Adapting Video Foundation Models to 3D Medical Volumes via Latent Trajectory Priors

Guowei Dai[1], Duwei Dai[2], Yulong Ji[3], Yi Zhang[4], and Hu Chen[1⋆]

[1] College of Computer Science, Sichuan University, China
[2] Second Affiliated Hospital of Xi'an Jiaotong University, China
[3] School of Aeronautics and Astronautics, Sichuan University, China
[4] School of Cyber Science and Engineering, Sichuan University, China

Abstract. Adapting video foundation models to volumetric medical segmentation by treating axial depth as a temporal dimension is attractive, but the resulting slice propagation can be unstable when anatomical structures branch, disappear, or become poorly contrasted across slices. Static geometric prompts provide little guidance for newly appearing or disconnected components, and SAM2-style streaming memory may accumulate errors in these regions. We propose DiffuPrompt, a backbone-frozen, single-prompt framework that augments SAM2 with latent trajectory priors for volumetric propagation. Given a user prompt on a representative slice, an LLM-conditioned Neural ODE predicts a continuous, patient-conditioned latent trajectory and converts it into a persistent prior memory. A dual-stream spatial gate then fuses this prior memory with the standard observation-driven working memory, allowing the model to use image evidence when it is reliable and trajectory priors when local propagation becomes ambiguous. Experiments on MSD, BTCV, and BraTS show improved promptable volumetric segmentation over SAM/SAM2-based baselines, especially on boundary-sensitive metrics.

Keywords: Video Foundation Models · 3D Medical Image Segmentation · Neural Ordinary Differential Equations · Latent Trajectory Priors · Interactive Segmentation

1 Introduction

Visual Foundation Models (VFMs) such as the Segment Anything Model 2 (SAM2) [18, 21] have made promptable video object segmentation practical through streaming memory architectures [3, 4]. A natural way to use these models for 3D medical volumes is to treat the axial direction as a temporal sequence [1, 26]. This setting is useful when a clinician can provide one prompt on a representative slice but dense labels for a new organ, modality, or site are not available. It is also different from fully supervised, closed-set automatic segmentation: our goal is single-prompt volumetric propagation with a frozen SAM2

⋆ Corresponding author.

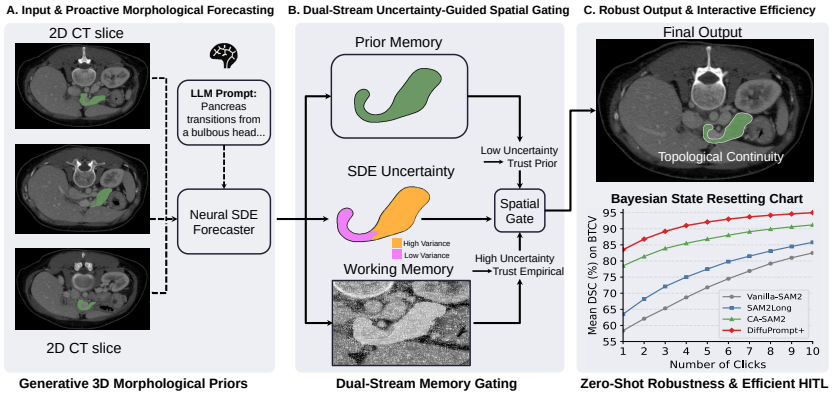

Fig. 1: Fig. 1. Overview of DiffuPrompt. (A) A single prompt on a representative 2D CT slice is used to forecast cross-slice anatomical evolution with an LLM-guided Neural SDE/ODE prior. (B) The predicted prior memory and observation-driven working memory are adaptively fused by an uncertainty-guided spatial gate. (C) This dual-stream memory gating improves topological continuity and robust volumetric propagation with fewer interactive corrections.

backbone, not replacing task-specific models such as nnU-Net when same-domain annotations are abundant.

Directly applying SAM2 to axial slices can be unstable because anatomical structures may branch, merge, disappear, or become poorly contrasted across depth. Static prompts such as points or boxes are spatially local, and a prompt on one slice provides limited guidance for disconnected components that appear later. With a greedy First-In-First-Out (FIFO) memory queue [20, 25], early errors in ambiguous slices may be reused as memory and then propagate to subsequent slices. We refer to this failure mode as memory drift or memory collapse in the remainder of the paper.

