This document provides supplementary details for our main paper. We begin with the preliminaries of our method (Sec. A), followed by more details on the training data production (Sec. B), specific settings for our Block-wise Analysis (Sec. C), and the full experimental evaluation protocol (Sec. D). Furthermore, we present additional ablation studies (Sec. F) and an analysis of failure cases (Sec. G). Full video comparisons are available on our project page.

A. Prelinimaries

WAN2.1 Model Structure. Our work is built upon WAN2.1 [38], a video diffusion model constructed as a stack of customized Diffusion Transformer (DiT) blocks [29]+. Each block is composed of a visual self-attention layer for modeling spatio-temporal relationships within the video, and a cross-attention layer to incorporate textual conditioning, although text prompts are not used in our specific application.

To adapt this architecture for the video inpainting task, we follow a similar input formulation to Gen-Omnimatte [15]. The process is as follows: first, the input video $\boldsymbol{V}$ is encoded into a latent representation $\mathbf{z}_v$ using a pre-trained Variational Autoencoder (VAE). This video latent is then concatenated along the channel dimension with the downsampled frame-wise binary mask $\boldsymbol{M}$ and a noise latent of the same spatial dimensions. The concatenated tensor is then passed through a linear projection layer to compress its channel dimension, forming the final sequence of visual tokens. These tokens are then duplicated and concatenated to serve as the input to the DiT model for the diffusion process.

LoRA and Training Objectives. To efficiently fine-tune the model for the Omnimatte task, we integrate Low-Rank Adaptation (LoRA) [10] into the model's self-attention layers. For a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, the update is represented by a low-rank decomposition $W_0+\Delta W = W_0+BA$, where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are trainable matrices with a low rank $r \ll \min(d, k)$. The forward pass is modified as:

$$h = W_0x + BAx. \tag{8}$$

In our framework, the Branch DiT applies this LoRA computation selectively to the copied tokens that are designated to learn the alpha matte, leaving the original tokens to be processed by the frozen, pre-trained weights.

Our model is trained with direct supervision on the alpha matte prediction. We employ the standard flow matching objective [22], which trains the model to predict the noise $\boldsymbol{\epsilon}$ added to a clean latent $\mathbf{z}_0$ at timestep t. The loss function is defined as:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,\mathbf{z}_0,\mathbf{z}_1}\left[||(\mathbf{z}_1 - \mathbf{z}_0) - \mathbf{v}_\theta(\mathbf{z}_t, t)||^2\right]. \tag{9}$$

where $\mathbf{z}_0$ is the VAE-encoded ground-truth alpha matte, $\mathbf{z}_1$ is the sampled Gaussian noise, t is the diffusion timestep, and $\mathbf{v}_\theta$ is our network (with trainable LoRA parameters) that predicts the velocity from the noisy latent $\mathbf{z}_t$.

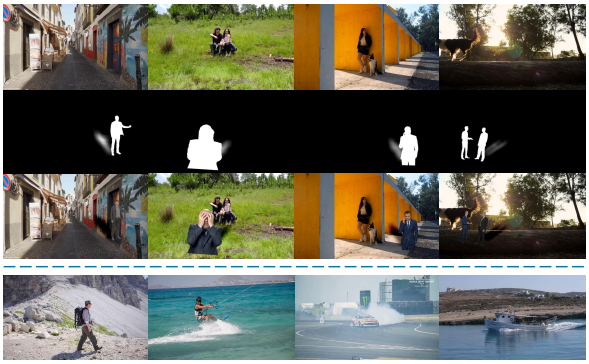

Figure 10. **Generalization from simple synthetic data to complex real-world scenes.** Our method is trained exclusively on synthetic data (top) featuring only basic shadow effects. Despite this, it successfully decomposes challenging in-the-wild videos (bottom) with a variety of unseen effects, including reflections and smoke. This highlights the model's ability to bridge a substantial domain gap by learning a true separation principle.

B. Training Data

While the sources and types of our input data are outlined in the main paper, this section provides a detailed description of our data augmentation process and presents visual examples of our training data.

Data Augmentation. We employ a temporally coherent video augmentation pipeline adapted from [20, 21] to improve model robustness and generalization. This involves simulating camera motion by first embedding the sequence within a larger canvas using asymmetric padding (sampled from a uniform distribution $\mathcal{U}(0.3, 0.5)$ of the target width), followed by a smooth affine transformation. Rather than per-frame randomization, we interpolate between two affine states (A and B) using an easing function. To create pronounced lateral movement, the horizontal translations of A and B are set to be of opposite sign with magnitudes ranging from 15-30% of the image width, while rotation ($\pm5°$), scale (0.95-1.05), and shear ($\pm3°$) are varied subtly.

