EvoCDS: UNIFYING VIDEO CONTEXT-DEPENDENT CONCEPT SEGMENTATION VIA EVOLVING CONCEPT DISCRIMINATIVE SUBSPACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified image context-dependent (ICD) concept segmentation has made substantial progress, yet extending it to video requires more than temporal modeling. Across natural videos, medical videos, and 3D medical scenarios, video context-dependent (VCD) concept segmentation must handle heterogeneous concepts whose discriminative evidence evolves with the visual context, together with severe cross-task imbalance in video and frame counts. We characterize this temporal evolution as concept drift and propose ***EvoCDS***, the first unified framework specifically designed for VCD concept segmentation. Motivated by the compact low-dimensional structure of concept-specific discriminative information, ***EvoCDS*** constructs an evolving concept discriminative subspace that preserves reliable historical directions while adapting to current evidence. We further introduce temporal clip balancing to stabilize joint learning across heterogeneous sequence datasets. Across eight tasks spanning natural and medical scenarios, ***EvoCDS*** consistently outperforms existing unified models and achieves competitive performance against specialized methods. We also construct ***Video-CD***, a benchmark of 201 videos and 34,562 frames for evaluating robustness to concept drift. Together, ***EvoCDS*** and ***Video-CD*** provide a unified framework and benchmark for studying evolving CD concepts in dynamic visual scenarios.

1 INTRODUCTION

In philosophy and cognitive science, concepts are commonly categorized as context-independent (CI) or context-dependent (CD) (Barsalou, 1982; Martial et al., 2018; Lachmann & Van Leeuwen, 2005). Since Spider (Zhao et al., 2024c) first unified image context-dependent (ICD) concept segmentation tasks, unified CD concept segmentation has received increasing attention (Zhao et al., 2026; Liang et al., 2026; Zhao et al., 2024b; Zhang et al., 2026). Most existing unified ICD concept segmentation methods (Chang et al., 2025; Zhao et al., 2024c; Xu et al., 2026b; Lu et al., 2025) model semantic relationships in static images, primarily capturing spatial dependencies between objects and context within a single frame. Video context-dependent (VCD) concept segmentation, however, introduces an additional temporal dimension. Whether a target is perceived as a salient object (Fan et al., 2019), camouflaged object (Yao et al., 2026), shadow (Chen et al., 2021), or lesion region (Yang et al., 2024) (§ A) depends on both spatial context and temporal cues such as inter-frame motion, scene evolution, and cross-frame consistency.

We argue that the central challenge of VCD concept segmentation is therefore not temporal modeling alone, but learning concept representations whose discriminative evidence evolves with the visual context. Existing studies often attribute errors to factors such as occlusion, deformation, scene transitions, and appearance variation. We instead view these seemingly distinct phenomena from a unified perspective: the discriminative evidence associated with a CD concept evolves over time, giving rise to concept drift (§ B). As illustrated in Fig. 1(a), concept drift may manifest as appearance drift, semantic shift, shape drift, or occlusion, altering the relationship among the target, non-target regions, and reference context. In practice, scene transitions, target disappearance, occlusion, and severe deformation can change the evidence available for recognizing a target while leaving its semantic identity unchanged. Such changes reshape foreground–background separability over time. Existing methods typically rely on fixed concept representations or fixed foreground–background discrimina-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