To address this issue, we propose DiffuPrompt, a trajectory-prior framework for promptable volumetric segmentation (conceptually illustrated in Fig. 1). Instead of relying only on observed historical masks, DiffuPrompt builds a persistent prior memory from a continuous latent trajectory. Large Language Models (LLMs) provide organ-level anatomical descriptions [6, 19], and a Neural Ordinary Differential Equation (ODE) parameterizes a smooth latent flow along the axial direction. This trajectory prior is not used as a hard anatomical template; it is fused with the observation-driven working memory through a learned spatial gate.

This design keeps the contribution at the system level: a frozen SAM2 backbone is combined with trainable lightweight adapters, a semantic-conditioned latent ODE, and a dual-stream memory gate. During test-time inference, the user provides one center-slice prompt and the model propagates the segmentation bidirectionally through the volume. In summary, our main contributions are as follows:

- We analyze a practical failure mode of SAM2-style axial propagation for medical volumes: static prompts and FIFO memory can accumulate errors when structures change shape or visibility across slices.
- We propose DiffuPrompt, which constructs LLM-conditioned Neural ODE trajectory priors and stores them as a persistent prior memory for single-prompt volumetric propagation.
- We introduce a dual-stream spatial gate that adaptively combines observation-driven working memory with trajectory prior memory, and validate the resulting system on MSD, BTCV, and BraTS under a promptable backbone-frozen protocol.

2 Related Work

Medical Image Segmentation via SAM and SAM2. Early adaptations of the Segment Anything Model (SAM) [2,7,12] focused on bridging the natural-to-medical domain gap via parameter-efficient fine-tuning (e.g., MoE-SAM [13]) or 3D spatial context aggregation (e.g., GA-SAM [14], 3D-SAutoMed [15]). SAM2 [9,16,18] extends this paradigm to videos, and surgical tracking adaptations such as VP-SAM [19] and SAM-I2V [17] add spatio-temporal side networks. These methods are strong promptable baselines, but most axial-volume adaptations still depend heavily on local appearance consistency. DiffuPrompt complements this line by adding an explicit trajectory prior memory for structures that are hard to propagate from observation memory alone.

Latent Priors for Volumetric Consistency. Generative medical models have been used for modality translation, augmentation, and enforcing volumetric consistency. For example, MedM2G [22] studies medical multi-modal generation, VAP-Diffusion [10] uses multimodal language descriptions for medical generation, and ISCS [5] improves inter-slice consistency in 3D medical image synthesis. Our work uses this broader idea of latent anatomical regularity differently: rather than generating a full image volume for augmentation, we predict a compact latent trajectory and convert it into SAM2 memory features for promptable segmentation.

Memory Mechanisms in Video Object Segmentation. The memory bank is central to VOS architectures [20]. Traditional approaches employ greedy, continuously expanding memory, which can amplify error accumulation [23,25]. Recent methods restrict or filter memory updates. For instance, RMem [25] limits memory capacity using temporal positional embeddings, SAM2Long [4] maintains multiple hypothesis paths through a constrained memory tree, and MA-SAM2 [20] uses mask-quality rejection to prevent degraded features from entering the memory pool. These mechanisms mainly operate on observed frames. In contrast, DiffuPrompt keeps a separate prior memory $\mathcal{B}^{\mathrm{pm}}$ and uses a spatial confidence gate to combine it with the working memory $\mathcal{B}^{\mathrm{wm}}$.

