EntitySAM: Segment Everything in Video
— Supplementary Material —

Mingqiao Ye[1,*] Seoung Wug Oh[2] Lei Ke[3] Joon-Young Lee[2]
[1]EPFL [2]Adobe Research [3]Carnegie Mellon University

In this supplementary material, Section 1 presents additional experimental analysis of our EntitySAM, including detailed ablation studies on different visual prompts on class specific segmentation performance and comparison using the same classification method for SOTA methods. Section 2 provides comprehensive implementation details and method limitation analysis, where we thoroughly discuss the technical specifications and potential constraints of our approach. Section 3 presents the pseudo-code of decoder layers, offering a detailed technical breakdown of our architecture's key components and computational workflow. For extensive visual comparisons of our results in Section 4, please refer to the project page.

1. Supplementary Experiments

Vision Language Model Prompt Design We employ GPT-4O [1] for entity mask classification using a carefully designed vision-language input approach. Our text prompts, detailed in Table 1, utilize a pre-defined set of category labels to ensure outputs align with the target categories.

For visual prompts, we conducted an extensive ablation study exploring various input designs to highlight regions of interest for classification. Figure 1 illustrates five distinct visual prompt types, with corresponding performance results shown in Table 2. Initial experiments with uncropped images (prompts a and b) yielded suboptimal results, even with explicit red box references in the text prompts. This revealed VLM's limitations in pixel-level understanding of localized regions. While cropping significantly improved performance, redundant content within the cropped regions occasionally interfered with classification accuracy. This was particularly problematic when classifying background regions that included foreground elements within the cropped box. Our final design, prompt (e), implements Masked Cropping Image, which isolates the region of interest and highlights it with a red box. This approach achieves optimal results and surpasses the previous state-of-the-art in open-vocabulary video panoptic segmentation.

Class Specific Evaluation The optimized prompt design enables us to extend EntitySAM's class-agnostic video mask output to class-specific video panoptic segmentation masks, with results presented in Table 2 of the main paper. Notably, our approach of decoupling segmentation and classification represents a generalizable design principle that extends beyond EntitySAM to other models. As demonstrated in Table 3 , when we applied our mask classification modules to OV-DVIS++, we observed significant improvements. Specifically, for the ResNet50 backbone implementation, we replaced the end-to-end open vocabulary training classification with VLM-based classification as a post-processing step. This modification yielded an approximate 3.0 VPQ improvement, validating the effectiveness of our classification module design. Furthermore, EntitySAM outperformed OV-DVIS++ when using identical classification methods, which can be attributed to our superior class-agnostic mask quality.

2. More Implementation Deatils

More Implementation Deatils EntitySAM incorporates three newly designed modules, as detailed in the main paper. First, the dual visual encoder uses an additional DINOv2 (VIT-S / VIT-L) for feature supplementation. These features pass through a linear projector, are repeated for each mask in the batch, and concatenated with SAM 2's memory features. Second, the PromptGenerator employs 50 learnable queries to perform cross-attention with the enhanced features, predicting both input prompts and their corresponding confidence scores. Third, the entity mask decoder uses 50 mask queries and 50 IoU queries concatenated with prompt queries. This process involves query-level self-attention, followed by query-feature and feature-query cross-attention. The final mask output is generated through dynamic convolution of updated queries and upsampled

*This work was done during an internship at Adobe Research.

Figure 1. Ablation on Entity Mask Classification Visual Prompt Design. (a) Image + Draw Box (b) Image + Draw Box + Draw Mask (c) Croped Image + Draw Box (d) Croped Image + Draw Box + Draw Mask (e) Masked Cropping Image

Table 1. Entity Mask Classification Text Prompt Design.

