# Time-Contrastive Pretraining for In-Context Image and Video Segmentation

Assefa Wahd, Jacob Jaremko, and Abhilash Hareendranathan

Department of Radiology and Diagnostic Imaging
University of Alberta
{wahd, jjaremko, hareendr}@ualberta.ca

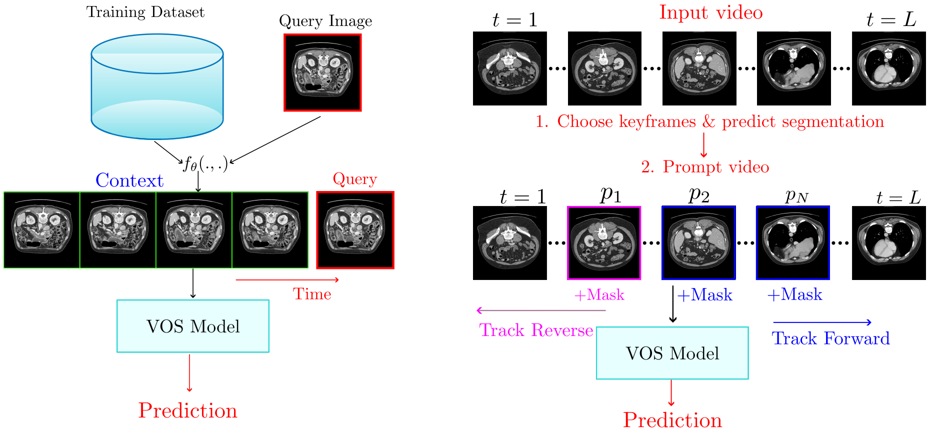
**Abstract.** In-context learning (ICL) has shown promise for generalizing to new visual tasks using a few examples, but current methods are limited. They typically rely on a rigid gridding strategy that restricts the number and resolution of context images. We propose **Temporal**, a novel approach that overcomes these limitations by reformulating visual ICL as a video object segmentation (VOS) problem. This VOS-based approach naturally handles a variable number of full-resolution context images. To automatically select the most relevant context for a given query, we introduce a prompt retriever pretrained on videos using a time-contrastive objective. This objective learns from the temporal coherence of video, using adjacent frames as positive examples (i.e., useful context images) and distant frames as negatives. For image segmentation, our retriever builds a pseudo-video by prepending the retrieved context images to the query image, which is then processed by the VOS model. For video segmentation, the retriever identifies keyframes, our ICL pipeline generates their masks, and these masks are propagated through the video. On the MICCAI FLARE 2022 challenge, Temporal significantly outperforms baselines, achieving a Dice score of 90.95% for image segmentation (+10.64%) and 92.45% for video segmentation (+14.88%).

**Keywords:** Time-Contrastive Learning · In-Context Learning · Video Object Segmentation

## 1 Introduction

Medical imaging generates vast quantities of unlabeled data, yet acquiring dense annotations remains costly. This challenge has motivated the development of approaches requiring fewer labels, such as *self-supervised learning* [2,5,13,9,10,1], *semi-supervised learning* [17], and *transfer learning* [20]. Recently, foundation models have introduced new paradigms such as *visual prompting* and *in-context learning* (ICL) [21,7,8,3,22], enabling adaptation to new tasks with minimal or no fine-tuning.

In-context learning (ICL)—where models perform new tasks or generalize to unseen domains by conditioning on input-output examples without parameter updates—has been remarkably successful in natural language processing [4]. It



a) Image segmentation

b) Video segmentation

**Fig. 1. Model architecture.** (a) *Visual in-context learning via VOS:* The retriever $f_\theta$ selects the optimal context set $P$ for a query image $x_q$. The query image $x_q$ is appended to the context to create a video sequence, and the VOS model predicts the segmentation mask for $x_q$ using this sequence. (b) *Video segmentation with ICL:* 1) Select keyframes (Section 2.3), 2) Predict masks for keyframes using step (a), 3) Propagate masks across the video.

offers a compelling vision for medical imaging: a single model could generalize across diverse segmentation tasks by "reasoning" over relevant annotated examples. However, adapting ICL to computer vision presents unique challenges. We highlight two critical issues: (1) selecting the most relevant context images, and (2) effectively modeling interactions between the context set and the query image. An ideal visual ICL model must support a variable number of context images and preserve their full resolution. While textual input-output pairs can be trivially concatenated, it is less clear how to best handle a variable number of images for visual ICL.