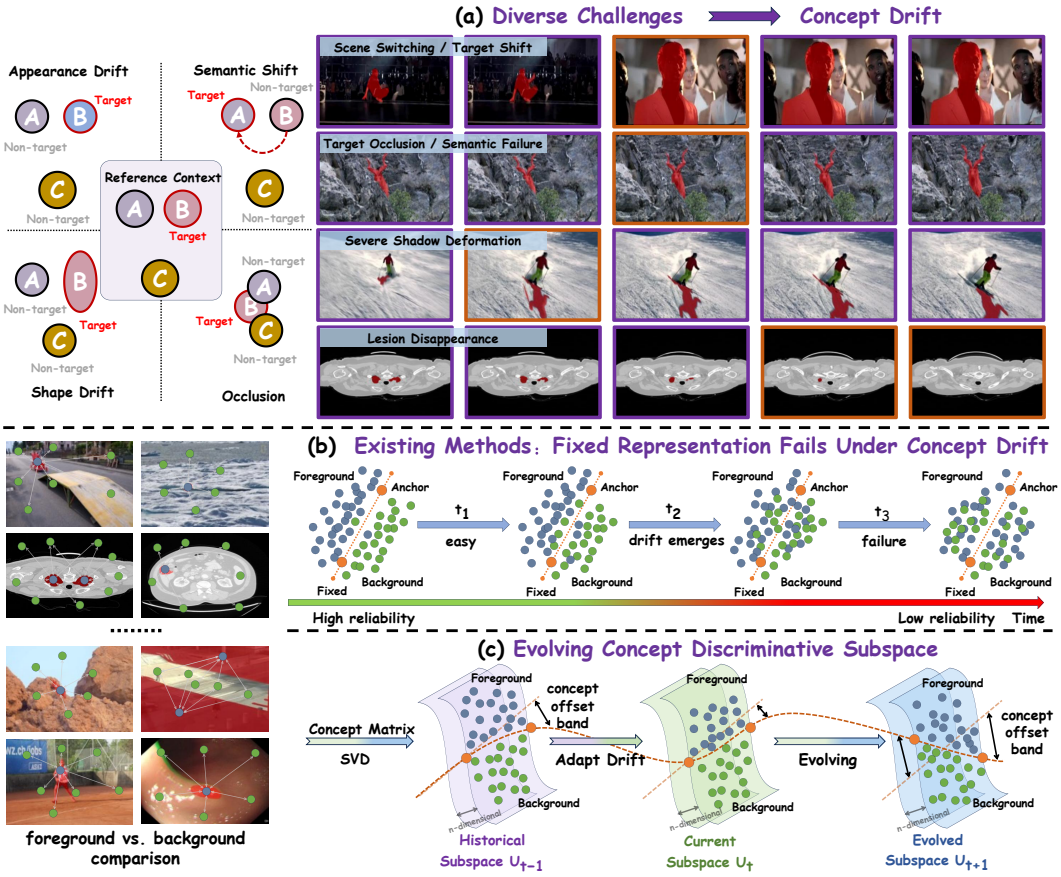

Figure 1: (a) Temporal concept drift of context-dependent (CD) concepts. (b) Existing methods rely on fixed representations, limiting their adaptability to concept drift. (c) The proposed evolving concept discriminative subspace.

tive relationships throughout a video. As shown in Fig. 1(b), these representations may be reliable at early time steps but become progressively misaligned as concept drift emerges, eventually leading to segmentation failure. From this perspective, VCD concept segmentation can be formulated as learning temporally evolving CD concept representations that adapt to scene dynamics.

In this paper, we propose ***EvoCDS*** for unified VCD concept segmentation. In CD concept segmentation, a feature vector encodes shared visual information, while different concepts rely on different components or directions of this representation for discrimination. We decompose the feature representation into two components as $F_{VI} = G_{RI} + T_{RI}$. General representation information (G_{RI}) captures basic semantic, texture, structural, and spatial cues, while task-specific representation information (T_{RI}) captures the discriminative factors associated with the current concept. Furthermore, T_{RI} typically lies on a low-dimensional manifold within the high-dimensional feature space (§ C). Although target appearance, spatial location, and background may vary substantially across frames, the discriminative directions associated with the same CD concept tend to evolve continuously and can therefore be approximated by a small set of dominant directions.

Motivated by this property, ***EvoCDS*** models concept-specific discrimination through an **evolving concept discriminative subspace**. The concept definition stream extracts foreground and background prototypes and constructs their difference vectors to characterize concept discrimination. Local, global, and historical discriminative directions are then organized into a concept matrix, from which a compact subspace is derived. As illustrated in Fig. 1(c), this subspace evolves from U_{t-1} to U_t and further to U_{t+1} by retaining reliable historical directions while incorporating current evidence. For prediction, query features are projected into the subspace and matched against foreground and background prototypes, while high-confidence predictions update the current discriminative direction. The reconstruction error measures its consistency with the historical subspace; when the current direction can no longer be adequately represented, a drift-aware reset discards outdated evidence and reconstructs the subspace. In this way, ***EvoCDS*** continuously adapts concept representations to evolving visual contexts instead of relying on fixed representations.

Beyond concept evolution, unified VCD learning faces severe cross-task imbalance in video and frame counts, motivating temporal clip balancing (TCB) across heterogeneous datasets. For example, liver tumor datasets (Bilic et al., 2023) contain up to 20× more frames than COVID-19 (Jun et al., 2020) and 10× more than shadow datasets (Chen et al., 2021). Without balancing, data-rich tasks dominate gradient updates and underrepresent smaller datasets. TCB enables stable joint optimization while preserving temporal dependencies across VCD scenarios.