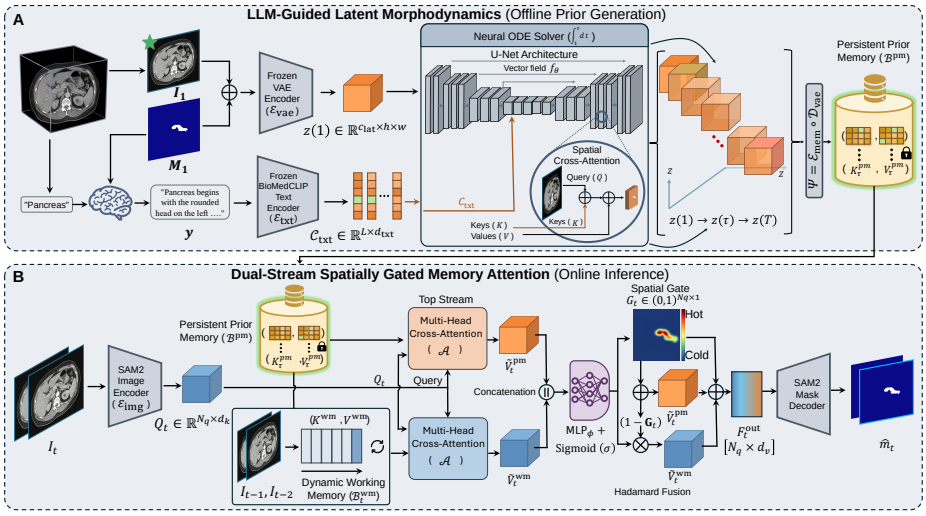

Fig. 2: Overall architecture of DiffuPrompt. **(A) Prior generation:** an LLM-conditioned Neural ODE predicts a continuous latent trajectory and converts it into a persistent prior memory ($\mathcal{B}^{\mathrm{pm}}$). **(B) Inference:** a dual-stream spatial gate fuses observation-driven working memory ($\mathcal{B}^{\mathrm{wm}}$) with trajectory prior memory for volumetric propagation.

3 Methodology

We formulate promptable 3D medical segmentation as a temporally conditioned propagation problem coupled with a persistent memory-augmented visual foundation model. Given an anisotropic 3D medical volume $\mathcal{V} \in \mathbb{R}^{T\times C_{\mathrm{in}}\times H\times W}$, where T denotes the number of transverse slices, C_{in} represents the imaging modality channels, and $H \times W$ is the spatial resolution, an interactive prior set $\mathcal{P}_{t_0} = \{p_i, c_i\}_{i=1}^{N_p}$ is provided on a representative slice t_0. Our objective is to infer $\Phi_\Theta : (\mathcal{V}, \mathcal{P}_{t_0}) \to \hat{\mathcal{M}}$, where $\hat{\mathcal{M}} \in \{0,1\}^{T\times 1\times H\times W}$ is the dense volumetric binary mask sequence. DiffuPrompt augments the SAM2 working memory with an LLM-conditioned Neural ODE prior memory and a dual-stream spatial gate, as shown in Fig. 2.

3.1 Latent Morphodynamics via Conditional Neural ODEs

The prior module estimates a patient-conditioned anatomical trajectory along the Z-axis without directly using future manual prompts. Instead of frame-by-frame static priors, we represent the organ evolution as a continuous flow in a compressed latent space.

Let $\mathcal{E}_{\mathrm{vae}} : \mathbb{R}^{C_{\mathrm{in}}\times H\times W} \times \{0,1\}^{1\times H\times W} \to \mathbb{R}^{c_{\mathrm{lat}}\times h\times w}$ be a frozen pre-trained variational autoencoder encoder. The trajectory is initialized from the prompted slice by concatenating image I_{t_0} and mask M_{t_0}: $z(t_0) = \mathcal{E}_{\mathrm{vae}}(I_{t_0} \,\|\, M_{t_0})$, where $\|$ denotes channel-wise concatenation and $c_{\mathrm{lat}} \ll C_{\mathrm{in}}$ keeps the trajectory compact.

We model the axial drift of the latent anatomical representation with a Neural ODE. The continuous-time vector field $f_\theta : \mathbb{R}^{c_{\text{lat}} \times h \times w} \times \mathbb{R} \times \mathbb{R}^{d_{\text{txt}}} \to \mathbb{R}^{c_{\text{lat}} \times h \times w}$ is parameterized by a semantic-conditioned UNet. The text condition is encoded from an anatomical description y (e.g., "pancreas with varying cross-sections") as $\mathcal{C}_{\text{txt}} = \mathcal{E}_{\text{txt}}(y) \in \mathbb{R}^{d_{\text{txt}}}$. The latent state evolves according to $\frac{dz(t)}{dt} = f_\theta\big(z(t), t, \mathcal{C}_{\text{txt}}\big)$.

To obtain a discrete trajectory for the volume, we solve the initial value problem from t_0 to each target slice index τ and project the resulting latent states into the SAM2 memory space. By combining the VAE decoder $\mathcal{D}_{\text{vae}}$ and the SAM2 memory encoder $\mathcal{E}_{\text{mem}}$ into $\Psi = \mathcal{E}_{\text{mem}} \circ \mathcal{D}_{\text{vae}}$, the prior memory bank $\mathcal{B}^{\text{pm}}$ is

$$\mathcal{B}^{\text{pm}} = \left\{ \Psi \left(z(t_0) + \int_{t_0}^{\tau} f_\theta\big(z(s), s, \mathcal{C}_{\text{txt}}\big) ds \right) \right\}_{\tau=1}^{T}, \tag{1}$$

where each element in $\mathcal{B}^{\text{pm}}$ comprises prior key-value pairs $(K_\tau^{\text{pm}}, V_\tau^{\text{pm}})$. Here, $K_\tau^{\text{pm}} \in \mathbb{R}^{N_m \times d_k}$ and $V_\tau^{\text{pm}} \in \mathbb{R}^{N_m \times d_v}$, with N_m denoting the number of spatial memory tokens per slice.