Early attempts at visual ICL relied on a *gridding* strategy, arranging input-output image pairs in a grid alongside a query image [3,23,22,18]. This strategy, however, limits the number of context images and necessitates downsampling, reducing image detail—a critical limitation for medical applications. More recent work has focused on context selection. Zhang et al. [23] found that performance is significantly impacted by the choice of context examples and proposed both unsupervised (CLIP-based [14]) and supervised retrieval techniques. Subsequent innovations included padding images with learnable vectors [22] and combining channel and spatial features for computing similarity [16].

A promising alternative emerged when Foster et al. [8] framed ICL as a video object segmentation (VOS) task. In this approach, context images are treated as frames in a synthetic video, and a VOS model [6] segments the final frame (the query image). This allows for handling full-resolution images and variable-sized

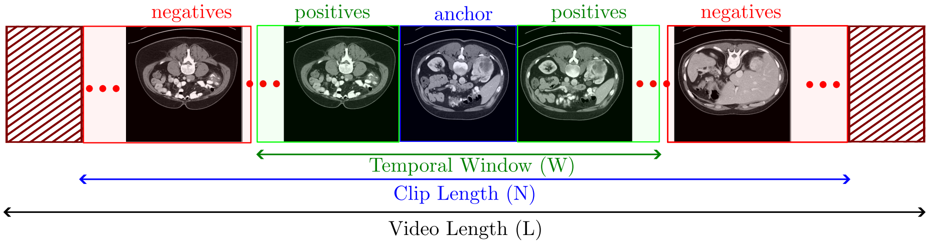


**Fig. 2.** *Multi-positive sampling strategy.* For each anchor image in a clip, we sample $M$ temporally adjacent frames as positives and all other frames in the clip as negatives. Shaded regions are frames that were not sampled in a minibatch.

context sets. However, this method relied on pre-trained CLIP embeddings for context retrieval, which may not be optimal for unseen domains like medical imaging.

Our work, named **Temporal**, extends the VOS-based ICL framework by introducing a self-supervised, time-contrastive approach for context retrieval and a unified pipeline for both image and video segmentation. We pretrain a retriever on videos using a multipositive time-contrastive objective, which leverages temporal coherence as a natural supervisory signal. This avoids the need for manual labels by treating adjacent frames as positive examples. Furthermore, we provide a comprehensive analysis of context selection and an efficient keyframe-based method for video segmentation. To our knowledge, this is the first work to combine time-contrastive networks with a multipositive cross-entropy loss for self-supervised learning from videos. Our approach achieves up to 92.45% Dice on video benchmarks—a 14.88% improvement over semi-supervised VOS baselines. Figure 1 illustrates our approach. The code is available at github.com/aswahd/Temporal.

## 2 Method

Our method consists of three main components: (1) a formulation of visual ICL as a video object segmentation (VOS) task, (2) a prompt retriever pretrained with a multipositive time-contrastive objective, and (3) an inference pipeline for both image and video segmentation. We use SAM2 [15] as our base VOS model.

### 2.1 Visual In-Context Learning as VOS

Given a dataset of input-output pairs $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$ and a query $x_q$, ICL aims to predict $\hat{y}_q$ by conditioning on a context set $P = \{(x_{c_1}, y_{c_1}), \ldots, (x_{c_K}, y_{c_K})\} \subseteq \mathcal{D}$. To overcome the limitations of gridding strategies, we frame ICL as a VOS problem. The context set $P$ and the query $x_q$ are treated as frames in a synthetic video, where each pair $(x_{c_k}, y_{c_k})$ is a labeled frame and $x_q$ is the final, unlabeled frame. A VOS model $\mathcal{V}_\phi$ then segments $x_q$:

$$\mathcal{V}_\phi(P, x_q) = \mathcal{V}_\phi\big(x_{c_1}, y_{c_1}, \ldots, x_{c_K}, y_{c_K}, x_q\big). \tag{1}$$

The selection of the optimal context set $P$ is handled by our prompt retriever, described next.