Our contributions are summarized as follows:

- We identify concept drift as a central challenge in VCD concept segmentation and formulate the problem as learning temporally evolving concept representations across heterogeneous scenarios.
- We propose ***EvoCDS***, which models concept-specific discrimination through an evolving concept discriminative subspace, enabling segmentation as discriminative evidence changes over time.
- We construct ***Video-CD***, a benchmark specifically designed to evaluate robustness under temporal concept drift, covering eight CD concepts with 201 videos and 34,562 frames.
- Extensive experiments across eight tasks spanning natural videos, medical videos, and 3D medical scenarios validate our formulation and demonstrate that ***EvoCDS*** consistently outperforms existing unified models while remaining competitive with state-of-the-art specialized methods.

2 Related Work

Video Context-dependent (VCD) Concept Segmentation. Unlike image context-dependent (ICD) concept segmentation (Zhao et al., 2024c), VCD concept segmentation requires both spatial and temporal context to identify and segment concept targets in videos. In particular, target appearance, motion, and interactions with the environment jointly determine its recognition and segmentation. In this work, we consider eight VCD concept segmentation tasks across natural and medical scenarios (§ A) to investigate unified modeling of VCD concept segmentation. Existing methods mainly adopt U-shaped encoder-decoder architectures and incorporate techniques such as temporal feature modeling (*e.g.*, optical flow or temporal fusion) (Luo et al., 2026; Zhao et al., 2024a; Liu et al., 2023a; Hu et al., 2024), multi-scale spatio-temporal feature extraction (Chen et al., 2025; Yang et al., 2024; Xing et al., 2025; Ji et al., 2025), boundary refinement (Zhou et al., 2026b; Shaker et al., 2024; Xu et al., 2025), and architecture integration (Pang et al., 2024; Chen et al., 2021; Su et al., 2025). To support cross-domain VCD concept segmentation, we develop a unified framework that handles diverse tasks with shared parameters in a single end-to-end training process.

Unified Vision Models. In recent years, unified segmentation frameworks have shown strong potential for multi-task visual modeling (Hao et al., 2025; Li et al., 2025a; Zhao et al., 2024b; Zhang et al., 2026). Inspired by In-Context Learning (ICL) (Brown et al., 2020; Park et al., 2025; Agarwal et al., 2024), the vision community has also explored visual ICL (Sun et al., 2025b; Li et al., 2025b; Wan et al., 2026). Painter (Wang et al., 2023a) and SegGPT (Wang et al., 2023b) unify segmentation through random coloring, while UniverSeg (Butoi et al., 2023), SR-ICL (Chang et al., 2025), and TP-Seg (Xu et al., 2026b) extend visual ICL to medical image segmentation. Spider (Zhao et al., 2024c) leverages contextual demonstrations to unify ICD concept segmentation across natural and medical scenarios. More recently, ConceptSeg-R1 (Zhao et al., 2026) formulates concept segmentation as rule-induced grounding and employs MetaGRPO to learn transferable rules, together with proxy reasoning and concept translation across three concept hierarchies. Unlike ConceptSeg-R1, which focuses on image-level concept generalization, ***EvoCDS*** addresses VCD concept segmentation by modeling temporal concept drift with an evolving discriminative subspace across heterogeneous video domains.

3 Method

Framework Overview. ***EvoCDS*** consists of two components: *the concept segmentation stream* and *the concept definition stream*. As shown in Fig. 2, the concept segmentation stream adopts a general encoder-decoder architecture (Ronneberger et al., 2015) with different backbones to maximize knowledge sharing while minimizing task-specific designs. Video sequences from all eight tasks are jointly fed into the network, with frames processed sequentially within each sequence. All tasks share the same feature extraction and fusion process, producing a unified representation Q_t, where t denotes the current temporal index. In the concept definition stream, video frames and con-

