Towards Explainable Video Camouflaged Object Detection: SAM2 with Eventstream-Inspired Data

Anonymous submission

Abstract

Video Camouflaged Object Detection (VCOD) poses significant challenges due to the subtle appearance of camouflaged objects, especially under dynamic motion and occlusion. Existing methods predominantly rely on optical flow or black-box features for motion modeling, which often entail substantial computational costs and suffer from limited interpretability. Inspired by the human strategy of identifying abnormal movements between frames and the principle of event camera image formation, we propose an eventstream-inspired dual-branch framework for VCOD. Specifically, we design an eventstream-inspired data extraction module to capture pixel-level motion variations, effectively distinguishing object motion from background dynamics. This event-based representation is integrated into SAM2 through a dual-branch memory-augmented framework, consisting of Time Bridge Attention and Visual Bridge Attention, enabling joint modeling of motion and appearance cues. In addition, we introduce a Prompt Embedding Generator to eliminate the need for human-provided interactive prompts, facilitating fully automatic VCOD. Extensive experiments on MoCA-Mask and CAD2016 demonstrate that our approach significantly outperforms state-of-the-art methods, achieving both superior segmentation accuracy and interpretable motion modeling. To the best of our knowledge, this is the first work to incorporate eventstream-inspired representations into the VCOD task. Code and related resources will be released.

1 Introduction

Camouflage, a sophisticated survival strategy widely observed in nature, enables organisms to blend seamlessly into their surroundings and avoid detection by predators (Stevens and Merilaita 2009). This biological phenomenon has inspired the computer vision community to formulate the Camouflaged Object Detection (COD) problem, which aims to detect objects with minimal visual contrast against their backgrounds. Compared to conventional object detection and segmentation tasks, COD presents unique challenges due to its inherent visual ambiguity, where targets are intentionally designed to remain imperceptible. Beyond its scientific significance for understanding visual perception, COD has promising real-world applications, including industrial quality inspection (Fan et al. 2023) and medical image analysis (Fan et al. 2020b).

Recently, research interest has shifted toward Video Camouflaged Object Detection (VCOD), which extends COD to dynamic scenes and introduces new challenges due to temporal complexity. Videos often involve object motion, background variations, and frequent occlusions that may cause temporary object disappearance. Nevertheless, camouflaged objects typically exhibit temporally consistent and predictable motion patterns. Human observers performing VCOD tasks tend to rely on motion cues—systematically comparing adjacent frames to identify subtle differences that static appearance alone fails to reveal. Inspired by this human strategy, we introduce an eventstream-inspired representation that explicitly encodes pixel-level inter-frame variations, serving as a motion-sensitive input. This design enables our model to focus on subtle but temporally coherent changes that are essential for localizing camouflaged objects in complex video scenarios.

To further enhance the generalization and robustness of our approach, we build upon the powerful visual priors provided by the Segment Anything Model (SAM) series (Kirillov et al. 2023). However, directly applying SAM

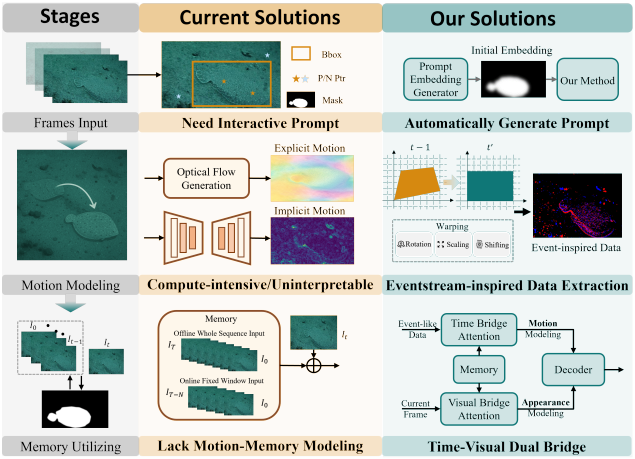

Figure 1: Comparison of existing solutions and our method across three VCOD stages: frame input, motion modeling, and memory utilization. Our approach removes manual prompts, leverages interpretable eventstream motion modeling, and integrates motion cues into long-term memory.

to VCOD is non-trivial due to its dependence on interactive prompts (e.g., points, bounding boxes, or masks) for segmentation initialization (Meeran, Mantha et al. 2024). This limitation arises for two reasons: (1) in camouflaged scenes, even human annotators may struggle to identify objects without temporal context, making prompt-based initialization unreliable; and (2) prompt-based strategies are impractical for large-scale video processing as they require manual intervention for each sequence. Moreover, existing VCOD frameworks often rely on memory mechanisms that primarily aggregate appearance cues while underutilizing motion information, which is crucial for achieving temporally consistent and robust segmentation in scenarios with fast motion, occlusion, or background clutter.

To address these limitations, we propose a unified and interpretable framework specifically designed for VCOD, as illustrated in Fig. 1. The framework consists of three key components: (1) a **Prompt Embedding Generator (PEG)** that autonomously generates informative prompts, enabling fully automatic segmentation while leveraging SAM's generalization capabilities; (2) an **Eventstream-Inspired Motion Modeling** module, inspired by event cameras, which captures pixel-level intensity changes to provide interpretable and reliable motion representations; and (3) a **Dual-Branch Memory-Augmented Framework** that integrates both appearance and motion information into long-term memory for temporally consistent predictions.