### 2.2 Multipositive Time-Contrastive Pretraining

Our objective is to sample multiple positive examples along with a substantial number of negatives for each anchor frame in a video, and to train the network using a contrastive loss. We formalize this as follows:

*Sampling Strategy:* For each video $V$ of length $L$, we sample a contiguous clip $C = \{t_{\text{start}}, t_{\text{start}} + 1, \ldots, t_{\text{start}} + N - 1\}$, where $N$ denotes the clip length and $t_{\text{start}} \sim \mathcal{U}(1, L - N)$. All frames in the clip are used in a minibatch.

*Generating Label Matrix:* For each anchor frame $i \in C$, positive pairs are defined as frames within a temporal window $\mathcal{W}_i = [\max(1, i - \lfloor W/2 \rfloor), \min(N, i + \lfloor W/2 \rfloor)]$, where $W$ specifies the window size. We sample $M$ frames from the window. This generates a binary label matrix $\mathbf{A} \in \{0,1\}^{N \times N}$:

$$\mathbf{A}_{i,j} = \begin{cases} 1 & \text{if } |i - j| \leq \lfloor W/2 \rfloor \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

To prevent trivial self-similarity matches, diagonal elements $\mathbf{A}_{i,i}$ are masked to 0.

*Multi-View Augmentation:* We generate two augmented views $\tilde{C}_1$ and $\tilde{C}_2$ through random geometric (e.g., translation) and photometric (e.g., Gaussian blurring) transformations. The label matrix expands to encode multi-view relationships:

$$\hat{\mathbf{A}} = \begin{bmatrix} \mathbf{A} & \mathbf{A} + \mathbf{I}_N \\ \mathbf{A} + \mathbf{I}_N & \mathbf{A} \end{bmatrix} \tag{3}$$

where $\mathbf{I}_N$ denotes the $N \times N$ identity matrix. This formulation ensures that:

- Temporal neighbors within each view remain positives
- Temporal neighbors across views are positives
- Two views of the same frame are positives (i.e., $+\mathbf{I}_N$).

For minibatch training, we sample $B$ video clips, and extend the above label matrix to consider frames from different videos as negatives. The final label matrix is a block diagonal matrix with $B$ blocks. See Fig. 1 for the overall architecture and 2 for an illustration of a sampled minibatch.

*Multipositive Time-Contrastive Loss:* Each anchor, in a minibatch of shape $(B \times N, C, H, W)$, has $2M - 1$ positives and $2B(N - M)$ negatives. The encoder $f_\theta$ projects the input batch into a set of embeddings $\mathbf{z} = \{z_1, z_2, \ldots, z_{2BN}\}$.