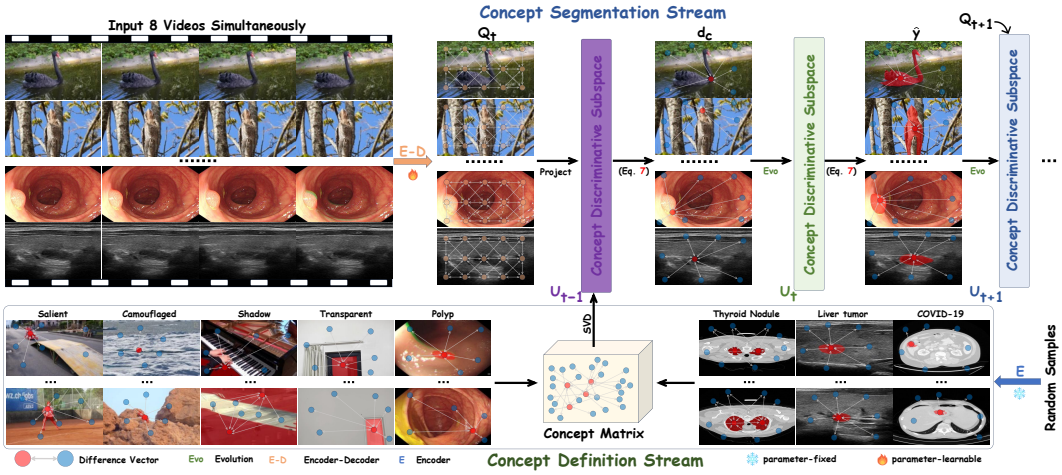

Figure 2: Overall pipeline of ***EvoCDS***: (1) the concept segmentation stream extracts shared features from multi-task frames, and (2) the concept definition stream constructs concept matrices from frames and masks. Shared features are projected into a low-rank concept discriminative subspace, where pixel-wise distances to foreground and background concept prototypes yield logits $\hat{y}$.

cept masks are processed by a frozen encoder to extract visual features, which are aligned with the masks to obtain concept prototypes and construct the concept matrix D. The unified representation Q_t is first passed through the concept discriminative head to obtain preliminary logits y. Based on pixel–prototype relationships, the current difference vector d_c is computed to evolve the concept discriminative subspace. Pixel features are then projected into the evolved subspace, where distances to the updated prototypes are recomputed to produce the final logits $\hat{y}$.

Concept Discriminative Subspace (CDS). To capture task-specific representation T_{RI}, we construct the concept matrix D from the concept definition samples. Unlike previous methods (Wang et al., 2019; Cong et al., 2026a;b) relying on category-level foreground/background prototypes, CD concept segmentation requires modeling foreground and background as relative concepts within the same visual context. Therefore, the key T_{RI} is expected to lie in the discriminative directions separating foreground from background. Specifically, to capture local discriminative information, for the i-th[1] concept definition sample, let p_f^i and p_b^i denote its foreground and background prototypes, respectively. We define the sample-level difference vector as:

$$d_i = p_f^i - p_b^i \in \mathbb{R}^C. \tag{1}$$

To incorporate global discriminative information, we aggregate foreground and background features across all concept definition samples to obtain global prototypes. However, simple averaging is sensitive to local noise, imbalanced annotated regions, and appearance variations across sample sets. We therefore introduce a cross-attention mechanism to refine the foreground and background prototypes with cross-sample contextual information, yielding the global difference vector:

$$d_g = p_f^g - p_b^g \in \mathbb{R}^C. \tag{2}$$

To improve the concept stability during continuous inference, we introduce historical discriminative information. The historical difference vectors $\{d_1^{hist}, \ldots, d_m^{hist}\}$ from previous frames preserve earlier discriminative directions, reducing fluctuations in definition samples estimation and providing temporally consistent concept priors for the current frame. We then construct the task concept matrix by concatenating global, local, and historical difference vectors:

$$D = [\underbrace{d_g}_{\text{global}}, \underbrace{d_1, \ldots, d_i}_{\text{local}}, \underbrace{d_1^{hist}, \ldots, d_m^{hist}}_{\text{historical}}] \in \mathbb{R}^{C \times (1+i+m)}. \tag{3}$$

Since D contains both G_{RI} and T_{RI} and its discriminative information is typically concentrated along a few dominant directions, we decompose D via singular value decomposition (SVD) as $D = U\Sigma V^\top$ and retain the dominant left singular vectors to construct the CDS:

$$U_r = [u_1, \ldots, u_r] \in \mathbb{R}^{C \times r}, \tag{4}$$

[1] $i \leq 12$, please refer to Tab. 4(c) for details.