Comprehensive experiments on two widely used VCOD benchmarks, MoCA-Mask and CAD2016, demonstrate that our method not only achieves state-of-the-art performance but also improves interpretability by explicitly modeling motion cues.

Our main contributions are as follows:

- We develop a prompt-free online VCOD framework by introducing a Prompt Embedding Generator that eliminates the need for interactive prompts while retaining SAM's strong visual priors.
- We propose an interpretable eventstream-inspired motion modeling strategy that captures pixel-level motion variations, enhancing the detection of camouflaged objects.
- We design a dual-branch memory augmentation mechanism that jointly models visual and motion cues, enabling robust and temporally coherent segmentation.
- Our method achieves new state-of-the-art performance on public VCOD benchmarks, with over 15% relative improvement compared to the second-best approaches.

2 Related Works

2.1 Video Semantic Segmentation

Video Semantic Segmentation (VSS) (Su et al. 2023) aims to assign semantic labels to each pixel across video frames while maintaining temporal coherence. Extending static image segmentation to videos is challenging due to appearance variations, motion blur, and the need for consistent temporal modeling. Early approaches relied on optical flow (Tokmakov, Alahari, and Schmid 2017) to estimate pixel-level motion, while more recent methods utilize memory-based (Oh et al. 2019) and transformer-based architectures (Yang et al. 2022) to achieve improved temporal consistency.

The Segment Anything Model (SAM) (Cheng et al. 2023) demonstrates impressive generalization through prompt-based interactive segmentation. SAM2 (Ravi et al. 2024) extends this concept to videos by incorporating temporal memory propagation, while recent adaptations (Chen et al. 2024) explore fine-grained tasks such as camouflaged object detection (Fan et al. 2020a). However, SAM-based approaches remain fundamentally dependent on manual prompts, which hinders scalability for fully automatic video segmentation.

2.2 Eventstream-inspired Object Detection

Event cameras (Lichtsteiner, Posch, and Delbruck 2008) asynchronously record pixel-level intensity changes, generating event streams that provide high-temporal-resolution motion cues with minimal latency. Event-based representations have been successfully applied to video object tracking (Iaboni et al. 2021), segmentation (Jiang, Moreau, and Davoine 2024), and applications like gaze estimation (Li, Chang, and Raychowdhury 2024).

Despite their advantages, traditional motion modeling techniques—whether based on optical flow or implicit temporal aggregation—are often computationally expensive and lack interpretability. Event-based methods provide fine-grained, noise-resilient motion cues but face challenges due to the absence of large-scale datasets specifically tailored for camouflaged object detection (COD). This gap limits their direct application to VCOD tasks.

2.3 Image/Video Camouflaged Object Detection

COD has evolved from early handcrafted features (Hou and Li 2011) to deep learning-based frameworks that leverage visual attention (Mei et al. 2021) and auxiliary cues like edges (Lyu et al. 2023) or depth (Wu et al. 2023). Recent works also explore camouflaged-specific attributes (Zhang et al. 2024) and generative modeling (Zhang et al. 2025) to improve detection performance.

Extending COD to videos (VCOD) introduces challenges such as motion blur, occlusion, and object disappearance. Existing methods typically rely on appearance-based memory (Yang et al. 2021), which underutilizes fine-grained motion cues and appearance-motion interactions essential for temporally consistent segmentation.

3 Methods

3.1 Overview

An overview of our proposed framework is illustrated in Fig. 2. We first introduce the motivation behind our design, followed by a detailed description of each module.

Motivation. Humans are highly sensitive to biological motion and can recognize actions even from sparse or degraded visual information (Johansson 1973). This ability is supported by the hierarchical structure of the human visual cortex and the specialized pathways for motion perception (Van Essen and Maunsell 1983). In the context of VCOD, humans tend to identify camouflaged targets by

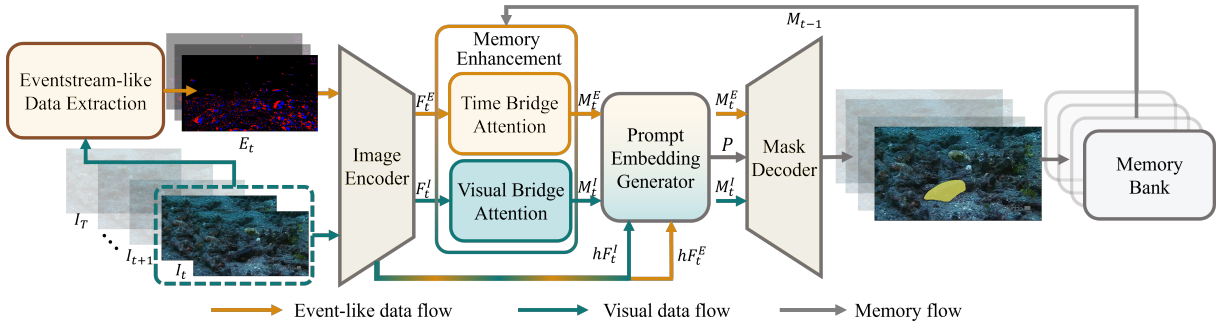

Figure 2: Overview of our framework with two branches: an eventstream-inspired branch for motion modeling and a visual branch for appearance. Both interact with memory through bridge attention and generate adaptive prompts for mask prediction.

carefully comparing adjacent frames to discover subtle and abnormal motion patterns that deviate from the regular background dynamics. This observation motivates us to develop a framework capable of explicitly modeling motion information and its interaction with long-term memory, thereby enabling a comprehensive understanding of camouflaged object behavior throughout the video.