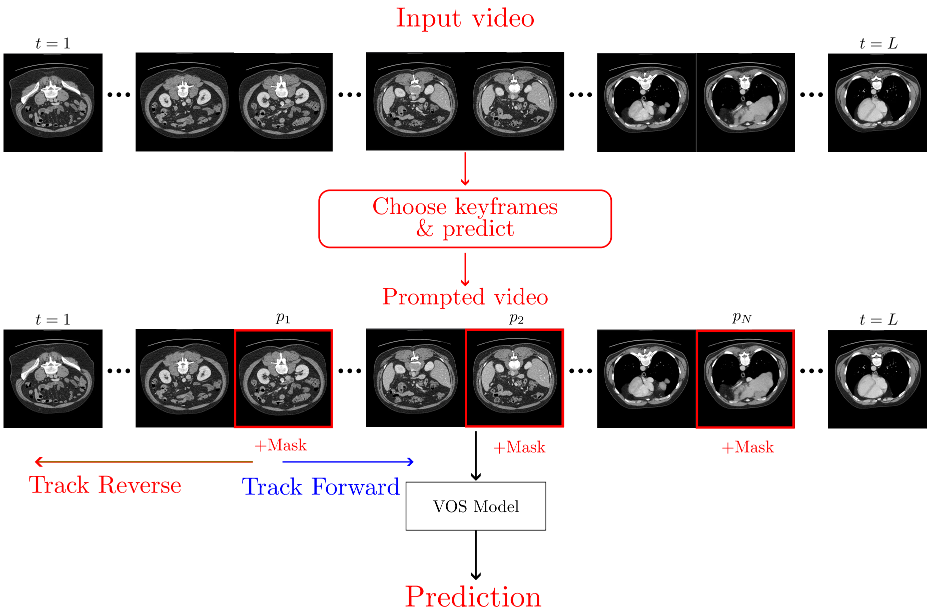


**Fig. 3.** Inference pipeline for video segmentation.

Each embedding is L2 normalized to obtain $\hat{z}_i = z_i/\|z_i\|$. The multipositive contrastive loss is defined as:

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{2BN}\sum_{i=1}^{2BN}\frac{1}{|P(i)|}\sum_{j\in P(i)}\log\frac{\exp\left(\frac{\hat{z}_i^\top \hat{z}_j}{\tau}\right)}{\exp\left(\frac{\hat{z}_i^\top \hat{z}_j}{\tau}\right)+\sum_{k\in N(i)}\exp\left(\frac{\hat{z}_k^\top \hat{z}_j}{\tau}\right)}, \tag{4}$$

where $P(i)$ denotes the set of positive samples for the $i$-th anchor, $N(i)$ represents the corresponding set of negatives, and $\tau$ is the temperature parameter.

### 2.3 Inference

*Image segmentation:* For a given test image $x_q$, we first compute its embedding $z_q = f_\theta(x_q)$ and retrieve the top-$K$ most similar images from the training database $\mathcal{D}$ based on cosine similarity. These retrieved contexts are then used to construct a synthetic video sequence $V = \{x_{c_1}, \ldots, x_{c_K}, x_q\}$, with the query image $x_q$ appended as the final frame. Finally, the VOS model $\mathcal{V}_\phi$ processes this sequence to produce the segmentation mask $\hat{y}_q$ for the query image, as illustrated in Fig. 1(a).

*Video Segmentation:* Given an input video $V = \{x_1, x_2, \ldots, x_L\}$ with $L$ frames, our goal is to generate segmentations $\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_L\}$. We can apply the image-based inference procedure frame-by-frame, and concatenate the results but this is computationally expensive for long sequences and it doesn't consider temporal coherence when making predictions. Instead, we propose a keyframe-based approach that leverages the temporal structure of videos. The inference proceeds as follows:

*1. Choose frames that are most useful if used as prompts (keyframes).* First, each frame $x_t$ is mapped to an embedding $z_t = f_\theta(x_t) \in \mathbb{R}^d$. We then measure how "similar" each frame is to the training database by computing $s_t = \max_{d \in \mathcal{D}} \frac{z_t \cdot d}{\|z_t\| \, \|d\|}$. We form a candidate set $\mathcal{S} \subseteq V$ of size $K$ by selecting the top-$K$ frames with the highest similarity scores $\{s_t\}$. Intuitively, these are the frames that best match the training data distribution and thus are more likely to yield accurate segmentations using the visual ICL (Figure 1).

*2. Choose $Q$ representative frames from $\mathcal{S}$.* Although all frames in $\mathcal{S}$ are strongly matched to the training set, some may be temporally redundant. Hence, we select a subset of size $Q$ that is more diverse. Let $\mathcal{S} = \{x_{(1)}, x_{(2)}, \ldots, x_{(K)}\}$ be sorted in descending order by $s_t$. We impose a minimum temporal distance $\tau$ between chosen frames to encourage coverage of different time points. Starting with $\mathcal{Q} = \{x_{(1)}\}$, we add $x_{(i)}$ to $\mathcal{Q}$ if $\min_{x_j \in \mathcal{Q}} \big| t_i - t_j \big| > \tau$, where $t_i$ is the frame index of $x_{(i)}$. This process continues until we have selected $Q$ frames or exhausted $\mathcal{S}$.