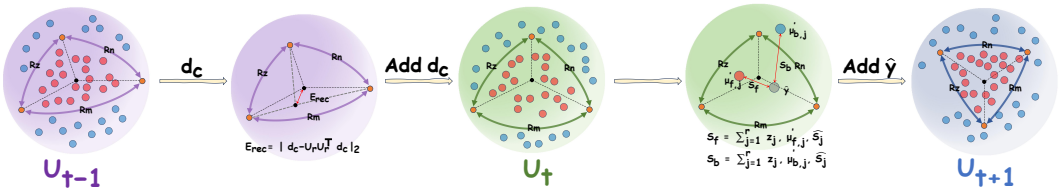

Figure 3: Illustration of the temporal evolution of the concept discriminative subspace ($r = 8$).

where $u_1, \ldots, u_r$ denote the first r columns of U, corresponding to the r largest singular values. We use U_r ($r = 8$) as the basis of the CDS, preserving the dominant discriminative directions while suppressing less informative variations. Given the query feature $Q_t \in \mathbb{R}^{B\times C}$ and the global foreground and background prototypes p_f^g and p_b^g, we project them into the CDS:

$$z \leftarrow \text{Norm.}(Q_t U_r),\ z \in \mathbb{R}^{B\times r}, \quad (\mu_f, \mu_b) \leftarrow \text{Norm.}(p_f^g U_r, p_b^g U_r),\ \mu_f, \mu_b \in \mathbb{R}^{1\times r}, \tag{5}$$

where z denotes the representation of Q_t in the CDS, μ_f and μ_b are the projected foreground and background concept prototypes, and Norm. indicates the normalization (Ba et al., 2016) operation.

Concept Discriminative Head (CDH). Compared with relatively constrained medical scenes, natural scenes exhibit profound dynamic variations and more complex background distractors (§ C). To adapt the CDS to such diverse real-world scenarios, we design a lightweight dynamic CDH. Given the concept feature z, the foreground and background concept prototypes μ_f and μ_b, and the singular values S_r, we construct a comprehensive task-conditioned vector:

$$V = [\mu_f, \mu_b, \mu_f - \mu_b, \mu_f \odot \mu_b, S_r], \tag{6}$$

which is fed into a lightweight MLP to generate dynamically adapted prototypes μ'_f and μ'_b, together with reweighted singular values $\hat{S}_r$. Meanwhile, z is dynamically adjusted through a separate MLP to obtain a feature representation adapted to the task concept. Segmentation is then performed through dynamic concept matching. Specifically, the foreground and background scores at each spatial location are computed as:

$$S_f = \sum_{j=1}^{r} z_j\, \mu'_{f,j}\, \hat{S}_j, \qquad S_b = \sum_{j=1}^{r} z_j\, \mu'_{b,j}\, \hat{S}_j. \tag{7}$$

Finally, the segmentation logits $\hat{y}$ is obtained from the difference between the foreground and background responses, *i.e.*, $\hat{y} = S_f - S_b$.

Concept Subspace Evolution (CSE). To address concept drift, we introduce the CSE mechanism. The predicted logits are converted into a soft mask y using the sigmoid function. Pixels with $y(f) > 0.7$ and $y(b) < 0.3$ are selected as high-confidence foreground and background pixels, respectively, while uncertain pixels are ignored. Given the current query feature Q_t, the foreground and background concept vectors are computed as:

$$y_{\text{fg}}^c = \frac{\sum_i \mathbb{I}(y(f))\, Q_t(f)}{\sum_i \mathbb{I}(y(f)) + \epsilon}, \qquad y_{\text{bg}}^c = \frac{\sum_i \mathbb{I}(y(b))\, Q_t(b)}{\sum_i \mathbb{I}(y(b)) + \epsilon}, \tag{8}$$

where $\mathbb{I}(\cdot)$ is the indicator function and ϵ prevents division by zero. The current difference vector is then computed as $d_c = y_{\text{fg}}^c - y_{\text{bg}}^c$. If either the foreground or background region is empty, d_c is considered invalid and excluded from the concept matrix. For a valid d_c, its reconstruction error (§ H) under the current CDS is computed as (Fig. 3):

$$E_{\text{rec}} = \left\| d_c - U_r U_r^\top d_c \right\|_2. \tag{9}$$

This error measures how well the current discriminative direction is represented by the historical concept basis. It serves as an internal subspace-consistency criterion rather than a standalone frame-level concept-drift classifier. Specifically, μ_{rec} and σ_{rec} denote the mean and standard deviation of historical reconstruction errors, respectively. When $E_{\text{rec}} > \mu_{\text{rec}} + 2\sigma_{\text{rec}}$, the current discriminative direction is considered insufficiently represented by the historical subspace, indicating substantial inconsistency with the accumulated historical evidence and triggering a subspace reset. The historical difference vectors are discarded, and the concept matrix D is reconstructed from the difference

Table 1: Representative datasets and metrics used in our experiments for **natural** and **medical** tasks.