A crucial component of our data synthesis is the addition of realistic associated effects. We focus on simulating shadows, as they are one of the most common and challenging effects to separate. For a given foreground alpha matte α, we generate a shadow matte α_s by applying strong vertical compression (to 10-30% of original height), significant horizontal shear ($30° - 60°$), semi-transparent rendering (30-70% opacity) and a final blur. This shadow matte is then used to darken the corresponding region on the background video before the final foreground is composited on top. This process forces the model to learn to identify and separate regions that are visually part of the foreground layer but are not captured by the original object mask.

Demonstration and Discussion. In Fig. 10, we provide a visual comparison between our synthetic training data and

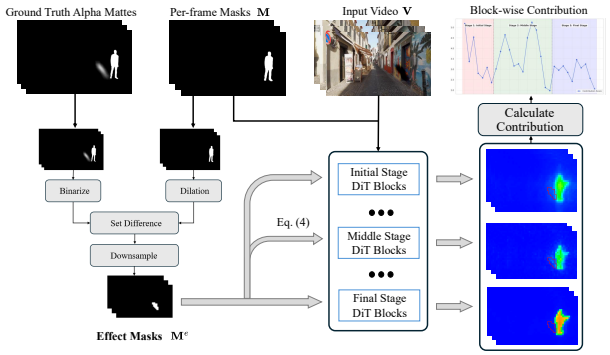

Figure 11. Pipeline of our Block-wise Contribution Analysis Method. The ground-truth alpha matte enables us to precisely localize the regions containing effects and subsequently compute the contribution score.

real-world videos from our test set. The top row showcases examples of our generated training frames, including the final composited video and the corresponding ground-truth alpha matte, which includes both the object and its simulated shadow. The bottom row presents frames from real videos, where the effects (e.g., natural shadows, reflections) are significantly more complex and subtle.

As is visually evident, a substantial *domain gap* exists between our synthetic training data and the real-world test scenarios. Our training set, despite the augmentations, features clean-cut objects and simplified, programmatic shadows. In contrast, real videos contain complex lighting, soft and intricate effects, and various image artifacts. Despite this gap, our model demonstrates strong performance on these real-world examples, successfully isolating both the foreground object and its nuanced, naturally occurring effects. This robust generalization capability suggests that our end-to-end training approach has enabled the model to learn the fundamental, underlying logic of foreground-effect separation, rather than merely memorizing the specific characteristics of our synthetic dataset.

C. Analysis Details

Our block-wise analysis, presented in the main paper, was conducted on a test set of 1000 synthetic videos generated using the same pipeline as our training data but with held-out foreground and background clips. To obtain the effect mask $\boldsymbol{M}^e$ used in the analysis, we isolate the rendered shadow region from our synthetic data generation process. This provides a ground-truth spatial map of where the associated effect is located, allowing us to quantitatively measure each block's sensitivity to effect-related features. The whole pipeline is illustrated in Fig. 11.

D. Evaluation Protocols

Video Reconstruction Quality. This protocol assesses the fidelity of the decomposition. We take a set of P videos to carry out this experiment. The predicted foreground layer $\hat{\boldsymbol{F}}$ is composited back onto the predicted background layer using the predicted alpha matte $\hat{\alpha}$. The quality is measured by comparing this reconstructed video with the original input video $\mathbf{V}$. A high-quality decomposition should allow for a near-perfect reconstruction.

We use three standard metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Warping Loss to calculate temporal misalignments.

Background Composition Plausibility. This protocol evaluates how well the separated foreground layer can be composited onto novel, unseen backgrounds. We take a set of P predicted foreground layers and compose them with a set of Q different background videos. The key is that the separated foreground layer must be clean and free of artifacts from its original background to ensure a seamless new composition.

We use the Fréchet Video Distance (FVD) to quantitatively measure the quality of the newly generated videos. We compute the FVD between the set of our composited videos and the set of original background videos. A lower FVD score indicates that the distribution of the composited videos is closer to that of real videos, suggesting a higher-quality and more plausible decomposition.

We present cases in Fig. 12 to demonstrate the effectiveness of our method in the above experiments, where P and Q are 40 and 200, respectively. Our method demonstrates superiority in preserving the completeness and fine-grained details of the foreground. More critically, it possesses the capability to capture clean and intact associated effects, free from background artifacts.