LLM-Driven Anatomical Semantic Injection. Instead of using only one-hot class labels, we use an LLM to construct anatomical descriptors y. For each target organ class, the prompt requests concise shape and cross-slice transition descriptions, such as *The pancreas transitions from a bulbous head to a tapering tail.* These text descriptors are encoded by BioMedCLIP into $\mathcal{C}_{\text{txt}} \in \mathbb{R}^{L \times d_{\text{txt}}}$. Within the Neural ODE vector field, $\mathcal{C}_{\text{txt}}$ is injected into spatial latent features $z(t)$ at multiple resolution scales through cross-attention layers, so the trajectory prior is conditioned on both the prompted slice and organ semantics.

3.2 Dual-Stream Spatially Gated Memory Attention

After constructing the trajectory prior, we modify the auto-regressive decoding process of SAM2 to keep the prior memory separate from observation-driven updates. The model uses a dynamic working memory $\mathcal{B}_t^{\text{wm}} = \{(K_i^{\text{wm}}, V_i^{\text{wm}})\}_{i=t-W}^{t-1}$ of window size W and the persistent prior memory $\mathcal{B}^{\text{pm}}$.

For the current slice I_t, the image encoder extracts the visual query feature $Q_t \in \mathbb{R}^{N_q \times d_k}$ (where $N_q = H' \times W'$). We define a generalized multi-head cross-attention functional $\mathcal{A}(\cdot, \cdot)$ parameterized by the memory stream $s \in \{\text{wm}, \text{pm}\}$. Incorporating the scaling factor $\sqrt{d_k}$ and omitting the multi-head splitting for brevity, the feature retrieval from stream s is:

$$\tilde{V}_t^s = \mathcal{A}(Q_t, \mathcal{B}^s) = \text{Softmax}\left(\frac{Q_t (K^s)^\top}{\sqrt{d_k}} \right) V^s, \quad \tilde{V}_t^s \in \mathbb{R}^{N_q \times d_v}, \tag{2}$$

where $K^s \in \mathbb{R}^{L_s \times d_k}$ and $V^s \in \mathbb{R}^{L_s \times d_v}$ denote the concatenated keys and values across the respective memory sequence length L_s ($L_{\text{wm}} = W \cdot N_m$ and $L_{\text{pm}} = T \cdot N_m$).

A fixed linear combination of $\tilde{V}_t^{\text{wm}}$ and $\tilde{V}_t^{\text{pm}}$ can dilute useful image evidence when the observation is clear. We therefore use a spatial confidence gate to adaptively weight the two memory streams. We concatenate the query and retrieved

features along the feature dimension and map them to a dense probability map using an MLP (MLP_ϕ):

$$\mathbf{G}_t = \sigma\Big(\mathrm{MLP}_\phi\big(Q_t \parallel \tilde{V}_t^{\mathrm{wm}} \parallel \tilde{V}_t^{\mathrm{pm}}\big)\Big) \in (0,1)^{N_q \times 1}, \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid activation. Larger values of $\mathbf{G}_t$ increase the contribution of the trajectory prior, while smaller values keep the prediction closer to the observation-driven working memory. The final memory context $F_t^{\mathrm{out}} \in \mathbb{R}^{N_q \times d_v}$ injected into the mask decoder is the convex combination

$$F_t^{\mathrm{out}} = \mathbf{G}_t \odot \tilde{V}_t^{\mathrm{pm}} + (\mathbf{1} - \mathbf{G}_t) \odot \tilde{V}_t^{\mathrm{wm}}. \tag{4}$$

This soft fusion lets DiffuPrompt preserve local image evidence while using the trajectory prior as an additional cue in ambiguous regions.

3.3 Training Objective and Inference Protocol

DiffuPrompt is not training-free. The SAM2 backbone, VAE, LLM, and BioMedCLIP text encoder are frozen, while the LoRA adapters, Neural ODE vector field, and gate MLP are trained on the training partitions of the evaluated benchmarks. At test time, no annotation from the target volume is used beyond the single user prompt. We therefore describe the setting as backbone-frozen, single-prompt inference rather than training-free inference.