Network Overview. We propose a dual-branch VCOD framework built on SAM2 for online video processing (Fig. 2). The framework comprises:

(1) The *eventstream-inspired branch* (yellow arrows), which models motion by extracting an eventstream-inspired representation E_t from consecutive frames I_{t-1} and I_t (Sec. 3.2). This representation is encoded into features F_t^E, which interact with the memory bank via the *Time Bridge Attention* module to retrieve motion-relevant information M_t^E and generate prompts.

(2) The *Visual branch* (green arrows), which processes the current frame I_t to obtain visual features F_t^I that are refined through *Visual Bridge Attention* for appearance consistency.

The outputs from both branches, including high-resolution features (hF_t^E, hF_t^I), are fused and passed to the *Prompt Embedding Generator* (Sec. 3.4) to produce adaptive prompts. The *Mask Decoder* then uses the prompt P_t ($t = 0$ for initialization), motion-enhanced features M_t^E, and appearance-enhanced features M_t^I to predict the segmentation mask, which is stored in the memory bank along with an occlusion prediction. This design yields a compact yet effective VCOD framework with interpretable motion-appearance modeling.

3.2 Eventstream-Inspired Data Acquisition

Extracting reliable eventstream-inspired data is fundamental for constructing interpretable and effective motion representations in VCOD. In dynamic scenes, apparent motion arises from both camouflaged foreground objects and global effects such as background changes or camera movement, which can easily overshadow subtle object motion. To address this, we employ homography-based global motion compensation, which aligns frames and removes large-scale motion caused by camera movement. This step is simple, efficient, and interpretable, and it significantly enhances motion discriminability by isolating object-induced cues.

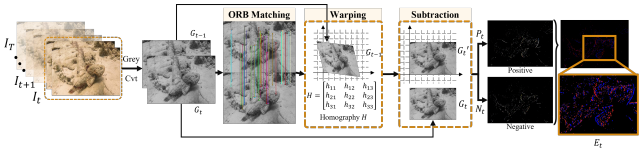

Figure 3: eventstream-inspired Data Extraction Pipeline. The pipeline compensates for global motion and highlights local motion residuals to obtain eventstream-inspired representations from adjacent frames. (zoom in for details)

Although the planar assumption of homography is an approximation, it is highly effective in typical VCOD scenarios—such as top-down surveillance, long-range imaging, underwater videos, and natural habitats—where background depth variations are relatively small. By warping adjacent frames with an estimated homography, we suppress camera-induced motion, thereby exposing the fine-grained residuals corresponding to moving camouflaged objects.

As illustrated in Fig. 3, given two consecutive RGB frames I_{t-1} and I_t, we first convert them into grayscale images G_{t-1} and G_t to suppress color noise and reduce computational complexity. A homography matrix H is then estimated using ORB-based feature matching, with RANSAC employed to reject outliers and ensure robustness. The previous frame G_{t-1} is subsequently warped by H to obtain the globally aligned frame G'_t. The motion residual map is computed as:

$$D_t = G_t - G'_t. \tag{1}$$

To extract meaningful motion signals from D_t, we apply an adaptive thresholding scheme based on local Gaussian statistics, generating positive and negative motion maps:

$$\begin{aligned} P_t(u,v) &= 1[D_t(u,v) > \tau_t(u,v)], \\ N_t(u,v) &= 1[D_t(u,v) < -\tau_t(u,v)], \end{aligned} \tag{2}$$

where $\tau_t(u,v)$ is a spatially adaptive threshold and $1[\cdot]$ denotes the indicator function.

The eventstream-inspired representation E_t is then formed by projecting P_t and N_t onto RGB channels:

$$E_t^1(u,v) = P_t(u,v)\cdot R_t(u,v), \quad E_t^2(u,v) = N_t(u,v)\cdot B_t(u,v), \tag{3}$$

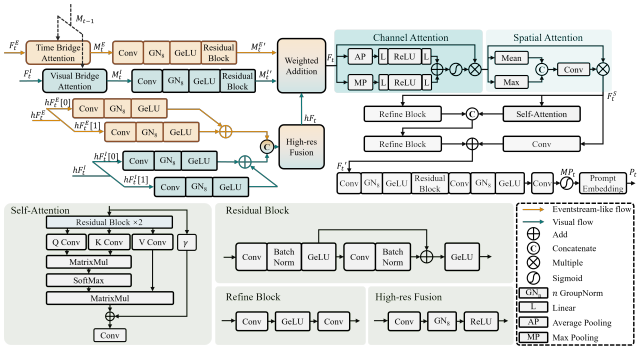

Figure 4: Architecture of the Prompt Embedding Generator (PEG). It fuses event, image, memory, and high-resolution features to generate a feature-level prompt embedding P_t for segmentation. (zoom in for details)

with $E_t = \text{stack}[E_t^1, E_t^2]$. Finally, morphological operations (e.g., erosion and dilation) are applied to eliminate noise and refine motion regions. The resulting E_t provides an interpretable, motion-sensitive representation, which serves as the input to our eventstream branch. Additional visualizations and implementation details are available in Supplementary Materials (Sec. 1 and Sec. 2).

