Note that spatially-uniform albedos form tighter clusters in PCA space, while highly textured ones exhibit greater variation.

4.3 Real-world Examples

Figure 7 and 8 shows additional examples with online images and real-world videos. For each input, we segment the objects and mask the background before passing it to the model.

Figure 7 illustrates how our model jointly interprets shape and albedo from static inputs. In the top row, the carpet uses distorted textures and shading to mimic a ‘visual cliff’. In the bottom row, our model generalizes to an artist’s painting, producing either a flat ‘postcard’ interpretation or a shaded 3D one. When the shape is perceived as three-dimensional, the predicted albedo appears more uniform.

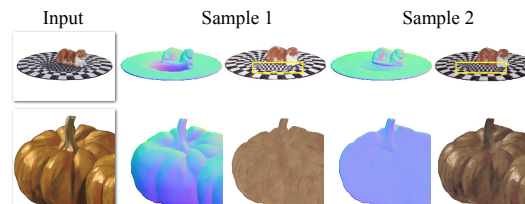


Figure 7: Multiple shape and material hypotheses on an ambiguous online image. A cat on an illusion carpet painted to resemble a hole. When interpreted as a hole (Sample 1), the predicted albedo is brighter in the center region (yellow box) than when interpreted as a plane (Sample 2).

Figure 8 compares our model to RGB-X [61] and DiffusionRenderer [40] on real-world videos of various objects under natural outdoor or indoor lighting. We test them in two settings (a) moving objects, same as our training paradigm and (b) moving camera, which is out-of-distribution for our model. For the latter we use the captured images in the Stanford-ORB dataset [38], by feeding our model sets of three images from nearby camera views. In both scenarios, we show that our model produces better visual quality for both shape and albedo, especially for shiny and metallic objects.

To test the overall generalization of our model to unseen motions, we include additional quantitative results in the appendix A.6 and A.7.

Summary. Experimentally we find that our model generalizes quite well to out-of-distribution inputs, including to ambiguous perceptual stimuli (Fig. 4) and to captured videos that contain multiple objects or novel types of motions (Fig. 8). We attribute this to: (i) our use of local, shift-invariant attention at high resolutions, which helps avoid overfitting to global object features; and (ii) our training set, which includes 3D assets composed of multiple objects and self-occluding object-parts.

Limitations. Our model currently focuses on restricted classes of motions and materials. While it can handle multiple objects undergoing the same rigid motion, it does not apply to scenes that contain deforming objects, or that contain multiple distinct motions. Nor does it apply to objects that exhibit translucency or transparency. Another limitation of our model is its reliance on objects being pre-segmented from the background. Extending the model to perform perceptual grouping in addition to inferring object attributes is an important future direction.

5 Conclusion

Inspired by the interdependence of shape and material perception, and by the human ability to resolve ambiguities through motion, we introduce a generative perception approach that produces combined samples of shape, texture, and reflectance which explain a short video of differential object motion. Our model recovers complete descriptions of an object’s shape and materials while capturing and modeling inherent ambiguities. We show that it successfully maintains a distribution



Figure 8: Outputs from captured videos of real world objects. Our model effectively uses differential motion to produce higher-quality estimations of shape and albedo compared to RGB-X and DiffusionRenderer.

of plausible interpretations under ambiguous static observations, and that it naturally converges to more concentrated and accurate hypotheses in the presence of motion. By introducing an effective conditional video diffusion model that operates directly in pixel space, we also provide a potential new backbone for space-time generative perception. We hope this inspires future research in ambiguity-aware perception and dynamic vision for embodied systems.

Acknowledgments

We thank Jianbo Shi for reviewing the manuscript. We also thank Kohei Yamashita for guidance about the data rendering pipeline and Boyuan Chen for discussion about video diffusion modeling. This work was supported in part by the NSF cooperative agreement PHY-2019786 (an NSF AI Institute, <http://iaifi.org>) and by JSPS 21H04893 and JST JPMJAP2305.