*3. Predict segmentation and filter out non-confident predictions.* For each chosen frame $x_q \in \mathcal{Q}$, we apply the image inference procedure (read above or Figure 1(b)) to obtain a predicted segmentation mask $\hat{y}_q$ and an associated confidence score $c_q$: $(\hat{y}_q, c_q) = \text{Temporal}(x_q, \mathcal{D})$.. Since some predictions may be unreliable, we only keep those that exceed a threshold $\gamma$: $P = \big\{ \big(x_q, \hat{y}_q\big) \mid c_q \geq \gamma \big\}$. Hence, $P$ contains high-confidence and representative prompts that will guide the subsequent video-wide segmentation.

*4. Propagate predictions for the rest of the video.* We feed the retained prompts $P$ into a promptable VOS model $\mathcal{V}_\phi$. Since the selected keyframes may be in the middle of the video, we propagate the prompts in both forward and reverse directions to generate segmentations for the entire video: $\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_L\} = \mathcal{V}_\phi\big(V, P\big)$. The process is illustrated in Figure 3.

*Fine-tuning:* To bridge the domain gap between natural and medical images, we fine-tuned all parameters of the underlying VOS model using "synthetic" videos constructured from 2D images in the training set. For each image $\mathbf{x}_i$, we: 1) select top-$K$ context images via $f_\theta$, 2) construct video $V_i = \{\mathbf{x}_1, \ldots, \mathbf{x}_K, \mathbf{x}_Q\}$ sorted by similarity, and 3) train $\mathcal{V}_\phi$ on random clips $v_i \subset V_i$ to minimize $\min_\phi \mathcal{L}_{seg}(\mathcal{V}_\phi(v_i), y)$. Fine-tuning significantly improves both image and video segmentation as shown in Tables 1 & 2.

## 3 Experiments

*Representative methods.* We compare against grid-based [23] and VOS-based ICL [8]. We adopt SAM-2 as the video foundation model since it represents the current SOTA in generic video segmentation.

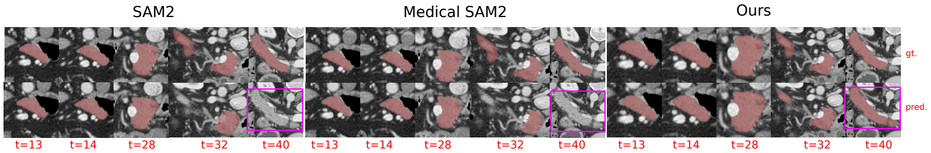


**Fig. 4.** *Qualitative segmentation results* on Pancreas. Our approach (right) produces more accurate organ segmentations and tracking compared to baselines. The baselines fail to track the organ after $t = 32$.

*Datasets.* We evaluate our approach on the MICCAI FLARE 2022 dataset, a benchmark dataset for CT scan segmentation. The dataset includes both annotated scans and a substantial collection of unlabeled data, which we utilize for self-supervised pretraining. Specifically, we use 40 scans to construct the context set and 10 scans for testing. These 10 scans comprise 922 slices containing foreground annotations.

*Implementation Details.* Our time-contrastive pretraining employs a ResNet-50 [11] backbone. We sample 16-frame video clips with a stride of 2 and apply random augmentations (e.g., rotation, horizontal flipping, and Gaussian blurring). For the cross entropy loss, the temperature is set to $\tau = 0.1$, and two adjacent frames are selected as positives per anchor. Training runs for 500 epochs with a minibatch of 4 clips (resulting in a batch size of $16 \times 2 \times 4 = 128$ images) using the AdamW optimizer [12] (learning rate $lr = 10^{-3}$ and weight decay $10^{-4}$). The hyperparameters were chosen via grid search over $lr \in \{10^{-4}, 10^{-3}, 10^{-2}\}$ and $\tau \in \{0.1, 0.5, 1.0\}$. The VOS model utilizes a SAM2-B+ backbone [15]. For image segmentation, we empirically found that 10 context images work best (see Table 1). For video segmentation, we sample 20 keyframes. Fine-tuning is performed for 100 epochs with a learning rate of $10^{-5}$, a batch size of 1, and the AdamW optimizer.