3.3 Memory Enhancement Attention

The Memory Enhancement Attention module jointly models motion and appearance cues by interacting with the memory bank. It includes two submodules: *Time Bridge Attention* and *Visual Bridge Attention*, denoted as $\mathcal{A}_{\text{Time}}$ and $\mathcal{A}_{\text{Visual}}$. Given the eventstream feature F_t^E and visual feature F_t^I, these modules incorporate information from the previous memory state M_{t-1} to produce enhanced features:

$$M_t^E = \mathcal{A}_{\text{Time}}(F_t^E, M_{t-1}), \quad M_t^I = \mathcal{A}_{\text{Visual}}(F_t^I, M_{t-1}). \tag{4}$$

Both attention modules share the same architecture with independent parameters. The general form of $\mathcal{A}$ is:

$$\mathcal{A}(F, M) = \text{MLP}\left(\text{CA}(\text{SA}(F), M) + \text{D}(\text{SA}(F))\right), \tag{5}$$

where $\text{SA}(\cdot)$ is a self-attention block:

$$\text{SA}(F) = \text{SA}(\text{LN}(F)) + \text{Drop}(F). \tag{6}$$

Here, $\text{CA}(\cdot, M)$ is cross-attention conditioned on memory M, $\text{D}(\cdot)$ is a feed-forward layer, and $\text{MLP}(\cdot)$ a multi-layer perceptron. This design aligns and refines motion and appearance features through temporal memory interactions.

3.4 Prompt Embedding Generator

The Prompt Embedding Generator (PEG) eliminates the need for explicit user prompts (e.g., points, boxes) by converting motion and appearance cues into a dense feature-level embedding $P_t \in \mathbb{R}^{C\times H\times W}$. PEG integrates eventstream features F_t^E, visual features F_t^I, and memory-enhanced outputs M_t^E and M_t^I from Time and Visual Bridge Attentions.

First, M_t^E and M_t^I are refined via a residual enhancement block $\mathcal{R}(\cdot)$:

$$M_t^{E'} = \mathcal{R}(M_t^E), \quad M_t^{I'} = \mathcal{R}(M_t^I). \tag{7}$$

To retain fine details, high-resolution features are extracted and fused:

$$hF_t = \mathcal{F}_{\text{high-res}}(\mathcal{H}(F_t^E), \mathcal{H}(F_t^I)). \tag{8}$$

The refined memory and high-resolution features are combined via weighted addition:

$$F_t = \mathcal{F}_{\text{add}}(M_t^{E'}, M_t^{I'}, hF_t). \tag{9}$$

Attention modules enhance discriminative capacity, where channel and spatial attention highlight key regions:

$$F_t^S = \mathcal{A}_{\text{Spatial}}(\mathcal{A}_{\text{Channel}}(F_t)). \tag{10}$$

Finally, We generate the prompt embedding feature by concatenating outputs, refining them, and applying a convolutional projection:

$$F_t' = \text{Add}\left(\text{Concat}\left(\mathcal{A}_{\text{Self}}(F_t^S),\ \text{Refine}(F_t^S)\right),\ \text{Conv}(F_t^S)\right). \tag{11}$$

The mask prediction MP_t is produced by a lightweight segmentation head with residual and convolutional blocks:

$$MP_t = \sigma \circ \text{Conv}_2 \circ \text{GeLU} \circ \text{GN}_8 \circ \text{Conv}_1 \circ \mathcal{R}(F_t'), \tag{12}$$

where σ is the sigmoid activation. The predicted mask MP_t is then encoded into the final prompt embedding P_t, which guides the mask decoder. Additional visualizations and intermediate results are shown in Sec. 4.3.

3.5 Optimization Objective

We optimize the proposed network by minimizing the following joint loss function:

$$L(t) = \alpha L_{\text{emb}} + \beta L_{\text{mask}} + 20 L_{\text{focal}} + L_{\text{Dice}}, \tag{13}$$

where L_{emb} is the structure-aware loss on the PEG-generated prompt embedding, combining weighted binary cross-entropy and IoU losses as in (Wei, Wang, and Huang 2020). L_{mask} is the segmentation loss between the predicted and ground-truth masks. We also adopt L_{focal} and L_{Dice} from SAM2 (Ravi et al. 2024), with weights 20 and 1, respectively. The hyperparameters α and β are set to 1 and 0.5, with further analysis in Supplementary Materials Sec. 3.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate our method on two VCOD benchmarks: MoCA-Mask (Cheng et al. 2022) and CAD2016 (Bideau and Learned-Miller 2016), following standard protocols. MoCA-Mask contains 87 videos (22,939 frames) of camouflaged animals in natural scenes, with 71 sequences for training and 16 for testing. CAD2016 is a