Prompt Type	Text
System	You are a highly advanced image classification system. You have been trained on a vast array of visual data and can accurately identify objects, scenes, and concepts across a wide range of categories. You will be presented with {batch_size} images for classification. Your task is to analyze each image carefully, considering multiple aspects such as shape, color, texture, context, and any distinguishing features. Note that if different parts are unconnnected in the image, it might be a background/stuff category. Draw upon your extensive knowledge to determine the most accurate label for each image from the provided set of {n_classes} classes. Output your classifications in JSON format, with each image number as the key and the corresponding **index** of the class as the value. Be as precise and specific as possible in your classifications. You **must** only select from the provided classes, and any answer outside this set is **incorrect and unacceptable**. If you're unsure, choose the most likely class based on the visual information available. Here's the expected output format: {{"1": 0, "2": 1, ..., "{batch_size}": "index_value"}} Remember, only output the JSON object with your classifications. Do not include any explanations or additional text. Classes (index: label): {class_mapping}.
User	Please identify the index of the class for the object in the image provided. **Ignore the black empty region**, the interested region has a red line as a bounding box. If the different parts are unconnected, the initial guess is a background/stuff class. Only classify the main object within this region. **The class index must be one of the provided classes, strictly from the system prompt.** Output the answer in a JSON format, with the key {`j+1`}.

Table 2. Ablation Study on Entity Mask Classification Visual Prompt Design. We use a subset of 80 videos from the original VIPSeg validation set with 343 videos for visual prompt ablation.

Model	Visual Prompt	COCO → VIPSeg VPQ	VPQTh	VPQSt
EntitySAM (ours)	(a) Image + Draw Box	12.1	16.7	8.2
	(b) Image + Draw Box + Draw Mask	12.8	17.7	8.8
	(c) Croped Image + Draw Box	19.3	25.1	14.7
	(d) Croped Image + Draw Box + Draw Mask	18.7	23.6	14.8
	(e) Masked Cropping Image	**26.1**	**29.1**	**23.7**

features at a 256 × 256 resolution. We implement tube-mask Hungarian matching and compute IoU L1 loss and Mask Loss solely on matched predictions.

Our training procedure, as outlined in Section 4.1, is entirely class-agnostic. We uniformly label all COCO dataset categories as "Entity" during training. The process consists of two stages: first, a 40K-iteration image-level training phase with frame length of 1, followed by a 10K-iteration video stage. In the video stage, we employ Large Scale Jittering to create pseudo-videos of length 8 with 1024 × 1024 square resolution. We freeze the image encoder, memory encoder, and memory attention parameters, while detaching gradients for the memory encoder during temporal propagation. Training is

Table 3. Comparison of Zero-shot Video Panoptic Segmentation on COCO → VIPSeg evaluation. We also add our designed VLM Entity Mask Classification Module for SOTA models.

Method	Backbone	VLM Classification	VPQ	VPQ^{Th}	VPQ^{St}	STQ
FC-CLIP [5]	ResNet-50		22.3	25.5	19.1	19.7
OV-DVIS++(Online) [6]	ResNet-50		24.4	26.8	22.4	22.0
OV-DVIS++(Offline) [6]	ResNet-50		23.8	26.4	21.4	24.4
FC-CLIP [5]	ResNet-50	✓	25.2	29.4	21.6	23.1
OV-DVIS++(Online) [6]	ResNet-50	✓	27.1	30.6	24.0	25.5
OV-DVIS++(Offline) [6]	ResNet-50	✓	26.8	29.7	24.2	27.2
EntitySAM (ours)	ViT-S	✓	**28.7**	**32.9**	**25.1**	**31.4**

conducted using 8 A100 GPUs. In inference visualization, the first frame is generated without relying on any prior memory. We recommend propagating the first frame twice to produce more stable and consistent visualization results.

Limitation Analysis EntitySAM effectively segments "entities" in videos under our proposed task of Video Entity Segmentation. We have validated its effectiveness across multiple benchmarks. Currently, our zero-shot implementation is trained on COCO to maintain consistency with existing state-of-the-art models [5, 6]. However, the COCO dataset has inherent biases and limited scale, which may restrict generalization capabilities. The model might show decreased performance when handling completely new scenarios, like underwater scenes. Due to computational constraints, we leave larger-scale training for future work, such as developing a SAM-like data engine to demonstrate the scalability of both the task and model.