References

- [1] Adobe Stock. <https://stock.adobe.com/>.
- [2] HDRIs: poly haven. <https://polyhaven.com/hdris>.
- [3] Gwangbin Bae and Andrew J. Davison. Rethinking inductive biases for surface normal estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [4] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1670–1687, 2014.
- [5] Ronen Basri and Darya Frolova. A two-frame theory of motion, lighting and shape. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE.
- [6] Peter N Belhumeur, David J Kriegman, and Alan L Yuille. The bas-relief ambiguity. *International Journal of Computer Vision*, 35(1):33–44, 1999.
- [7] Sean Bell, Kavita Bala, and Noah Snaveley. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)*, 33(4):1–12, 2014.
- [8] Yanrui Bin, Wenbo Hu, Haoyuan Wang, Xinya Chen, and Bing Wang. Normalcrafter: Learning temporally consistent normals from video diffusion priors. *arXiv preprint arXiv:2504.11427*, 2025.
- [9] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.
- [10] Brent Burley. Physically-based Shading at Disney. In *siggraph*, 2012.
- [11] Chris Careaga and Yağız Aksoy. Colorful diffuse intrinsic image decomposition in the wild. *ACM Transactions on Graphics (TOG)*, 43(6):1–12, 2024.
- [12] Manmohan Chandraker. The information available to a moving observer on shape with unknown, isotropic BRDFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(7):1283–1297, 2015.
- [13] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [14] Katherine Crowson, Stefan Andreas Baumann, Alex Birch, Tanishq Mathew Abraham, Daniel Z Kaplan, and Enrico Shippole. Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers. In *Forty-first International Conference on Machine Learning*, 2024.
- [15] Valentin Deschaintre, Miika Aittala, Fredo Durand, George Drettakis, and Adrien Bousseau. Single-image svbrdf capture with a rendering-aware deep network. *ACM Transactions on Graphics (ToG)*, 37(4):1–15, 2018.
- [16] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.
- [17] 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*, pages 21795–21806, 2024.
- [18] Yuto Enyo and Ko Nishino. Diffusion reflectance map: Single-image stochastic inverse rendering of illumination and reflectance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.
- [19] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3D geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2024.
- [20] Ruiqi Gao, Emiel Hooeboom, Jonathan Heek, Valentin De Bortoli, Kevin P. Murphy, and Tim Salimans. Diffusion meets flow matching: Two sides of the same coin. 2024. URL <https://diffusionflow.github.io/>.
- [21] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090*, 2017.