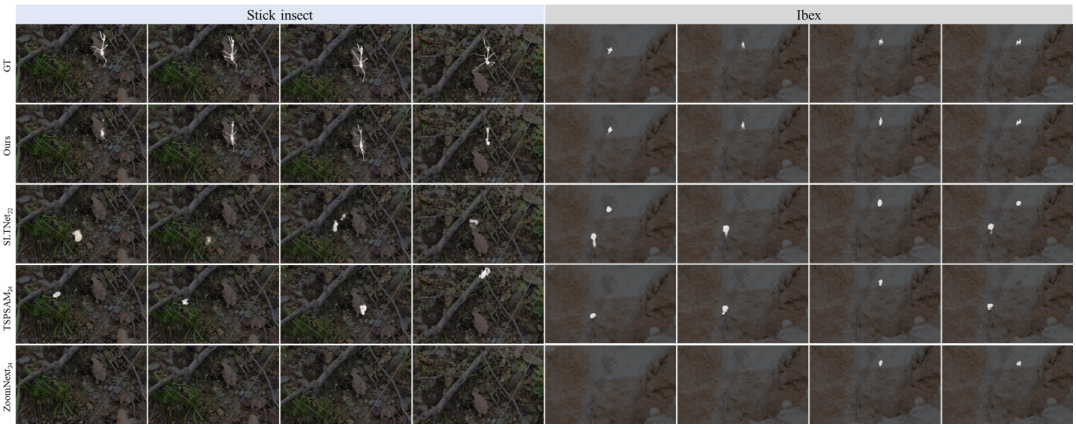

Figure 5: Qualitative results on MoCA-Mask for two challenging scenarios. From top to bottom: Ground Truth (GT), Ours, SLTNet, TSP-SAM, and ZoomNeXt. Each column corresponds to a different frame index. (zoom in for details)

smaller dataset of 9 YouTube clips, with pixel-wise masks annotated on every fifth frame.

Training. We adopt a two-stage strategy: (1) fine-tune SAM2 on COD10K (Fan et al. 2020a); (2) fine-tune on MoCA-Mask using only the additional modules of our framework. We exclude MoCA-Mask due to their low quality during fine-tuning and achieve strong performance. Additionally, we observe that directly mixing COD10K and MoCA-Mask during training leads to performance degradation. We thus follow this two-stage training scheme. See Supplementary Materials Sec. 4 for detailed training step discussion.

Evaluation Metrics. We use six standard metrics: structure measure (S_α) (Fan et al. 2017), weighted F-measure (F_β^ω) (Margolin, Zelnik-Manor, and Tal 2014), enhanced alignment (E_ϕ) (Fan et al. 2018), mean absolute error (MAE), mean Dice (mDice), and mean IoU (mIoU). Higher S_α, F_β^ω, E_ϕ, mDice, and mIoU, and lower MAE indicate better performance.

Implementation Details. All experiments are implemented in PyTorch on NVIDIA RTX 8000 GPUs. Input frames are resized to 1024×1024 to match SAM2. Data augmentation includes random flips, color jitter, and grayscale conversion. Both training stages use a cosine learning rate schedule with a 0.1 decay. In the first stage (SAM2 fine-tuning), the image encoder uses 3×10^{-6} and other modules 5×10^{-6}. In the second stage (video fine-tuning), the embedding generator and bridge attention modules use 5×10^{-4}.

4.2 Comparison with State-of-the-Art Methods

Qualitative Results We qualitatively evaluate our method on two challenging MoCA-Mask scenarios in Fig. 5: (1) a stick insect hidden in dense weeds and (2) an ibex camouflaged against a rocky cliff. The stick insect's fine structures and slender body cause other methods (SLTNet, TSP-SAM, ZoomNeXt) to misidentify or miss the object entirely, while our approach accurately recovers details such as antennae and limbs. In the ibex case, the small target size further challenges existing methods, which fail to detect it reliably. In contrast, our method successfully localizes and reconstructs the object, demonstrating strong capability in capturing subtle, small-scale camouflage. Additional qualitative results on CAD2016 are included in Supplementary Sec. 5.

Quantitative Results Tab. 1 compares our method with nine state-of-the-art approaches on MoCA-Mask and CAD2016 using six metrics. Our approach achieves the best performance across most metrics, with notable improvements of 20.4% in F_β^ω, 15.5% in mDice, and 17.5% in mIoU on MoCA-Mask. On CAD2016, we observe similar gains, including a 5.9% increase in S_α and over 20% improvements in F_β^ω, mDice, and mIoU.

Despite being an online method using only historical frames, we outperform all offline baselines, demonstrating strong robustness. The higher absolute scores on CAD2016 stem from its larger, less challenging objects, unlike MoCA-Mask's smaller targets (see Supplementary Sec. 6). While our MAE is not the lowest, our method excels in structure-aware metrics (S_α, F_β^ω, E_ϕ) due to sharper object boundaries, yielding better object-level segmentation.

4.3 Ablation Study

Effectiveness of Proposed Modules Tab. 2 presents the ablation results verifying the contributions of each component. Compared to the baseline SAM2, which is fine-tuned using a blank mask, the proposed prompt generation module provides adaptive and informative guidance by automatically generating frame-wise prompts instead of relying on fixed zeros, leading to consistent improvements across all metrics. The introduction of eventstream-inspired data further enriches the representation with motion-sensitive cues extracted from adjacent frames, enhancing spatial and temporal modeling. However, as it lacks explicit temporal reasoning, its benefit is significantly boosted when combined with the Time Bridge module. The Time Bridge explicitly models motion-memory interactions, leading to substantial gains, especially in F_β^ω (over 50% relative improvement) and notable boosts in all other metrics, underscoring its critical role in the proposed framework.

Table 1: Quantitative comparison with state-of-the-art (SOTA) methods under six commonly used metrics. The best results are marked in **bold**, and the second-best are underlined. In the "Online" column, ✓ indicates that the method only uses past frames, while ✗ denotes reliance on future frames.