### 3.1 Results

*Image segmentation:* The grid-based method with unsupervised context retrieval performs poorly on medical images, achieving only around 5% Dice score. So, we abandon this approach in favor of VOS-based methods. Table 1 shows that Temporal outperforms unsupervised prompt retrieval methods including DINOv2 [13], CLIP [14], and MedCLIP [19]. Our method achieves 83.55% Dice versus 80.31% for the best baseline. After fine-tuning, Temporal reaches **90.95%** Dice, a **10.64%** improvement. Performance gains are particularly significant for challenging organs like Pancreas (79.53% versus baseline's 49.10%).

*Video Segmentation:* We evaluate on 10 abdominal CT videos (average length: 95 frames) comparing two paradigms:

- **Semi-supervised baselines**: SAM-2 [15] and Medical SAM-2 [24] require first-frame manual prompts. So, we provide ground-truth masks for the first annotated slice of each organ.

**Table 1. Image segmentation** results (Dice, %) of Temporal and baselines on the FLARE 2022 dataset. "ctx" denotes the number of context images used. All baseline models are evaluated with ctx=10. (RK: Right Kidney, Spleen, Panc: Pancreas, Aorta, IVC: Inferior Vena Cava, Gall: Gallbladder, Eso: Esophagus, Stom: Stomach, LK: Left Kidney). Fine-tuning (ft) improves performance.

| Model | Liver | RK | Spleen | Panc. | Aorta | IVC | Gall. | Eso. | Stom. | LK | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Grid-based methods** | | | | | | | | | | | |
| Zhang et al. [23] (CLIP) | | | | | | N/A | | | | | |
| **VOS-based methods** | | | | | | | | | | | |
| Foster et al. [8] (Random, 5 evals) | 0.8999 | 0.8904 | 0.8561 | 0.3006 | 0.9278 | 0.6998 | 0.5378 | 0.4127 | 0.6349 | 0.9003 | 0.7060 |
| Foster et al. [8] (CLIP) | 0.9234 | 0.8968 | 0.9136 | 0.4886 | 0.9490 | 0.8252 | 0.6142 | 0.6434 | 0.7291 | 0.8583 | 0.7842 |
| Foster et al. [8] (MedCLIP) | 0.8950 | 0.8486 | 0.8605 | 0.4464 | 0.9535 | 0.8207 | 0.6647 | 0.6758 | 0.7501 | 0.8097 | 0.7724 |
| Foster et al. [8] (DINOv2) | 0.9161 | 0.9003 | 0.8815 | 0.4910 | 0.9485 | 0.8432 | 0.7519 | 0.6730 | 0.7509 | 0.8747 | 0.8031 |
| **Ours** | | | | | | | | | | | |
| Temporal (ctx=5) | 0.9088 | 0.9003 | 0.9222 | 0.6105 | 0.9526 | 0.8185 | 0.7579 | 0.7096 | 0.7194 | 0.8881 | 0.8188 |
| Temporal (ctx=10) | 0.9243 | 0.9105 | 0.9221 | 0.6211 | 0.9560 | 0.8364 | 0.8130 | 0.7228 | 0.7499 | 0.8991 | 0.8355 |
| Temporal (ctx=5,ft) | 0.9504 | 0.9115 | 0.9578 | 0.7872 | 0.9669 | 0.9204 | 0.8361 | 0.8619 | 0.8585 | 0.8980 | 0.8949 |
| Temporal (ctx=10, ft, topk) | 0.9506 | 0.9032 | 0.9371 | 0.7860 | 0.9669 | 0.9214 | 0.8598 | 0.8772 | 0.8671 | 0.8958 | 0.8965 |
| Temporal (ctx=10,ft) | 0.9631 | 0.9202 | 0.9650 | 0.7953 | 0.9671 | 0.9218 | 0.9073 | 0.8697 | 0.8764 | 0.9090 | 0.9095 |