The trainable modules are optimized with supervised volumetric masks from the training split. The objective combines a segmentation term for the decoded masks, a latent consistency term between predicted and encoded slice states, and a smoothness regularizer on adjacent latent states:

$$\mathcal{L} = \mathcal{L}_{\mathrm{seg}}(\hat{\mathcal{M}}, \mathcal{M}) + \lambda_{\mathrm{lat}} \mathcal{L}_{\mathrm{lat}}(\hat{z}, z) + \lambda_{\mathrm{sm}} \sum_{t=2}^{T} \|\hat{z}(t) - \hat{z}(t-1)\|_2^2. \tag{5}$$

Here, $\mathcal{L}_{\mathrm{seg}}$ is implemented with the standard Dice and binary cross-entropy losses. The latent and smoothness terms regularize the ODE trajectory so that the prior memory changes smoothly across nearby slices while still being supervised by the observed anatomical masks.

4 Experiments

4.1 Experimental Setup

Datasets. We benchmark DiffuPrompt on three 3D medical datasets that cover different organs, modalities, and shape variations. All volumes are standardized to $1.0 \times 1.0 \times 1.0$ mm^3 isotropic spacing and intensity-clipped (0.5–99.5 percentile) to reduce preprocessing differences. Medical Segmentation Decathlon (MSD): Spanning 10 heterogeneous tasks across CT/MRI modalities and target geometries.

BTCV (Beyond the Cranial Vault): Comprising 47 portal venous CT scans with 13 annotated abdominal organs.
BraTS 2023 (Brain Tumors): Multi-parametric MRI scans with variable tumor appearance and infiltrating boundaries.

Evaluation Metrics. We evaluate segmentation performance using the Dice Similarity Coefficient (DSC) for overlap accuracy. To account for clinically significant boundary errors and topological fractures, we employ the 95th percentile Hausdorff Distance (HD_{95}) and the Normalized Surface Distance (NSD) at a 1 mm tolerance threshold.

Implementation Details. The framework is implemented in PyTorch and trained on NVIDIA A100 GPUs. We adopt the pre-trained SAM2-Hiera-Large [18] as the vision backbone and keep it frozen. Low-Rank Adaptation (LoRA, $r = 8$) is injected into the image encoder's query/value projections, while the Neural ODE vector field and gate MLP are trained with the objective in Eq. (5). `Llama-3-8B-Instruct` produces organ descriptions that are embedded by BioMedCLIP [24]. We integrate the Neural ODE using `torchdiffeq`. During inference, following standard promptable VOS protocols, a single interactive prompt (box or click) is provided on the target's center slice, from which the model decodes the bidirectional volumetric sequence.

Table 1: Quantitative comparison under a single center-slice bounding-box prompt. We report mean DSC with standard deviation in the subscript, plus HD95 and NSD. **Bold** indicates the best promptable result within this table, while underline denotes the second-best. *Fully-supervised models use closed-set training and are included as reference upper bounds rather than direct promptable competitors.

Method	MSD (Macro-Avg 10 Tasks)			BraTS 2023			BTCV (13 Organs)		
	DSC (%) ↑	HD95 (mm) ↓	NSD (%) ↑	DSC (%) ↑	HD95 (mm) ↓	NSD (%) ↑	DSC (%) ↑	HD95 (mm) ↓	NSD (%) ↑
Fully-supervised Specialist Baselines (Static, Closed-set Upper Bounds)									
Swin UNETR* [8]	$82.3_{\pm 12.8}$	4.49	81.5	$83.5_{\pm 4.8}$	4.75	81.2	$84.1_{\pm 6.6}$	3.89	82.5
nnU-Net* [11]	$84.5_{\pm 10.2}$	3.86	83.2	$85.4_{\pm 3.1}$	4.25	83.6	$86.2_{\pm 5.8}$	3.15	85.0
Natural Video Tracking Foundation Models (Single-Prompt Inference)									
Vanilla SAM2 [18]	$60.8_{\pm 24.5}$	18.40	58.4	$62.2_{\pm 18.6}$	16.30	59.4	$58.4_{\pm 22.3}$	19.50	54.6
SAM2Long [4]	$66.4_{\pm 19.6}$	14.25	63.5	$67.1_{\pm 14.8}$	13.85	63.8	$63.5_{\pm 18.2}$	16.20	59.4
Efficient-SAM2 [23]	$65.2_{\pm 20.2}$	15.10	62.2	$66.4_{\pm 15.5}$	14.60	62.2	$62.8_{\pm 19.8}$	17.10	58.2
Medical-Specific Promptable/Foundation Models									
VRP-SAM [19]	$71.5_{\pm 16.8}$	11.45	68.2	$73.8_{\pm 9.5}$	10.50	70.1	$70.2_{\pm 14.2}$	12.40	67.5
SAM-Veteran [6]	$73.4_{\pm 15.5}$	9.25	70.8	$75.4_{\pm 8.2}$	9.25	72.6	$72.8_{\pm 12.5}$	10.10	69.2
SAM-CP [1]	$75.2_{\pm 14.8}$	8.60	72.1	$77.5_{\pm 7.4}$	8.40	75.2	$75.4_{\pm 12.1}$	9.80	72.6
ReSurgSAM2 [16]	$76.6_{\pm 14.1}$	7.95	73.5	$78.2_{\pm 6.6}$	7.85	76.5	$77.0_{\pm 11.3}$	8.60	74.8
CA-SAM2 [9]	$78.5_{\pm 13.6}$	6.60	75.8	$80.1_{\pm 5.0}$	6.50	78.4	$\underline{78.5}_{\pm 11.4}$	7.90	<u>76.4</u>
AoP-SAM [2]	$\underline{79.5}_{\pm 12.2}$	<u>5.75</u>	<u>76.5</u>	$\underline{80.5}_{\pm 4.8}$	<u>6.21</u>	<u>78.7</u>	$78.2_{\pm 10.8}$	<u>7.45</u>	76.2
DiffuPrompt (Ours)	$\mathbf{81.2}_{\pm 11.2}$	**4.92**	**79.6**	$\mathbf{82.4}_{\pm 3.6}$	**5.64**	**80.4**	$\mathbf{82.7}_{\pm 8.5}$	**4.46**	**80.5**