Segmentation Concept	Dataset	Type	#Train	#Test	Metrics
Camouflaged	CAMotion (Yao et al., 2026)	Video	23253	6775	MAE ↓ S_m ↑
Salient	DAVSOD (Fan et al., 2019) and DAVIS (Perazzi et al., 2016)	Video	10638	3419	MAE ↓ S_m ↑
Shadow	ViSha (Chen et al., 2021)	Video	4788	6897	MAE ↓ BER ↓
Transparent	VGSD (Liu et al., 2024)	Video	12314	6850	MAE ↓ BER ↓
Thyroid Nodule	VTUS (Yang et al., 2024)	Video	6651	2691	mDice ↑ mIoU ↑
Polyp	SUN-SEG (Ji et al., 2022; 2021; Fan et al., 2020)	Video	19544	29592	mDice ↑ mIoU ↑
COVID-19	COVID-19 data (Jun et al., 2020)	3D	2562	955	mDice ↑ mIoU ↑
Liver Tumor	LiTS (Bilic et al., 2023)	3D	43975	14661	mDice ↑ mIoU ↑

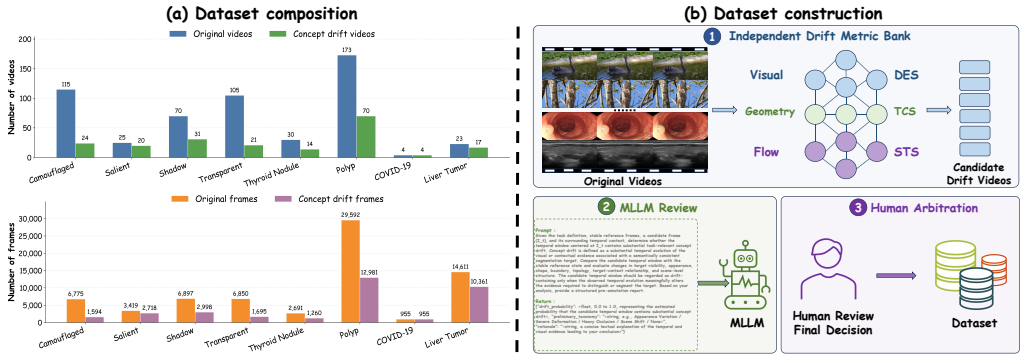

Figure 4: Composition (a) and construction (b) of our ***Video-CD*** benchmark dataset.

vectors of the concept definition samples together with the current difference vector d_c. Otherwise, d_c is incorporated into D. In either case, SVD is recomputed on the resulting concept matrix to update the CDS basis U_r.

Temporal Clip Balancing (TCB). We introduce the TCB strategy to mitigate optimization bias from varying video lengths. For each task, we sample a set of concept definition samples $C_d = \{(x_i, y_i)\}_{i=1}^{N}$ and a concept segmentation clip $C_s = \{x_j\}_{j=1}^{M}$, where x_i and y_i denote a concept definition sample and its annotation, respectively, and x_j denotes a concept segmentation frame. C_d provides the concept prior, while C_s consists of consecutive training samples and serves as an independent unit. Moreover, C_s can span multiple videos, exposing the model to diverse scenes, object states, and concept shifts. This dynamic clip construction provides rich learning signals for modeling concept evolution and complements the proposed CSE mechanism, leading to improved robustness and generalization across heterogeneous VCD concept segmentation tasks. During training, concept definitions from all tasks are aggregated as $C_d \in \mathbb{R}^{N\times 8\times C\times H\times W}$, while segmentation clips are processed in parallel. At each timestamp τ, the model predicts:

$$\hat{y}_\tau = \textit{EvoCDS}(C_s, C_d), \tag{10}$$

The loss (Liu et al., 2023b; He et al., 2023; Zhou et al., 2026a) consists of weighted binary cross-entropy and weighted IoU terms, which are summed across tasks and averaged over time. During inference, the concept segmentation stream extracts visual features, while the concept definition stream retrieves the corresponding concept representations. The historical concept matrix D_h stores historical descriptors to preserve temporal consistency during frame-by-frame processing.