Method	Pub	Online	MoCA-Mask						CAD2016					
			$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$	mDice ↑	mIoU ↑	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$	mDice ↑	mIoU ↑
RCRNet	ICCV'19	✗	.597	.174	.583	.025	.194	.137	†	†	†	†	†	†
PNS-Net	MICCAI'21	✗	.576	.134	.536	.038	.189	.133	.678	.369	.720	.043	.409	.309
MG	ICCV'21	✓	.547	.165	.537	.095	.197	.137	.613	.370	.537	.070	.351	.260
SLT-Net	CVPR'22	✗	.656	.357	.785	.021	.387	.310	.669	.481	.845	.030	.368	.268
IMEX	TMM'24	✓	.661	.371	.778	.020	.409	.319	.684	.452	.813	.033	.469	.370
TSP-SAM(M+P)	CVPR'24	✓	.673	.400	.766	.012	.421	.345	.705	.565	.836	.027	.591	.422
TSP-SAM(M+B)	CVPR'24	✓	.689	.444	.808	**.008**	.458	.388	.751	.628	.865	.021	.603	.496
ZoomNeXt(T=1)	TPAMI'24	✓	.690	.395	.702	.017	.420	.353	.721	.525	.759	.024	.523	.436
ZoomNeXt(T=5)	TPAMI'24	✗	.734	.476	.736	.010	.497	.422	.757	.593	.865	**.020**	.599	.510
EMIP	TIP'25	✓	.669	.374	†	.017	.424	.326	.710	.504	†	.029	.528	.415
EMIP-L	TIP'25	✓	.675	.381	†	.015	.426	.333	.719	.514	†	.028	.536	.425
Ours	—	✓	**.753**	**.573**	**.855**	.009	**.574**	**.496**	**.802**	**.717**	**.887**	.023	**.717**	**.615**

Table 2: Ablation study on MoCA-Mask to evaluate the effectiveness of each proposed module.

Prompt Gen	Eventstream	Time Bridge	$S_\alpha \uparrow$	$F_\beta^\omega \uparrow$	$E_m \uparrow$	$M \downarrow$	mDice ↑	mIoU ↑
			0.630	0.328	0.673	0.022	0.337	0.303
✓			0.640	0.336	0.713	0.017	0.341	0.308
✓	✓		0.665	0.381	0.773	0.013	0.390	0.343
✓		✓	0.715	0.513	0.813	0.012	0.514	0.437
✓	✓	✓	**0.753**	**0.573**	**0.855**	**0.009**	**0.574**	**0.496**

Intermediate Results Visualization To further demonstrate the effectiveness of our method, we visualize several intermediate results in Fig. 6 using two representative sequences. In the initial frame ($T = 0$), both the mask prompt and the dense embedding prompt align well with the ground truth (third row of the first and second columns of each instance), effectively removing the need for human-interactive prompts in the VCOD task. From $T = 25$ to $T = 125$, memory-enhanced visual features maintain consistent alignment with ground truth masks, illustrating the effectiveness of the memory refinement strategy.

Notably, in challenging occlusion cases such as $T = 50$ in *Arctic Fox* and $T = 50$, $T = 125$ in *Black Cat*, our method successfully recovers occluded objects, which is extremely challenging in conventional image-based COD (Zhang et al. 2024; Lyu et al. 2023). This ability is clearly reflected in both the predicted masks and memory-refined features. By introducing memory enhancement, our model effectively leverages historical object representations to handle occlusions, resulting in more reliable feature utilization and superior mask predictions under complex camouflage conditions.

Table 3: Performance comparison between our eventstream-inspired data and ESIM-based event-camera synthesis.

Method	$S_\alpha \uparrow$	$F_\beta^\omega \uparrow$	$E_m \uparrow$	$M \downarrow$	$D \uparrow$	IoU ↑	Acq FPS↑
w/ ESIM	.722	.527	.833	.012	.524	.449	5.73
Ours	.753	.573	.855	.009	.574	.496	35.65

Comparison with Event-Camera Synthesis Data To further validate the effectiveness of our eventstream-inspired data, we compare it with the widely used event-camera synthesis method ESIM (Rebecq, Gehrig, and Scaramuzza 2018), focusing on both performance and data acquisition speed (Table 3). We integrated ESIM (bin size 0.333s to match image capture FPS) and aligned its outputs with ours for a fair evaluation. ESIM runs at only 5.73 FPS versus our 35.65 FPS and shows a clear performance drop across all metrics. We attribute this to three factors: (1) lack of global motion compensation, (2) temporal resolution loss from event downsampling, and (3) synthetic noise, which is especially harmful in low-contrast VCOD scenarios. This comparison underscores the superiority of the proposed eventstream-inspired data.

Table 4: Comparison of model efficiency: parameters (Params), speed (FPS), and memory usage (Mem).

Method	Pub	Params (M)↓	FPS ↑	Mem (GB)↓
FSPNet	CVPR'23	274.24	1.56	4.47
TSP-SAM	CVPR'24	725.42	1.27	5.99
SAM-PM	CVPRW'24	313.33	2.34	4.55
Ours	–	98.10	2.56	3.78

Computation Efficiency We conduct a comprehensive analysis of computational efficiency in this section, with all inference performed on a single NVIDIA RTX 8000 GPU. As shown in Tab. 4, our method achieves the best segmentation performance while requiring significantly fewer parameters, offering higher frames per second (FPS), and consuming less memory compared to existing approaches. Specifically, our method utilizes 23.5 million trainable parameters and 74.6 million non-trainable parameters during second-