4.2 Comparison with State-of-the-Art

Table 1 compares DiffuPrompt with promptable SAM/SAM2-based baselines and fully supervised reference models. The natural video models, Vanilla SAM2

and SAM2Long, obtain lower DSC and larger variance on the medical volumes, consistent with the memory-drift failure mode discussed earlier. Compared with medical promptable baselines such as CA-SAM2 and AoP-SAM, DiffuPrompt improves the mean DSC on MSD, BraTS, and BTCV. The gains are also visible in boundary metrics: HD95 decreases from 5.75 mm to 4.92 mm on MSD and from 7.45 mm to 4.46 mm on BTCV compared with the strongest promptable baseline in the table.

The standard deviations reported with DSC suggest that the trajectory prior and dual-stream gate improve propagation stability. On BraTS 2023, for example, DiffuPrompt reaches 82.4% DSC with a 3.6% standard deviation, compared with $80.1\%_{\pm 5.0}$ for CA-SAM2. The fully supervised methods remain important closed-set references; our result on BTCV is below nnU-Net's 86.2% DSC, but it is obtained in a single-prompt propagation setting with a frozen SAM2 backbone.

4.3 Qualitative Analysis

To inspect propagation behavior, Fig. 3 presents a qualitative comparison on the multi-organ dataset.

As shown in the 3D renderings (top two rows), standard video foundation models such as Vanilla SAM2 and SAM2Long can produce fragmented structures when the target shape changes along the Z-axis, particularly for tubular structures such as vessels. DiffuPrompt produces more continuous 3D structures in these examples.

The bottom rows show zoomed-in 2D axial patches where low tissue contrast causes mask drift and boundary leakage in several baselines. The results of DiffuPrompt indicate that the gated prior memory can reduce some of these propagation errors while still preserving local boundary evidence.

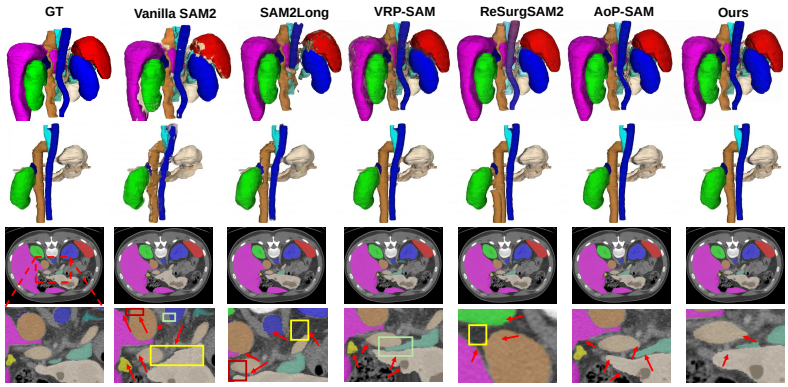

Fig. 3: Qualitative comparison of 3D multi-organ segmentation. Top: Baseline VFMs exhibit topological fragmentation and disconnected structures, while our method ensures morphological continuity. Bottom: 2D patches (colored boxes/red arrows) show memory collapse and boundary leakage in existing methods.