**Table 2. Video segmentation** results (Dice, %). SAM2 and Medical SAM2 are evaluated in a semi-supervised setting as they require prompts to propagate. Temporal can automatically select context without supervision and achieves superior performance.

| Method | Liver | RK | Spleen | Panc. | Aorta | IVC | Gall. | Eso. | Stom. | LK | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Baselines (Semi-supervised)** | | | | | | | | | | | |
| SAM-2 (mask prompt) | 0.9121 | 0.9638 | 0.8845 | 0.3557 | 0.9497 | 0.8021 | 0.8737 | 0.6497 | 0.4395 | 0.9257 | 0.7757 |
| Medical SAM-2 (mask prompt) | 0.9070 | 0.8991 | 0.9650 | 0.3402 | 0.8932 | 0.7828 | 0.7238 | 0.6970 | 0.6786 | 0.8358 | 0.7723 |
| **Ours** | | | | | | | | | | | |
| Temporal (ctx=10) | 0.9263 | 0.8867 | 0.9280 | 0.5970 | 0.9526 | 0.8633 | 0.8105 | 0.6281 | 0.7149 | 0.8853 | 0.8193 |
| Temporal (ctx=10, fine-tuned) | **0.9739** | **0.9625** | **0.9532** | **0.8132** | **0.9669** | **0.9219** | **0.8928** | **0.8686** | **0.9118** | **0.9581** | **0.9223** |
| Temporal (fine-tuned, mask prompt) | **0.9765** | **0.9625** | **0.9837** | **0.8203** | **0.9690** | **0.9351** | **0.8927** | **0.8875** | **0.8586** | **0.9586** | **0.9245** |

- **Automated VOS**: Temporal automatically identifies the target organs without manual prompts, and tracks them throughout the entire sequence.

Table 2 highlights our method's superiority: Temporal achieves **81.93%** Dice fully automatically compared to SAM2's 77.57% with manual prompts. After fine-tuning (Section 2), Temporal achieves **92.23%** Dice, outperforming Medical SAM-2 (77.83%) by **14.4%**, and **92.45%** in semi-supervised segmentation. The improvement is particularly dramatic on challenging organs like Pancreas, where the best baseline achieves only 35.57% Dice compared to Temporal's **82.03%** after fine-tuning. See Fig. 4 for qualitative results.

*Impact of Context Size.* To evaluate the effect of the number of annotated context images, we varied the number of context examples in our experiments. As shown in Table 1, increasing the number of context images results in improved performance: for example, 74.70% Dice with 5 context images versus 81.83% Dice with 10.

*Diversity-Aware Context Selection.* We compare standard top-$K$ context retrieval with a diversity-aware selection strategy. Our two-stage approach: (1) retrieves top-$K$ most similar training images to the query using cosine similarity on SSL features, then (2) greedily selects top-$Q$ diverse candidates. The core idea is to minimize redundancy in the context set, ensuring each selected example contributes unique information. For each candidate $\mathbf{z}_j$ in the $K$-subset, we iteratively select the sample maximizing:

$$\text{score}(z_j) = \text{sim}(z_j, z_{\text{query}}) - \lambda \cdot \max_{\mathbf{z}_k \in \mathcal{S}} \text{sim}(z_j, z_k),$$

where $\mathcal{S}$ contains already selected samples and $\lambda$ balances relevance versus diversity. In our experiments, $\lambda = 0.7$ was determined through validation. The results in Table 1 show that diverse sampling improves Dice by $\uparrow 2.1\%$ (context size 5) and $\uparrow 1.30\%$ (size 10) over top-$K$ selection, highlighting the importance of context diversity.

## 4 Conclusion

We introduced Temporal, a self-supervised learning objective for pretraining a prompt retriever, and we formulated visual in-context learning as a video object segmentation. When evaluated on the MICCAI FLARE 2022 dataset, our approach demonstrates substantial improvements, achieving a 9.23% and 14.88% increase in Dice scores in image and video segmentation, respectively.