Human Evaluation. To complement our quantitative metrics with a qualitative assessment of perceptual quality, we conducted a comprehensive user study. We recruited (28 as described in our main paper) participants with backgrounds in computer graphics or vision. The study was designed to compare our method against 4 leading baselines on a set of 20 challenging video sequences featuring a variety of objects and associated effects.

For each video sequence, participants were presented with a side-by-side comparison of the results from all methods, displayed in a randomized order to prevent bias. Participants were first briefed on the video decomposition task, and then they were shown how a clean separation is crucial for plausible compositing onto novel backgrounds.

Participants were asked to rate the quality of each method's output based on three predefined criteria, which were carefully explained to them beforehand:

1. **Foreground Integrity:** This criterion assesses the completeness and color fidelity of the main object, emphasiz-

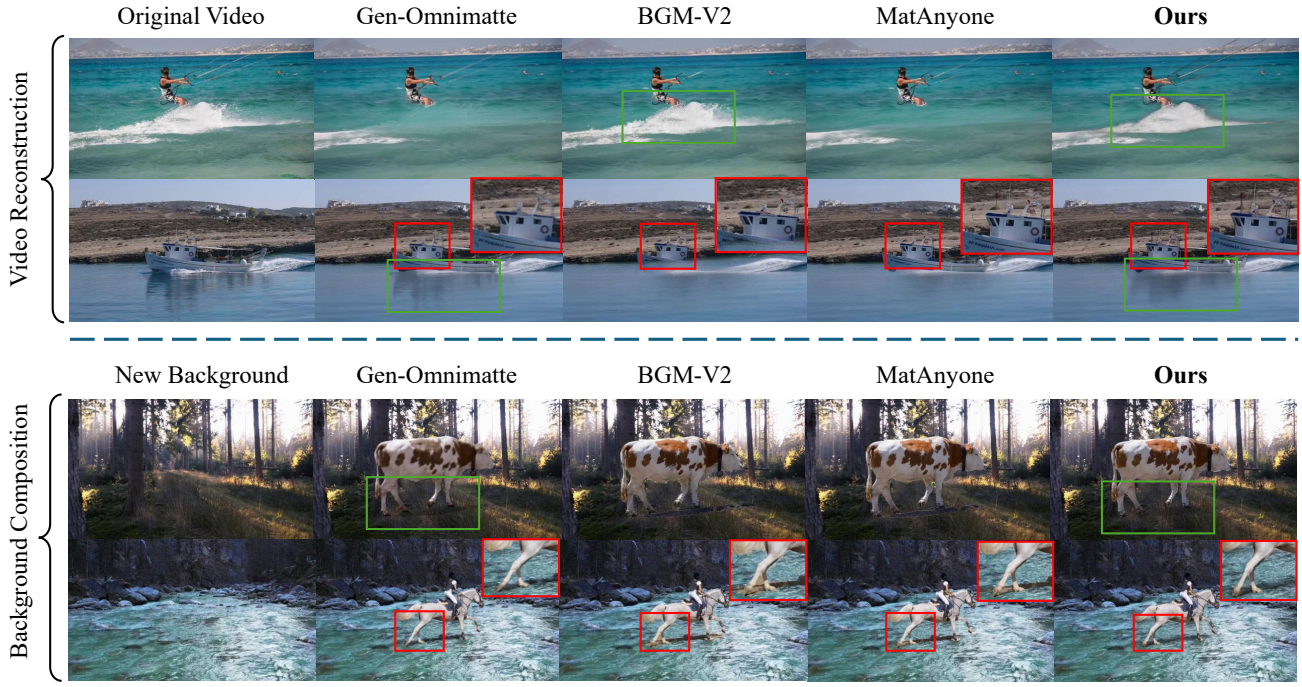

Figure 12. **Visualization of the Video Reconstruction and Background Composition.** The excellent performance of our method in both experiments stems directly from its superior ability to separate the foreground from its effects with high fidelity. We highlight correctly preserved effects in green boxes and magnify challenging details in red to demonstrate this capability. Zoom in for a better view.

ing the absence of background color bleeding.

2. **Effect Harmony:** This focuses on the integrity of the secondary effects, evaluating both their completeness and the plausibility of their rendered transparency.
3. **Temporal Consistency:** This evaluates the temporal consistency and aesthetic quality of the final decomposition, penalizing artifacts such as temporal flickering, jagged boundaries, or other visual instabilities.

For each criterion, participants provided a score on a 6-point scale, ranging from *0 (very poor)* to *5 (excellent)*. The final scores for each method were then averaged across all participants and video sequences to compute a **Overall Score** for each method aspect.