4 Experiment

4.1 Evaluation

Implementation Details. *EvoCDS* is implemented in PyTorch (Paszke et al., 2019) and trained on 8 NVIDIA RTX 5090 GPUs. We use Adam (Kingma & Ba, 2014) with an initial learning rate of 5×10^{-5} and a step learning-rate schedule with a step size of 20 and decay factor of 0.9. Inputs are resized to 512×512, and the model is trained for 35 epochs. Following unified visual models (Chang et al., 2025; Zhao et al., 2024c), we adopt pretrained Swin (Liu et al., 2021) and ConvNeXt (Liu et al., 2022b) backbones. During inference, sequences from all eight tasks are processed simultaneously, with one frame per task at each time step. Concept-definition examples are

Table 2: Comparisons with unified models and state-of-the-art specialized models on eight tasks. For fairness, all unified models are evaluated at an input resolution of 512×512. ‡: first-frame annotations are excluded during both training and testing. The best results are highlighted in **bold**.

Methods	Additional Cues	Backbone	Camouflaged		Salient		Shadow		Transparent		Thyroid Nodule		Polyp		COVID-19		Liver Tumor	
			MAE ↓	S_m ↑	MAE ↓	S_m ↑	MAE ↓	BER ↓	MAE ↓	BER ↓	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑
Specialized Models (One model for one task)																		
ZoomNeXt (Pang et al., 2024)	✓	PVTv2-B5 (Wang et al., 2022)	.033	.779	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CamSAM2‡ (Zhou et al., 2026b)	✓	Hiera-T (Ravi et al., 2024)	.075	.626	-	-	-	-	-	-	-	-	-	-	-	-	-	-
STDNet-A (Su et al., 2025)	✓	Customized Design (Su et al., 2025)	-	-	.087	.755	-	-	-	-	-	-	-	-	-	-	-	-
VSCode-V2 (Luo et al., 2026)	✓	Swin-B (Liu et al., 2021)	-	-	.076	.765	-	-	-	-	-	-	-	-	-	-	-	-
TVSD (Chen et al., 2021)	✓	MiT-B3 (Xie et al., 2021)	-	-	-	-	.033	.157	-	-	-	-	-	-	-	-	-	-
SODA (Liu et al., 2023a)	✓	MiT-B3 (Xie et al., 2021)	-	-	-	-	.029	.136	-	-	-	-	-	-	-	-	-	-
VGSD (Liu et al., 2024)	✓	ResNet-101 (He et al., 2016)	-	-	-	-	-	-	.099	.096	-	-	-	-	-	-	-	-
MVGD (Lu et al., 2026)	✓	SwinV2-B (Liu et al., 2022a)	-	-	-	-	-	-	.082	.081	-	-	-	-	-	-	-	-
FLA-Net (Lin et al., 2023)	✗	ResNet-101 (He et al., 2016)	-	-	-	-	-	-	-	-	.807	.716	-	-	-	-	-	-
Vivim (Yang et al., 2024)	✗	MiT-B3 (Xie et al., 2021)	-	-	-	-	-	-	-	-	.815	.727	-	-	-	-	-	-
SALI (Hu et al., 2024)	✗	PVTv2-B5 (Wang et al., 2022)	-	-	-	-	-	-	-	-	-	-	.833	.753	-	-	-	-
STDDNet (Chen et al., 2025)	✗	Res2Net-50 (Gao et al., 2019)	-	-	-	-	-	-	-	-	-	-	.844	.770	-	-	-	-
Decor-Net (Hu et al., 2023)	✗	Customized Design (Hu et al., 2023)	-	-	-	-	-	-	-	-	-	-	-	-	.961	.940	-	-
UNETR++ (Shaker et al., 2024)	✗	Customized Design (Shaker et al., 2024)	-	-	-	-	-	-	-	-	-	-	-	-	.963	.942	-	-
SegMamba (Xing et al., 2025)	✗	Customized Design (Xing et al., 2025)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.846	.839
LDMP (Ji et al., 2025)	✗	Swin-B (Liu et al., 2021)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.857	.850
Unified Models (One model for all tasks)																		
Spider (Zhao et al., 2024c)	✗	ConvNeXt-L (Liu et al., 2022b)	.034	.776	.079	.763	.036	.151	.086	.091	.810	.722	.825	.742	.965	.946	.871	.849
SR-ICL (Chang et al., 2025)	✗	Swin-B (Liu et al., 2021)	.035	.780	.079	.751	.035	.159	.090	.087	.809	.729	.823	.749	.963	.944	.869	.855
TP-Seg (Xu et al., 2026b)	✗	Hiera-L (Ravi et al., 2024)	.036	.783	.072	.758	.039	.163	.089	.085	.810	.723	.827	.751	.967	.952	.876	.851
SAM3 (Carion et al., 2026)	✗	SAM3 (Carion et al., 2026)	.079	.611	.158	.573	.076	.186	.092	.088	.697	.683	.612	.581	.722	.715	.757	.699
Concept-R1 (Zhao et al., 2026)	✗	SAM3 (Carion et al., 2026)	.029	.808	.071	.769	.072	.185	.076	.075	.792	.703	.813	.725	.736	.723	.802	.783
EvoCDS	✗	ConvNeXt-B (Liu et al., 2022b)	.028	.812	**.064**	.768	.031	.130	.071	.073	.832	.741	.837	.772	.972	.953	.878	.872
EvoCDS	✗	ConvNeXt-L (Liu et al., 2022b)	**.024**	**.837**	**.064**	**.783**	.032	.154	.073	.073	.819	.727	.831	.764	**.974**	**.956**	**.897**	**.886**
EvoCDS	✗	Swin-B (Liu et al., 2021)	.028	.802	.070	.742	.026	**.103**	.069	**.069**	.824	.731	**.848**	**.783**	.970	.951	.876	.870
EvoCDS	✗	Swin-L (Liu et al., 2021)	.026	.818	**.064**	.774	**.025**	.121	**.068**	**.069**	**.833**	**.742**	.832	.765	.970	.951	.894	.882