Table 2: Ablation study of core components in DiffuPrompt on the dense multi-target BTCV (13 Organs) dataset. We isolate the contributions of LLM-enriched semantics, Neural ODE-driven continuous trajectories, and the dual-stream spatial gate. $\mathcal{M}_{\text{FIFO}}$ denotes standard greedy memory, while $\mathcal{M}_{\text{Dual}}$ denotes our dual-stream memory.

Variant	Architectural Components				BTCV (13 Organs)		
	LLM	ODE	Memory	Gate ($\mathbf{G}_t$)	DSC (%) ↑	HD95 (mm) ↓	NSD (%) ↑
v_0	-	-	$\mathcal{M}_{\text{FIFO}}$	-	$58.4_{\pm 22.3}$	19.50	54.6
v_1	-	-	$\mathcal{M}_{\text{FIFO}}$	-	$63.2_{\pm 18.5}$	16.40	60.1
v_2	✓	-	$\mathcal{M}_{\text{FIFO}}$	-	$68.5_{\pm 15.2}$	12.80	65.4
v_3	✓	✓	$\mathcal{M}_{\text{FIFO}}$	-	$74.6_{\pm 11.5}$	8.90	71.8
v_4	✓	✓	$\mathcal{M}_{\text{Dual}}$	-	$78.9_{\pm 9.8}$	6.20	76.5
Ours	✓	✓	$\mathcal{M}_{\text{Dual}}$	✓	$\mathbf{82.7_{\pm 8.5}}$	**4.46**	**80.5**

Note: v_0 represents the Vanilla SAM2 baseline. v_1 uses static slice-wise latent priors without LLM semantics or continuous trajectory modeling. v_4 uses simple addition for memory fusion instead of the adaptive spatial confidence gate ($\mathbf{G}_t$).

4.4 Ablation Study

By isolating each component, we examine how DiffuPrompt reduces memory drift in the promptable BTCV setting. The quantitative results are summarized in Tab. 2. The vanilla SAM2 baseline (v_0) relies on standard greedy memory ($\mathcal{M}_{\text{FIFO}}$) and local appearance consistency, yielding 58.4% DSC and 19.50 mm HD95. Adding static slice-wise latent priors (v_1) improves DSC by 4.8 points, suggesting that prior information is useful but insufficient without semantic conditioning or continuous trajectory modeling.

Integrating LLM anatomical semantics into the prior (v_2) improves DSC to 68.5%, indicating that organ shape descriptions help condition the latent trajectory. Replacing the static prior with the continuous Neural ODE forecaster (v_3) further reduces HD95 from 12.80 mm to 8.90 mm and increases DSC to 74.6%, supporting the value of modeling slice-to-slice continuity.

Replacing standard memory with the dual-stream memory and simple addition (v_4) improves DSC to 78.9%, but fixed fusion can still dilute clear local image evidence. The learned spatial gate $\mathbf{G}_t$ improves DSC to 82.7%, reduces HD95 to 4.46 mm, and raises NSD to 80.5%, showing that adaptive memory weighting is important in this setting.

5 Conclusion

In this work, we presented DiffuPrompt, a trajectory-prior framework for adapting SAM2 to promptable 3D medical volume segmentation. The method combines an LLM-conditioned Neural ODE prior memory with a dual-stream spatial gate so that single-prompt axial propagation can use both local image evidence and a continuous latent trajectory prior. Experiments on MSD, BTCV, and

BraTS show consistent improvements over SAM/SAM2-based promptable baselines, especially on boundary-sensitive metrics. The method is intended for low-annotation or new-task interactive settings rather than as a replacement for fully supervised closed-set segmentation. A remaining limitation is that generated trajectory priors can be risky when the prior conflicts with genuine pathology or unusual anatomy; future work should add explicit uncertainty calibration and disagreement detection between image evidence and prior memory.

Acknowledgements

This work was supported in part by the Sichuan Natural Science Foundation under Grant 2024NSFSC0725.