- [22] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2335–2342. IEEE, 2009.
- [23] Xinran Han, Todd Zickler, and Ko Nishino. Multistable shape from shading emerges from patch diffusion. *Advances in Neural Information Processing Systems*, 37:34686–34711, 2024.
- [24] Bruce Hartung and Daniel Kersten. Distinguishing shiny from matte. *Journal of Vision*, 2(7):551–551, 2002.
- [25] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6185–6194, 2023.
- [26] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. *arXiv preprint arXiv:2409.18124*, 2024.
- [27] 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*, pages 770–778, 2016.
- [28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [29] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. Simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pages 13213–13232. PMLR, 2023.
- [30] Emiel Hoogeboom, Thomas Mensink, Jonathan Heek, Kay Lamerigts, Ruiqi Gao, and Tim Salimans. Simpler diffusion (sid2): 1.5 fid on imagenet512 with pixel-space diffusion. *arXiv preprint arXiv:2410.19324*, 2024.
- [31] Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, Merlin Nimier-David, Delio Vicini, Tizian Zeltner, Baptiste Nicolet, Miguel Crespo, Vincent Leroy, and Ziyi Zhang. Mitsuba 3 renderer, 2022. <https://mitsuba-renderer.org>.
- [32] James T. Kajiya. The Rendering Equation. *SIGGRAPH*, 20(4):143–150, August 1986.
- [33] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024.
- [34] Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. Marigold: Affordable adaptation of diffusion-based image generators for image analysis. *arXiv preprint arXiv:2505.09358*, 2025.
- [35] Diederik Kingma and Ruiqi Gao. Understanding diffusion objectives as the elbo with simple data augmentation. *Advances in Neural Information Processing Systems*, 36:65484–65516, 2023.
- [36] Peter Kocsis, Vincent Sitzmann, and Matthias Nießner. Intrinsic image diffusion for indoor single-view material estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5198–5208, 2024.
- [37] Peter Kocsis, Lukas Höllein, and Matthias Nießner. Intrinsix: High-quality PBR generation using image priors. *arXiv preprint*, 2025.
- [38] Zhengfei Kuang, Yunzhi Zhang, Hong-Xing Yu, Samir Agarwala, Elliott Wu, Jiajun Wu, et al. Stanford-ORB: a real-world 3d object inverse rendering benchmark. *Advances in Neural Information Processing Systems*, 36:46938–46957, 2023.
- [39] Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. SyncDiffusion: Coherent montage via synchronized joint diffusions. *Advances in Neural Information Processing Systems*, 36, 2024.
- [40] 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*, pages 26069–26080, 2025.
- [41] Stephen Lombardi and Ko Nishino. Reflectance and Natural Illumination from a Single Image. In *ECCV*, pages 582–595. Springer, 2012.

- [42] Stephen Lombardi and Ko Nishino. Reflectance and Illumination Recovery in the Wild. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 38(1):129–141, Jan. 2016.
- [43] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3D: Single image to 3D using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024.
- [44] M. S. Longuet-Higgins. Reflection and refraction at a random moving surface. i. pattern and paths of specular points. 50(9):838–844. Publisher: Optical Society of America.
- [45] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.
- [46] Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan de Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.
- [47] Geoffrey Oxholm and Ko Nishino. Shape and Reflectance Estimation in the Wild. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 38(2):376–389, Feb. 2016.
- [48] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.
- [49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [50] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.
- [51] Boxin Shi, Zhe Wu, Zhipeng Mo, Dinglong Duan, Sai-Kit Yeung, and Ping Tan. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3707–3716, 2016.
- [52] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *International Conference on Learning Representations*, 2021.
- [53] Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. *arXiv preprint arXiv:2502.06764*, 2025.
- [54] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.
- [55] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
- [56] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. MLP-Mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.
- [57] Yuriy Vasilyev, Todd Zickler, Steven Gortler, and Ohad Ben-Shahar. Shape from specular flow: Is one flow enough? In *CVPR 2011*, pages 2561–2568. IEEE.
- [58] Giuseppe Vecchio and Valentin Deschaintre. Matsynth: A modern PBR materials dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [59] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024.
- [60] Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. StableNormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (TOG)*, 43(6):1–18, 2024.
- [61] Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. RGB $\leftrightarrow$ X: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers, SIGGRAPH '24*, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705250. doi: 10.1145/3641519.3657445.

A Technical Appendices and Supplementary Material

A.1 Additional Related Work on 3D Reconstruction

Many recent learning-based multi-view scene reconstruction methods infer point clouds from images [71, 69, 70, 68]. These models are designed for deterministic reconstruction instead of modeling ambiguity, and they do not estimate material properties, so often have the surrounding illumination baked into their output point colors. Although some view-synthesis techniques based on NeRFs [67] and Gaussian splatting [63, 62] can disentangle reflectance and lighting, they require many viewpoints and per-scene optimization. In contrast, our method is feed-forward, and it leverages differential motion in short videos to successfully resolve texture–lighting ambiguity without the need for many wide-baseline images and explicit camera pose estimation.