3. EntiyDecoder

Algorithm Pseudocode We outline the pseudocode for our EntityDecoder in Algorithm 1, where we highlight the core components in both blue and green texts.

4. Video Entity Segmentation Visualization

For extensive visual comparisons of our results, please refer to the project page. We show the zero-shot video entity segmentation comparison of SAM2 [4] (Mask2Former [2] Init Prompt), DEVA [3] and EntitySAM (ours) with ResNet50 / ViT-Small Backbone.

Algorithm 1 EntityDecoder Forward Pipeline

procedure MASKDECODERFORWARD($\mathcal{I}, \mathcal{Q}, \mathcal{P}$)
  Input:
    Image features $\mathcal{I}$
    Object queries $\mathcal{Q}$
    Prompt embeddings $\mathcal{P}$
  Initialize Tokens:
  $\mathcal{T} \leftarrow \text{Concatenate}(\mathcal{Q}_{\text{object}}, \mathcal{Q}_{\text{IoU}}, \mathcal{P}_{\text{sparse}})$ ▷ Shape: $[N_q, 1, D]$
  Process Image Features:
  $\mathcal{F} \leftarrow \mathcal{I} + \mathcal{P}_{\text{dense}}$ ▷ Add dense prompts
  procedure TRANSFORMERFORWARD($\mathcal{F}$, PE, $\mathcal{T}$)
    $\mathcal{F}_{\text{seq}} \leftarrow \text{Flatten}(\mathcal{F})$ ▷ $B \times HW \times C$
    for layer $\in$ TransformerLayers **do**
      // Self-Attention Block
      $\mathcal{Q}' \leftarrow \text{SelfAttn}(\mathcal{T})$ ▷ Inter-object communication
      // Query Processing
      $\mathcal{Q}_{\text{split}} \leftarrow \text{SplitGroups}(\mathcal{Q}', K = 4)$ ▷ K groups
      // Cross-Attention Blocks
      $\mathcal{Q}_{\text{out}} \leftarrow \text{CrossAttn}(\mathcal{Q}_{\text{split}}, \mathcal{F}_{\text{seq}})$
      $\mathcal{Q}_{\text{mlp}} \leftarrow \text{MLP}(\mathcal{Q}_{\text{out}})$
      $\mathcal{F}_{\text{new}} \leftarrow \text{CrossAttn}(\mathcal{F}_{\text{seq}}, \mathcal{Q}_{\text{mlp}})$
      $\mathcal{T} \leftarrow \mathcal{Q}_{\text{mlp}}$
      $\mathcal{F}_{\text{seq}} \leftarrow \mathcal{F}_{\text{new}}$
    end for
    return $\mathcal{T}, \mathcal{F}_{\text{seq}}$
  end procedure
  Generate Predictions:
  // Separate tokens for different tasks
  $\mathcal{T}_{\text{mask}} \leftarrow \mathcal{T}_{1:N_q}$ ▷ Mask tokens
  $\mathcal{T}_{\text{IoU}} \leftarrow \mathcal{T}_{N_q:2N_q}$ ▷ IoU tokens
  // Generate outputs
  $\text{Masks} \leftarrow \text{HyperNetwork}(\mathcal{T}_{\text{mask}}, \mathcal{F}_{\text{upscaled}})$
  $\text{IoU} \leftarrow \text{PredictIoU}(\mathcal{T}_{\text{IoU}})$
  $\text{Class} \leftarrow \text{PredictClass}(\mathcal{P})$
  return Masks, IoU, Class
end procedure

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 3

[3] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *ICCV*, 2023. 3

[4] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[5] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. In *NeurIPS*, 2023. 3

[6] Tao Zhang, Xingye Tian, Yikang Zhou, Shunping Ji, Xuebo Wang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, and Yu Wu. Dvis++: Improved decoupled framework for universal video segmentation. *arXiv preprint arXiv:2312.13305*, 2023. 3