References

1. Chen, P., Xie, L., Huo, X., Yu, X., Zhang, X., Sun, Y., Han, Z., Tian, Q.: Sam-cp: Marrying sam with composable prompts for versatile segmentation (2025)
2. Chen, Y., Son, M., Hua, C., Kim, J.Y.: Aop-sam: Automation of prompts for efficient segmentation. In: AAAI. vol. 39, pp. 2284–2292 (2025)
3. Cuttano, C., Trivigno, G., Rosi, G., Masone, C., Averta, G.: Samwise: Infusing wisdom in sam2 for text-driven video segmentation. In: CVPR. pp. 3395–3405 (2025)
4. Ding, S., Qian, R., Dong, X., Zhang, P., Zang, Y., Cao, Y., Guo, Y., Lin, D., Wang, J.: Sam2long: Enhancing sam 2 for long video segmentation with a training-free memory tree. In: CVPR. pp. 13614–13624 (2025)
5. Du, C., Wu, Q., Tian, X., Yu, J., Wei, H., Zhang, Y.: Improving 2d diffusion models for 3d medical imaging with inter-slice consistent stochasticity. ICLR (2026)
6. Du, T., Li, H., Fan, Z., Zhang, J., Pan, P., Zhang, Y.: Sam-veteran: An mllm-based human-like sam agent for reasoning segmentation. In: ICLR (2026)
7. Gowda, S.N., Clifton, D.A.: Cc-sam: Sam with cross-feature attention and context for ultrasound image segmentation. In: ECCV. pp. 108–124. Springer (2024)
8. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: MICCAI. pp. 272–284. Springer (2021)
9. Huang, H., He, H., Xu, L., Zhu, X., Feng, S., Fu, G.: Ca-sam2: Sam2-based context-aware network with auto-prompting for nuclei instance segmentation. In: MICCAI. pp. 86–95. Springer (2025)
10. Huang, P., Fu, J., Guo, B., Li, Z., Wang, Y., Guo, Y.: Vap-diffusion: Enriching descriptions with mllms for enhanced medical image generation. In: MICCAI. pp. 624–633. Springer (2025)
11. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: ICCV. pp. 4015–4026 (October 2023)

13. Li, R., Wu, L., Gu, J., Xu, Q., Chen, W., Cai, X., Bu, J.: MoE-SAM: Enhancing SAM for Medical Image Segmentation with Mixture-of-Experts. In: MICCAI 2025, vol. 15961, pp. 367–377. Cham (2026)
14. Li, S., Zhang, J., Qi, L., Shi, Y.: GA-SAM: Geometry-Aware SAM Adaptation with Sparse Annotation-Driven Point Cloud Completion. In: MICCAI 2025, vol. 15967, pp. 214–224. Cham (2026)
15. Liang, J., Cao, P., Yang, W., Yang, J., Zaiane, O.R.: 3D-SAutoMed: Automatic Segment Anything Model for 3D Medical Image Segmentation from Local-Global Perspective. In: MICCAI 2024, vol. 15009, pp. 3–12. Cham (2024)
16. Liu, H., Gao, M., Luo, X., Wang, Z., Qin, G., Wu, J., Jin, Y.: Resurgsam2: Referring segment anything in surgical video via credible long-term tracking. In: MICCAI. pp. 435–445. Springer (2025)
17. Mei, H., Zhang, P., Shou, M.Z.: Sam-i2v: Upgrading sam to support promptable video segmentation with less than 0.2% training cost. In: CVPR. pp. 3417–3426 (June 2025)
18. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. ICLR (2024)
19. Sun, Y., Chen, J., Zhang, S., Zhang, X., Chen, Q., Zhang, G., Ding, E., Wang, J., Li, Z.: Vrp-sam: Sam with visual reference prompt. In: CVPR. pp. 23565–23574 (June 2024)
20. Yin, M., Wang, F., Ye, X., Meng, Y., Fu, Z.: Memory-augmented sam2 for training-free surgical video segmentation. In: MICCAI. pp. 328–337. Springer (2025)
21. Yin, Z., Nie, D., Li, S., Pan, J., Tang, Z.: Iterative foundation-dedicated learning: Optimized key frames, prompts and memories for semi-supervised segmentation. In: MICCAI. pp. 258–267. Springer (2025)
22. Zhan, C., Lin, Y., Wang, G., Wang, H., Wu, J.: Medm2g: Unifying medical multi-modal generation via cross-guided diffusion with visual invariant. In: CVPR. pp. 11502–11512 (2024)
23. Zhang, J., Li, Z., Liu, X., Gu, Q.: Efficient-sam2: Accelerating sam2 with object-aware visual encoding and memory retrieval (2026)
24. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
25. Zhou, J., Pang, Z., Wang, Y.X.: Rmem: Restricted memory banks improve video object segmentation. In: CVPR. pp. 18602–18611 (2024)
26. Zhou, Z., Hu, Y., Song, Y., Li, Z., Hu, S., Zhang, L.Y., Yao, D., Zheng, L., Jin, H.: Vanish into thin air: Cross-prompt universal adversarial attacks for sam2 (2025)