Some other recent studies [43, 64, 71] generate 3D assets from text or a conditioning image, including the generation of object parts that are not visible in the input. These models are often trained on artist-designed, cartoon-style 3D assets instead of photorealistic ones, and their output is typically a mesh plus a texture that has highlights and other illumination effects baked in.

A.2 Probing experiment on normal estimation from albedo prediction

We conduct a probing experiment with a low-resolution (64×64) diffusion UNet that is trained to infer the object albedo given a conditional image. We are interested in whether the model learns useful representations for predicting object surface normals. To do this, inspired by prior work [17], we first pre-train the albedo prediction UNet and then insert probes similar to the DPT decoder [66] using multiscale convolutions at intermediate layers. We train the lightweight probes to estimate shape from latent features using a small dataset of 10K ground-truth pairs for five epochs. At test time, we directly read out from the inserted probes for surface normal estimates.

As shown in Figure 9, the model trained for albedo prediction can indeed produce plausible shape estimates through probing. As a result, we hypothesize that building a unified framework for both shape and material estimation can leverage a shared representation and enable a more computationally efficient architecture.

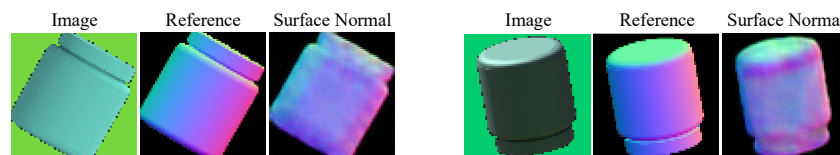


Figure 9: A model trained to predict surface albedo learns representation useful for estimating surface shapes. We show two examples with the ground truth shape (middle) and the readout shape from simple convolutional probes (right).

A.3 Ablation Studies

In Figure 10a, we compare the shape estimates from our base model (U-ViT3D) with those from our full model (U-ViT3D-Mixer). The base model uses only convolutions at high resolutions, so it tends to produce over-simplified geometry estimates. On the other hand, our full U-ViT3D-Mixer model captures finer details robustly, even for shiny objects.

In Figure 10b, we show the benefits of leveraging motion cues for material estimation. Texture estimation is highly ambiguous when the object is static: the appearance could come from painted texture, a reflected environment, or a mixture of the two. Figure 10b shows samples of our model’s textures from ambiguous static images, and it shows how object motion allows the model to separate texture more accurately.

A.4 Extension using Temporal Consistency Guidance

Our model generates short clips ($F = 3$ frames), but it can be used to create longer sequences by simultaneously generating multiple clips and using inference-time guidance (e.g., [39, 23]) to enforce