Method	Inference time		Peak VRAM	User Study Overall Score
	Stage 1	Stage 2		
MatAnyone	9.5s	6.2s	1915 mb	2.82
BGM-V2	4.6s	2.7s	12453 mb	2.26
Gen-Omni.	4.6s	389.2s	12453 mb	2.85
EasyOmni.	12.7s		24117 mb	4.08

Table 3. Comparison of training and inference cost.

Figure 13. **Importance of Inpainting Pre-training.** Fine-tuning a general video model (WAN Fun) instead of an inpainting model leads to performance degradation. The Quality Expert (top) exhibits severe color bleeding artifacts in the alpha matte. The Effect Expert (bottom) fails entirely to capture shadows. This ablation confirms that an inpainting foundation is critical, especially for capturing associated effects.

E. Training and Inference cost

Our models are trained on two H100 GPUs, requiring roughly 3 hours for the Effect Expert and 16 hours for the Quality Expert. As presented in the Tab. **??**, our ap-

proach demonstrates substantially faster inference than GenOmnimatte, matching the speed of typical video matting pipelines. Notably, the independent full inference time for the Effect Expert and the Quality Expert is 11.8s each, whereas the joint inference with Dual Expert Sampling only marginally increases the total time to 12.6s.

F. More Ablation Studies

In this section, we investigate the importance of the inpainting training of our base model. Specifically, we conduct an ablation study by fine-tuning the general-purpose conditional video generation model, WAN2.1 Fun, instead of our specialized video inpainting model. The primary goal is to determine whether the inherent knowledge of foreground removal is a prerequisite for successfully training a foreground decomposition model with our proposed strategy. The WAN2.1 Fun model shares the same input scheme with our inpainting base model but is trained for general conditional generation, not explicitly for object removal.

As shown in Fig. 13, we present the results of this ablation by training the WAN2.1 Fun model with our two expert modules:

- **Training as a Quality Expert:** When fine-tuned to predict the alpha matte, we observe that the model can successfully learn to transfer to this new domain and generate a coarse alpha matte. However, the output suffers from significant artifacts, most notably a tendency to leak colors from the original video directly into the alpha matte, corrupting its purity. We hypothesize that this issue could potentially be alleviated with a much larger training dataset and extended training steps, but it highlights a fundamental difficulty for a general model to learn this task.
- **Training as an Effect Expert:** When trained to capture associated effects, the model completely fails to acquire this capability. The fine-tuned model does not learn to identify or isolate effects like shadows or reflections, indicating that this skill does not emerge naturally from a general video generation prior.

These experiments collectively demonstrate that under our current training strategy, fine-tuning a general generative model can roughly predict a primary alpha matte but is incapable of cleanly separating its associated effects. This deficit serves as strong evidence that the pre-training objective of *object removal* is decisively important.

G. Failure Cases

Our method's performance is inherently dependent on the capabilities of the underlying inpainting model. As illustrated in Fig. 14, we observe two typical failure modes induced by this limitation:

- **Failure to Extract Effects:** If the inpainting model fails to perceive an associated effect, it will treat that region as part of the true background. Consequently, our method cannot separate this effect.
- **Failure to Reconstruct Foreground:** In cases where the inpainting model struggles to inpaint the object that is heavily occluded, its internal features for that region may become corrupted. This can lead to our method producing only a distorted foreground layer, as the decomposition is derived from these flawed features.

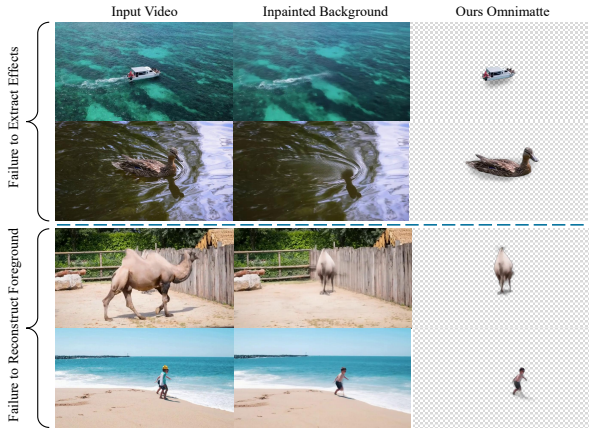

Figure 14. **Failure Cases.** Since our method is built upon an inpainting model, it naturally fails when the underlying inpainting model itself fails.

We believe that these limitations will be mitigated as the underlying video inpainting models become more powerful in their removal and completion capabilities.