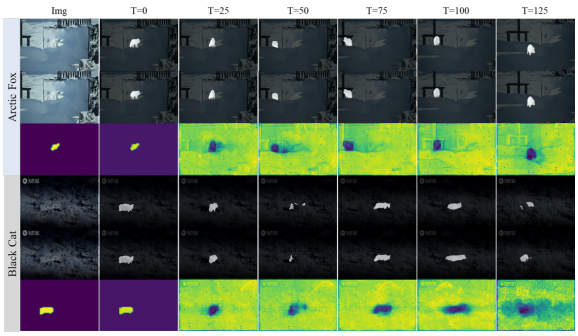

Figure 6: Visualization of temporal progression and memory enhancement. For each sequence, the first row shows input frames and ground truth masks sampled every 25 frames. The second row presents our predicted masks at the corresponding time steps. The third row visualizes the evolution of the prompt and feature representations, including the initial mask prompt, the dense embedding prompt after the prompt encoder, and the memory-refined features from $T = 25$ to $T = 125$. This visualization demonstrates that prompt prediction effectively guides the mask decoder, while memory refinement progressively enhances feature representations and ensures temporal consistency.

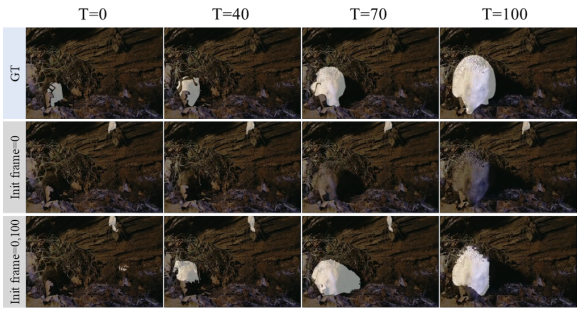

Figure 7: Failure case on a camouflaged hedgehog. Prompting with only frame 0 mis-segments the background, while adding frame 100 reveals corrects object.

stage training, making it substantially more lightweight than FSPNet (274.24M), TSP-SAM (727.1M total, including 89.75M trainable), and SAM-PM (313.33M), while being comparable in size yet outperforming SLTNet (82.41M) in terms of performance (Supplementary Sec. 7). Due to the unavailability of ZoomNeXt's video inference code, we adopt FPSNet as a reproducible baseline for comparison. Note that "online" refers to the causal setting in VCOD, where only current and past frames are used, unlike "offline" methods that access future frames. This does not imply real-time inference capability.

Failure Case and Analysis As shown in Fig. 7, when using frame 0 as the initial condition frame, the model incorrectly segments a background region due to the hedgehog being heavily occluded by leaves and hidden in a dark cave, making it difficult to distinguish even for humans. To address this, we experimented with introducing an additional initial frame at frame 100, where the hedgehog is more exposed. As seen in the third row, this adjustment significantly alleviates the mislocalization issue. However, this strategy is highly dependent on video content and temporal dynamics, which vary across scenes. For fairness and generalization, we consistently use frame 0 as the sole condition frame for all quantitative and qualitative evaluations.

Why Eventstream Features? Our eventstream-inspired representation encodes temporal intensity variations as

$$E_t(u, v) = I_t(u, v) - I_{t-1}(u, v), \tag{14}$$

where global camera motion is compensated via homography alignment:

$$D_t = G_t - G'_t. \tag{15}$$

This residual effectively isolates object-induced motion from background dynamics, which is crucial for detecting camouflaged objects with low appearance contrast.

Unlike optical flow, which enforces the brightness constancy constraint and estimates full 2D motion fields, our approach directly exploits temporal intensity changes E_t as event-like motion cues. Eventstream features are polarity-aware, emphasize sparse motion boundaries and directions, and provide a higher signal-to-noise ratio for subtle motion patterns. They are computationally efficient and less sensitive to global brightness variations due to homography compensation and adaptive thresholding. Theoretically, E_t approximates the temporal derivative:

$$E_t(u, v) \approx \frac{\partial I(u, v, t)}{\partial t}, \tag{16}$$

offering complementary motion cues to static appearance features and enabling our dual-branch framework to capture fine motion signals essential for VCOD.

Future Work Our study points to several promising directions for future research. First, as noted in Sec. 4.3, segmentation quality is highly sensitive to the initial frame, motivating the design of adaptive frame selection mechanisms that identify the most informative frame(s) either offline or via online self-correction during inference. Second, we aim to incorporate real eventstream data into VCOD to provide richer motion cues and improve robustness in challenging scenarios. Finally, given the limited availability of VCOD datasets (have MoCA and CAD2016 only), we plan to construct and release a dedicated dataset to advance research in this field and facilitate comprehensive evaluation.

5 Conclusion

In this work, we introduced an eventstream-inspired dual-branch framework for VCOD. Our method models motion and appearance cues jointly through an eventstream-inspired data extraction module and a memory-augmented dual-branch structure, while eliminating prompt dependency via a Prompt Embedding Generator. Experiments on MoCA-Mask and CAD2016 show that our approach achieves state-of-the-art performance with superior accuracy and interpretability. Beyond the empirical gains, our work offers new insights into motion modeling for VCOD and lays the foundation for future research exploring real event data for VCOD and other motion-sensitive video tasks.

References

Bideau, P.; and Learned-Miller, E. 2016. It's moving! a probabilistic model for causal motion segmentation in moving camera videos. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, 433–449. Springer.

Chen, T.; Lu, A.; Zhu, L.; Ding, C.; Yu, C.; Ji, D.; Li, Z.; Sun, L.; Mao, P.; and Zang, Y. 2024. Sam2-adapter: Evaluating & adapting segment anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more. *arXiv preprint arXiv:2408.04579*.