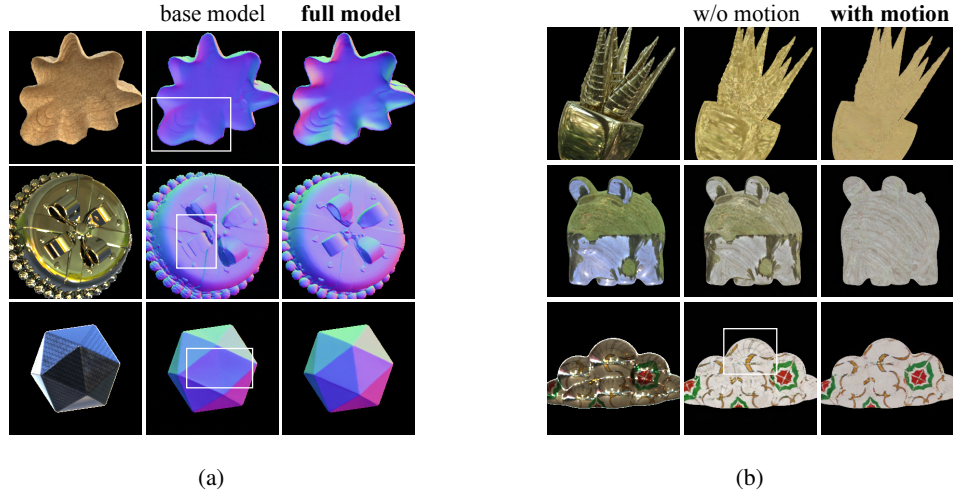


Figure 10: *Left*: Ablation of model components. The base U-ViT3D model, i.e., local attention and channel mixing layers ablated from our full model, produces over-smoothed results and struggles with specular surfaces due to complex inter-reflections. *Right*: Ablation of motion from input images. When the shiny objects are rendered as being static, the model can sometimes provide inaccurate estimates due to the inherent ambiguity between texture and illumination. Once the model observes the object undergoing motion, this ambiguity is resolved and the model produces more accurate albedo estimates.

temporal consistency between them. During inference, we apply a reconstruction loss (e.g., MSE loss) at one overlapping frame across adjacent temporal windows, encouraging the denoised predictions $\hat{x}$ to remain consistent at that frame. We find that this can provide temporally consistent shapes and materials over long time horizons, without the need for modifications or additional training. On the project page, we compare our long-horizon results to StableNormal [60], which shows substantially more temporal flickering and inconsistency.

A.5 Additional evaluation results for IID [36]

In the experiment section Table 1, we use the mean prediction for the IID baseline following the authors’ recommended setting because it achieves the best performance in their original paper [36]. The evaluation results with the alternative top-3 out of 10 evaluation protocol are below in Table 3. They are consistently worse than the mean prediction, consistent with the original paper’s claims.

Table 3: IID [36] performance under different evaluation protocols.

IID - Albedo Estimation	SSIM $\uparrow$	RMSE $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$
Mean prediction out of 10 samples	0.61	0.13	18.19	0.16
Best 3 out of 10 samples	0.44	0.15	17.01	0.20

A.6 Quantitative Results using the Stanford-ORB dataset

In the main paper we show qualitative results using images from the Stanford-ORB benchmark [38], where we feed our model sets of three images from nearby camera views. The effective “motion” in this experiment is very different from our training data because (i) the camera moves instead of the object; and (ii) the camera positions are separated by wide, irregular baselines. Despite this, we find that our model achieves high accuracy, especially in comparison to the two strongest comparison models—Marigold-e2e-ft [46] for shape and ColorfulShading [11] for albedo—using the same input images. Tables 4 and 5 show results for six of the benchmark’s scenes, where we see competitive performance for shape and superior performance for albedo. This is an example of our model generalizing beyond the horizontal object motion used for training.

Table 4: Stanford ORB real-world objects: shape inference results

Model	Shape (Median AE) ↓					
	Blocks	Cactus	Car	Cup	Grogu	Teapot
Marigold-e2e-ft [46]	10.19	14.59	13.69	12.14	13.27	13.25
Ours	11.22	19.50	13.86	7.85	10.84	15.02

Table 5: Stanford ORB real-world objects: albedo inference results

Model	Albedo (SSIM ↑ / PSNR ↑)					
	Blocks	Cactus	Car	Cup	Grogu	Teapot
ColorfulShading [11]	0.76 / 24.69	0.66 / 25.58	0.64 / 23.01	0.84 / 24.83	0.61 / 19.87	0.69 / 23.33
Ours	0.74 / 25.28	0.78 / 26.23	0.70 / 23.68	0.88 / 24.97	0.86 / 24.05	0.79 / 24.12

A.7 Generalization to random rotation and translation

Some examples in the main paper already suggest that our model can generalize to non-horizontal object motions. For instance, in the left video of Figure 6, the teapot rotates vertically, orthogonal to the motion seen during training. To examine this, we perform a simulation experiment using 174 unseen objects with randomized textures, each rendered under three different motion settings:

1. a vertical-axis rotation, as used for training;
2. a rotation about a randomly-chosen axis; and
3. a translation of up to 8% of the object size in a random direction.

