Gemini-SAM3: Semantic Planning and Amodal Tracking for Grounded Video Question Answering

Minseon Park[†] Sungkyunkwan University
Jian Kim[†] Soongsil University
Inhong Paek[†] UC Berkeley
Changwoo Jeon[†] Yonsei University
Jaeyun Shin KAIST
Jong Chul Ye[*] KAIST

Abstract

*Grounded video question answering requires consistent localization of question-referred objects over time, despite occlusion and visual ambiguity. We introduce **Gemini-SAM3**, a modular framework in which Gemini 3.7 Flash generates a shared tracking plan, Gemini provides sparse amodal localization, and SAM3 performs dense temporal propagation. We further adapt SAM3 for amodal tracking and combine the resulting branches using validation-based routing. Our final system achieves an official test HOTA of 0.6637.*

1. Introduction

Grounded video question answering (GVQA) requires identifying question-referred object instances and tracking them consistently over time, as exemplified by the Perception Test benchmark [7]. Unlike video object tracking, targets may be defined by roles, interactions, appearance order, or set membership rather than category labels or initial masks, requiring semantic identification and temporal localization.

Occlusion further complicates GVQA: dense trackers may drift, switch identities, or lose the target, while vision-language models can identify the correct object but produce sparse or spatially inconsistent localization. These limitations motivate amodal reasoning, which estimates the full object extent beyond visible regions [2, 4, 5, 9].

We propose **Gemini-SAM3**, which separates semantic target understanding from dense visual tracking. Gemini 3.7 Flash [3] generates a shared tracking plan and sparse amodal predictions, while SAM3 [1] provides dense trajectories; reliability-aware fusion and validation-derived routing combine zero-shot and amodal-adapted branches.

[†]Equal contribution.
[*]Corresponding author: jong.ye@kaist.ac.kr.

2. Method

As illustrated in Fig. 1, our framework combines question-guided target planning, dense and sparse tracking, reliability-aware fusion with routing, and amodal adaptation of SAM3. All branches share the target identities and anchor prompts produced during planning.

2.1. Question-Guided Target Planning

Given a video and question, Gemini 3.7 Flash identifies referred object instances and constructs a tracking plan containing target IDs, visibility intervals, anchor frames, and bounding boxes. A critic verifies the target description, and multiple anchors are used when necessary to cover targets.

Gemini predicts bounding boxes at the selected anchor frames, and the resulting shared tracking plan is reused by zero-shot SAM3, amodal-adapted SAM3, and sparse Gemini tracking to ensure consistent target definitions.

2.2. Dense and Sparse Tracking

For dense localization, SAM3 [1], following the promptable tracking paradigm of SAM2 [8], uses Gemini-generated anchor boxes to propagate targets bidirectionally through the video. Visibility intervals serve as gates to reduce drift, and SAM3 masks are converted into bounding boxes.

In parallel, Gemini predicts sparse amodal boxes at fixed temporal intervals, estimating full object extent under partial occlusion and position from temporal context under full occlusion. These predictions provide identity-aware references where dense tracking is unreliable.

2.3. Reliability-Aware Fusion and Routing

Each dense SAM3 track is accepted for fusion only if it satisfies:

$$\mathrm{IoU}(b_{\mathrm{anchor}}, b_{\mathrm{SAM}}) \geq 0.25, \quad N_{\mathrm{shared}} \geq 3,$$
$$\frac{1}{N_{\mathrm{shared}}} \sum_{t} \mathrm{IoU}(b_t^A, b_t^D) \geq 0.30.$$

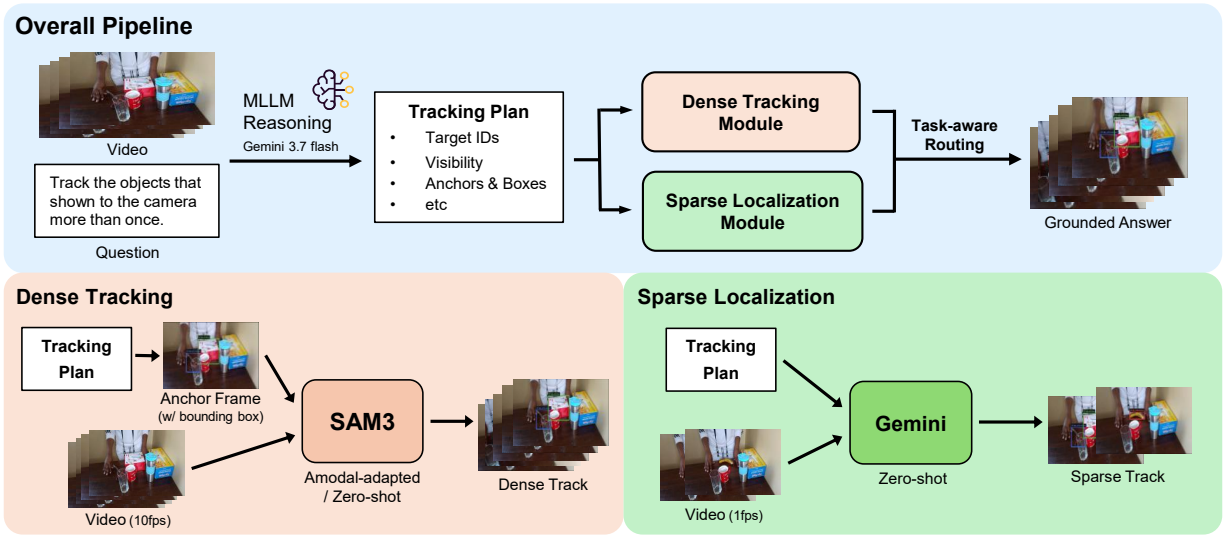


Figure 1. Overview of **Gemini-SAM3**. Gemini 3.7 Flash generates a shared tracking plan and sparse amodal predictions, while SAM3 performs dense tracking. Reliability-aware fusion and validation-derived routing produce the final grounded track.