Cheng, X.; Xiong, H.; Fan, D.-P.; Zhong, Y.; Harandi, M.; Drummond, T.; and Ge, Z. 2022. Implicit motion handling for video camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13864–13873.

Cheng, Y.; Li, L.; Xu, Y.; Li, X.; Yang, Z.; Wang, W.; and Yang, Y. 2023. Segment and track anything. *arXiv preprint arXiv:2305.06558*.

Fan, D.-P.; Cheng, M.-M.; Liu, Y.; Li, T.; and Borji, A. 2017. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, 4548–4557.

Fan, D.-P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.-M.; and Borji, A. 2018. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*.

Fan, D.-P.; Ji, G.-P.; Sun, G.; Cheng, M.-M.; Shen, J.; and Shao, L. 2020a. Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2777–2787.

Fan, D.-P.; Ji, G.-P.; Xu, P.; Cheng, M.-M.; Sakaridis, C.; and Van Gool, L. 2023. Advances in deep concealed scene understanding. *Visual Intelligence*, 1(1): 16.

Fan, D.-P.; Ji, G.-P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; and Shao, L. 2020b. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, 263–273. Springer.

Hou, J. Y. Y. H. W.; and Li, J. 2011. Detection of the mobile object with camouflage color under dynamic background based on optical flow. *Procedia Engineering*, 15: 2201–2205.

Iaboni, C.; Patel, H.; Lobo, D.; Choi, J.-W.; and Abichandani, P. 2021. Event camera based real-time detection and tracking of indoor ground robots. *IEEE Access*, 9: 166588–166602.

Jiang, C.; Moreau, J.; and Davoine, F. 2024. Event-based Semantic-aided Motion Segmentation. In *International Conference on Computer Vision Theory and Applications (VISAPP)*.

Johansson, G. 1973. Visual perception of biological motion and a model for its analysis. *Perception & psychophysics*, 14: 201–211.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Li, N.; Chang, M.; and Raychowdhury, A. 2024. E-Gaze: Gaze Estimation with Event Camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lichtsteiner, P.; Posch, C.; and Delbruck, T. 2008. A 128×128 120db $15\mu s$ latency asynchronous temporal contrast vision sensor. *IEEE journal of solid-state circuits*, 43(2): 566–576.

Lyu, Y.; Zhang, H.; Li, Y.; Liu, H.; Yang, Y.; and Yuan, D. 2023. UEDG: uncertainty-edge dual guided camouflage object detection. *IEEE Transactions on Multimedia*, 26: 4050–4060.

Margolin, R.; Zelnik-Manor, L.; and Tal, A. 2014. How to evaluate foreground maps? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 248–255.

Meeran, M. N.; Mantha, B. P.; et al. 2024. SAM-PM: enhancing video camouflaged object detection using spatio-temporal attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1857–1866.

Mei, H.; Ji, G.-P.; Wei, Z.; Yang, X.; Wei, X.; and Fan, D.-P. 2021. Camouflaged object segmentation with distraction mining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8772–8781.

Oh, S. W.; Lee, J.-Y.; Xu, N.; and Kim, S. J. 2019. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9226–9235.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Rebecq, H.; Gehrig, D.; and Scaramuzza, D. 2018. Esim: an open event camera simulator. In *Conference on robot learning*, 969–982. PMLR.

Stevens, M.; and Merilaita, S. 2009. Animal camouflage: current issues and new perspectives. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1516): 423–427.

Su, J.; Yin, R.; Zhang, S.; and Luo, J. 2023. Motion-state Alignment for Video Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3571–3580.

Tokmakov, P.; Alahari, K.; and Schmid, C. 2017. Learning motion patterns in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3386–3394.

Van Essen, D. C.; and Maunsell, J. H. 1983. Hierarchical organization and functional streams in the visual cortex. *Trends in neurosciences*, 6: 370–375.

Wei, J.; Wang, S.; and Huang, Q. 2020. F^3Net: fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12321–12328.

Wu, Z.; Paudel, D. P.; Fan, D.-P.; Wang, J.; Wang, S.; Demonceaux, C.; Timofte, R.; and Van Gool, L. 2023. Source-free depth for object pop-out. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1032–1042.

Yang, C.; Lamdouar, H.; Lu, E.; Zisserman, A.; and Xie, W. 2021. Self-supervised video object segmentation by motion grouping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7177–7188.

Yang, S.; Wang, X.; Li, Y.; Fang, Y.; Fang, J.; Liu, W.; Zhao, X.; and Shan, Y. 2022. Temporally efficient vision transformer for video instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2885–2895.

Zhang, H.; Lyu, Y.; He, T.; Li, X.; Li, Y.; Yuan, D.; and Yang, Y. 2025. CODdiff: Prior leading diffusion model for Camouflage Object Detection. *Knowledge-Based Systems*, 113381.

Zhang, H.; Lyu, Y.; Yu, Q.; Liu, H.; Ma, H.; Yuan, D.; and Yang, Y. 2024. Unlocking Attributes' Contribution to Successful Camouflage: A Combined Textual and Visual Analysis Strategy. In *European Conference on Computer Vision*, 315–331. Springer.

Reproducibility Checklist

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) NA

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) yes

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) NA

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) yes

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) no

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) no

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) yes

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) yes

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) NA

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) yes

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) yes

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) yes

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) yes

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) yes