Note that (2) and (3) are very different from training, and that (3) is equivalent to having a static image with translational misalignments between frames.

The results in Table 6 show there is limited degradation from in-distribution motions to out-of-distribution ones. We attribute this to the fact that local shape and material constraints derive from the relative angles between rotation axis, normals and curvatures (see [12]), and that since we use a wide variety of shapes during training, the model sees many of these local relationships, each affected by a local translation that depends on the point’s 3D position with respect to the rotation axis.

Table 6: Evaluation across different types of object motions.

Scenario	Shape		Albedo	
	Mean AE ↓	Median AE ↓	SSIM ↑	RMSE ↓
In-dist. vertical-axis rotation	12.267	10.415	0.868	0.051
Out-of-dist. random rotation	15.062	13.158	0.863	0.052
Out-of-dist. random translation	16.652	14.681	0.858	0.054

A.8 Reflectance Model

We describe the full reflectance model based on the Disney principled BSDF model [10]. Recall the reflectance equation

$$f_r(\omega_i, \omega_o; M) = (1 - \gamma) \frac{\rho_d}{\pi} [f_{\text{diff}}(\omega_i, \omega_o) + f_{\text{retro}}(\omega_i, \omega_o; r)] + f_{\text{spec}}(\omega_i, \omega_o; \rho_d, \rho_s, r, \gamma), \quad (6)$$

where $M = \{\rho_d, \rho_s, r, \gamma\}$ consists of the material parameters: diffuse albedo ρ_d , specular term, ρ_s roughness r , metallic-ness coefficient γ and n denotes the spatially varying surface normal at a point.

The diffuse term depends on the angle of incident light θ_i and outgoing direction θ_o :

$$f_{\text{diff}} = (1 - F_i/2) (1 - F_o/2), \quad \text{where } F_i = (1 - \cos \theta_i)^5, \quad F_o = (1 - \cos \theta_o)^5. \quad (7)$$

To capture retro-reflective highlights, we add

$$f_{\text{retro}} = R_R (F_i + F_o + F_i F_o (R_R - 1)), \quad \text{where } R_R = 2\gamma \cos^2 \theta_d, \quad \cos \theta_d = h \cdot \omega_i, \quad (8)$$

with h denoting the half-vector between the incident and outgoing directions.

The specular reflection is based on the microfacet model with GGX distribution:

$$f_{\text{spec}} = \frac{F D G}{4 (\omega_i \cdot n) (\omega_o \cdot n)} = \frac{F D G}{4 \cos \theta_i \cos \theta_o}. \quad (9)$$

Here, D is a microfacet distribution function defined by the roughness parameter r ,

$$D(h) = \frac{r^4}{\pi ((n \cdot h)^2 (r^4 - 1) + 1)^2}, \quad (10)$$

and G denotes the masking-shadowing function

$$G = G_1(\omega_i) G_1(\omega_o), \text{ where } G_1(\omega) = \frac{2}{1 + \sqrt{1 + r^4 (1 - n \cdot \omega)^2 / (n \cdot \omega)^2}}. \quad (11)$$

The Fresnel term F blends dielectric and metallic responses:

$$F = (1 - \gamma) F_{\text{dielectric}} + \gamma F_{\text{Schlick}}, \quad (12)$$

with

$$F_{\text{dielectric}} = \frac{1}{2} \left(\left(\frac{\cos \theta_i - \eta \cos \theta_t}{\cos \theta_i + \eta \cos \theta_t} \right)^2 + \left(\frac{\cos \theta_t - \eta \cos \theta_i}{\cos \theta_t + \eta \cos \theta_i} \right)^2 \right), \text{ where } \eta = \frac{2}{1 - \sqrt{0.08 \rho_s}} - 1, \quad (13)$$

where θ_t is the angle between the normal and the transmitted ray computed using Snell's Law and