Here, b_{anchor} and b_{SAM} denote the anchor box and the SAM3 box at the anchor frame, b_t^A and b_t^D denote sparse Gemini and dense SAM3 boxes, and N_{shared} is the number of shared frames. If the gate fails, the Gemini sparse trajectory is retained. Otherwise, SAM3 predictions are overlaid, and on reliable shared frames we compute the box residual

$$r(t) = b_t^A - b_t^D,$$

which is temporally interpolated along the dense SAM3 trajectory to preserve the sparse Gemini alignment while recovering higher-frequency localization. Residual propagation and gap interpolation are applied only over reliable short intervals to avoid extrapolating erroneous tracks across long occlusions.

Fusion produces zero-shot and amodal-adapted candidates, $F^{(z)}$ and $F^{(a)}$, respectively. A validation-derived deterministic router selects between them:

$$R(q,P) = \begin{cases} F^{(a)}, & C(q,P), \\ F^{(z)}, & \text{otherwise.} \end{cases}$$

Here, q and P denote the question and tracking plan, and $C(q,P)$ holds when P contains multiple targets and q is neither an occlusion-game nor an appear-twice question.

2.4. Amodal Adaptation of SAM3

To improve tracking under occlusion, we adapt SAM3 using visible and amodal supervision from SAIL-VOS [5]. The SAM3 backbone, decoders, and memory pathway are frozen, while only a residual head is trained to correct visible-mask logits to recover the full amodal object extent. The objective combines focal, Dice, IoU, presence, and containment losses:

$$\mathcal{L} = 20\mathcal{L}_{\text{focal}} + \mathcal{L}_{\text{dice}} + \mathcal{L}_{\text{iou}} + \mathcal{L}_{\text{presence}} + 5\mathcal{L}_{\text{contain}}.$$

Focal and Dice losses jointly supervise the amodal mask prediction, while the IoU and presence terms provide auxiliary supervision. The containment loss encourages the visible region to remain within the predicted amodal extent.

3. Experiment

The amodal branch is trained on SAIL-VOS [5] using 10 training and four disjoint validation sequences. Samples are eight-frame shot-aware causal clips, with fully occluded objects retained as positive amodal targets and ignored regions excluded from the loss. Training runs for 20 epochs at 1008-pixel resolution with batch size 1 using AdamW and cosine learning-rate decay from 10^{-4} to 10^{-5}. Only the residual head is optimized; frozen SAM3 components run in BF16, while the residual head is computed in FP32.

We evaluate with HOTA [6], which balances detection and association accuracy. Gemini 3.7 Flash [3] is used zero-shot for planning and sparse amodal localization. At inference, SAM3 performs dense tracking at 10 fps, while Gemini provides sparse localization at 1 fps. SAM3 uses the official checkpoint, while the amodal branch uses the SAIL-VOS-trained residual head. Routing is fixed using the challenge validation set and applied unchanged to the test set, without test annotations or per-example scores.

Table 1 shows gains from both amodal adaptation and sparse Gemini fusion, with selective routing achieving the best test HOTA of 0.6637.

Table 1. Test-set HOTA results.

Method	HOTA
Zero-shot SAM3 pipeline	0.6263
Amodal-adapted SAM3	0.6448
SAM3 + Gemini sparse fusion	0.6620
Selective routing	**0.6637**

4. Conclusion

We presented Gemini-SAM3, a modular framework for grounded video question answering that separates semantic target identification, dense visual tracking, and amodal reasoning. The final system achieves a test HOTA of 0.6637.

Acknowledgements

This work was supported by Snap through the project "Development of Generative AI-Based Object Removal and Scene Restoration Techniques for Advanced Image Editing."

References

[1] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment Anything with Concepts. In *International Conference on Learning Representations (ICLR)*, 2026.

[2] Kaihua Chen, Deva Ramanan, and Tarasha Khurana. Using Diffusion Priors for Video Amodal Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22890–22900, 2025.

[3] Google DeepMind. Gemini 3.7 Flash. Model Card, 2026. Published August 13, 2026.

[4] Cheng-Yen Hsieh, Kaihua Chen, Achal Dave, Tarasha Khurana, and Deva Ramanan. TAO-Amodal: A Benchmark for Tracking Any Object Amodally. *arXiv preprint arXiv:2312.12433*, 2023.

[5] Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang, and Alexander G. Schwing. SAIL-VOS: Semantic Amodal Instance Level Video Object Segmentation – A Synthetic Dataset and Baselines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3105–3115, 2019.

[6] Jonathon Luiten, Aljoša Ošep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. HOTA: A Higher Order Metric for Evaluating Multi-Object Tracking. *International Journal of Computer Vision*, 129:548–578, 2021.

[7] Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception Test: A Diagnostic Benchmark for Multimodal Video Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[8] 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. In *International Conference on Learning Representations (ICLR)*, 2025.

[9] Bo Zhang, Zhuotao Tian, Xin Tao, Songlin Tang, Jun Yu, and Wenjie Pei. Amodal SAM: A Unified Amodal Segmentation Framework with Generalization. *arXiv preprint arXiv:2604.20748*, 2026.