**Acknowledgments.** Dr. Jacob Jaremko is a Canada CIFAR AI Chair, and his academic time is made available by Medical Imaging Consultants (MIC), Edmonton, Canada. This work was supported by grants from TD Ready Grant, IC-IMPACTS, One Child Every Child, the Arthritis Society, and Alberta Innovates.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture (Apr 2023). https://doi.org/10.48550/arXiv.2301.08243
2. Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y., Schwarzschild, A., Wilson, A.G., Geiping, J., Garrido, Q., Fernandez, P., Bar, A., Pirsiavash, H., LeCun, Y., Goldblum, M.: A Cookbook of Self-Supervised Learning (Jun 2023). https://doi.org/10.48550/arXiv.2304.12210
3. Bar, A., Gandelsman, Y., Darrell, T., Globerson, A., Efros, A.A.: Visual Prompting via Image Inpainting

4. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language Models are Few-Shot Learners (Jul 2020). https://doi.org/10.48550/arXiv.2005.14165
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations (Jun 2020). https://doi.org/10.48550/arXiv.2002.05709
6. Cheng, H.K., Schwing, A.G.: XMem: Long-Term Video Object Segmentation with an Atkinson-Shiffrin Memory Model (Jul 2022). https://doi.org/10.48550/arXiv.2207.07115
7. Denner, S., Bujotzek, M., Bounias, D., Zimmerer, D., Stock, R., Jäger, P.F., Maier-Hein, K.: Visual Prompt Engineering for Medical Vision Language Models in Radiology (Aug 2024)
8. Foster, T., Croitoru, I., Dorfman, R., Edlund, C., Varsavsky, T., Almazán, J.: Flexible visual prompts for in-context learning in computer vision (Dec 2023). https://doi.org/10.48550/arXiv.2312.06592
9. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked Autoencoders Are Scalable Vision Learners. https://arxiv.org/abs/2111.06377v3 (Nov 2021)
10. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum Contrast for Unsupervised Visual Representation Learning (Mar 2020)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition (Dec 2015). https://doi.org/10.48550/arXiv.1512.03385
12. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization (Jan 2019). https://doi.org/10.48550/arXiv.1711.05101
13. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision (Feb 2024)
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision (Feb 2021). https://doi.org/10.48550/arXiv.2103.00020
15. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: SAM 2: Segment Anything in Images and Videos
16. Sun, Y., Chen, Q., Wang, J., Wang, J., Li, Z.: Exploring Effective Factors for Improving Visual In-Context Learning (Apr 2023). https://doi.org/10.48550/arXiv.2304.04748
17. van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. Machine Learning **109**(2), 373–440 (Feb 2020). https://doi.org/10.1007/s10994-019-05855-6
18. Wang, X., Wang, W., Cao, Y., Shen, C., Huang, T.: Images Speak in Images: A Generalist Painter for In-Context Visual Learning (Mar 2023). https://doi.org/10.48550/arXiv.2212.02499
19. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: Contrastive Learning from Unpaired Medical Images and Text (Oct 2022). https://doi.org/10.48550/arXiv.2210.10163

20. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. Journal of Big Data **3**(1), 9 (May 2016). https://doi.org/10.1186/s40537-016-0043-6
21. Yu, B.X., Chang, J., Wang, H., Liu, L., Wang, S., Wang, Z., Lin, J., Xie, L., Li, H., Lin, Z., Tian, Q., Chen, C.W.: Visual Tuning. ACM Computing Surveys **56**(12), 1–38 (Dec 2024). https://doi.org/10.1145/3657632
22. Zhang, J., Wang, B., Li, L., Nakashima, Y., Nagahara, H.: Instruct Me More! Random Prompting for Visual In-Context Learning (Nov 2023). https://doi.org/10.48550/arXiv.2311.03648
23. Zhang, Y., Zhou, K., Liu, Z.: What Makes Good Examples for Visual In-Context Learning? (Feb 2023). https://doi.org/10.48550/arXiv.2301.13670
24. Zhu, J., Hamdi, A., Qi, Y., Jin, Y., Wu, J.: Medical SAM 2: Segment medical images as video via Segment Anything Model 2 (Dec 2024). https://doi.org/10.48550/arXiv.2408.00874