$$F_{\text{Schlick}} = \rho_d + (1 - \rho_d) (1 - \cos \theta_d)^5. \quad (14)$$

A.9 Network Architecture Details

We use the following hyperparameters for the U-ViT3D-Mixer model.

```
channels          = [96, 192, 384, 768],
block_dropout    = [0, 0, 0.1, 0.1],
block_type       = ['Local3D'(1), 'Local3D'(1), 'Transformer'(3), 'Transformer'(8)],
noise_embedding_channels = 768,
attention_num_heads = 6,
patch_size       = 2,
local_attention_window_size = 7,
channel_mixer_expansion_factor = 3,
loss_type        = v-prediction (MSE)
```

We use the following training setup.

```
batch_size       = 64,
optimizer        = 'AdamW',
adam_betas       = (0.9, 0.99),
adam_weight_decay = 0.01
learning_rate     = 1e-4,
mixed_precision   = 'bfloat16',
max_train_steps   = 400k
```

A.10 Inference Hardware Requirement

We compare our model to other stochastic, diffusion-based baseline models in terms of inference time GPU requirement and runtime. We use each model's FP32 version and conduct the analysis on a single A100 GPU. As shown in Table 7, our model is compact, fast and memory efficient, due to its architectural design.

Table 7: Model comparison of inference requirements.

Model	Output	Model Size (FP32)	Inference Time	Inference GPU Memory
Ours	Shape and Materials, three frames	0.4 GB	2.7 s	2.8 GB
StableNormal	Shape only, single-frame	3.8 GB	1.1 s	19.5 GB
IID	Materials only, single-frame	6.2 GB	6.0 s	9.5 GB
RGB-X	Shape <i>or</i> Materials, single-frame	3.8 GB	10.5 s	6.0 GB

Appendix References

- [62] 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. *arXiv preprint arXiv:2311.17977*, 2023.
- [63] Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, and Kui Jia. Gs-ir: 3D gaussian splatting for inverse rendering. *arXiv preprint arXiv:2311.16473*, 2023.
- [64] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3D mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024.
- [43] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3D: Single image to 3D using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024.
- [66] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.
- [67] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [68] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [69] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. *arXiv preprint arXiv:2501.12387*, 2025.
- [70] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUST3R: Geometric 3D vision made easy. In *CVPR*, 2024.
- [71] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3D latents for scalable and versatile 3D generation. *arXiv preprint arXiv:2412.01506*, 2024.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: Our experiment section 4 reflect the claims in the abstract and introduction: (1) joint inference of shape and materials on real and synthetic objects (2) emergent ability to suggest multiple hypotheses for ambiguous input and (3) leverage motions cues for more accurate inference.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We discuss limitations and possible future work directions in the conclusion section Sec. 4.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not include new theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We include in the paper a detailed description of the model architecture, with implementation and model training details in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: Yes we include those details in the supplementary materials. We plan to open source the code, pretrained model as well as our synthetically generated dataset. We will also provide our data generation script for customized usage.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: Yes we include those details in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[No\]](#)

Justification: We do not report statistical significance because of the observation that object attributes disentanglement can be inherently ambiguous. We support our quantitative results of shape and albedo estimation by drawing 10 samples and taking the top three for all stochastic methods. For evaluation on motion disambiguation, we draw 100 samples to demonstrate the distribution shift in estimated albedo.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We include the compute resource requirement in the experiment section 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Our work conform, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We have discussed the potential positive impact of our model: our framework can serve a generative perception backbone for understanding time-varying visual inputs, useful for physically embodied agents. We do not perceive immediate negative societal impact from our work.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The data and model used in our work do not pose such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Yes we have credited and licensed the Adobe 3D assets used for generating our synthetic dataset.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: We will be releasing our dataset and codebase with proper documentation.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#)

Justification: This work does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[NA\]](#)

Justification: This work does not involve human participants.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer:[NA]

Justification: The development of the core method does not involve LLM usage.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.