Table 3: Quantitative comparisons with unified and specialized models on ***Video-CD***.

Methods	Additional Cues	Backbone	Camouflaged		Salient		Shadow		Transparent		Thyroid Nodule		Polyp		COVID-19		Liver Tumor	
			MAE ↓	S_m ↑	MAE ↓	S_m ↑	MAE ↓	BER ↓	MAE ↓	BER ↓	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑
Specialized Models (One model for one task)																		
ZoomNeXt (Pang et al., 2024)	✓	PVTv2-B5 (Wang et al., 2022)	.023	.726	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CamSAM2‡ (Zhou et al., 2026b)	✓	Hiera-T (Ravi et al., 2024)	.080	.603	-	-	-	-	-	-	-	-	-	-	-	-	-	-
STDNet-A (Su et al., 2025)	✓	Customized Design (Su et al., 2025)	-	-	.089	.752	-	-	-	-	-	-	-	-	-	-	-	-
VSCode-V2 (Luo et al., 2026)	✓	Swin-B (Liu et al., 2021)	-	-	.079	.761	-	-	-	-	-	-	-	-	-	-	-	-
TVSD (Chen et al., 2021)	✓	MiT-B3 (Xie et al., 2021)	-	-	-	-	.039	.197	-	-	-	-	-	-	-	-	-	-
SODA (Liu et al., 2023a)	✓	MiT-B3 (Xie et al., 2021)	-	-	-	-	.034	.175	-	-	-	-	-	-	-	-	-	-
VGSD (Liu et al., 2024)	✓	ResNet-101 (He et al., 2016)	-	-	-	-	-	-	.104	.100	-	-	-	-	-	-	-	-
MVGD (Lu et al., 2026)	✓	SwinV2-B (Liu et al., 2022a)	-	-	-	-	-	-	.092	.085	-	-	-	-	-	-	-	-
FLA-Net (Lin et al., 2023)	✗	ResNet-101 (He et al., 2016)	-	-	-	-	-	-	-	-	.814	.724	-	-	-	-	-	-
Vivim (Yang et al., 2024)	✗	MiT-B3 (Xie et al., 2021)	-	-	-	-	-	-	-	-	.830	.741	-	-	-	-	-	-
SALI (Hu et al., 2024)	✗	PVTv2-B5 (Wang et al., 2022)	-	-	-	-	-	-	-	-	-	-	.837	.760	-	-	-	-
STDDNet (Chen et al., 2025)	✗	Res2Net-50 (Gao et al., 2019)	-	-	-	-	-	-	-	-	-	-	.849	.778	-	-	-	-
Decor-Net (Hu et al., 2023)	✗	Customized Design (Hu et al., 2023)	-	-	-	-	-	-	-	-	-	-	-	-	.961	.940	-	-
UNETR++ (Shaker et al., 2024)	✗	Customized Design (Shaker et al., 2024)	-	-	-	-	-	-	-	-	-	-	-	-	.963	.942	-	-
SegMamba (Xing et al., 2025)	✗	Customized Design (Xing et al., 2025)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.792	.785
LDMP (Ji et al., 2025)	✗	Swin-B (Liu et al., 2021)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.810	.804
Unified Models (One model for all tasks)																		
Spider (Zhao et al., 2024c)	✗	ConvNeXt-L (Liu et al., 2022b)	.028	.749	.081	.750	.042	.181	.094	.092	.819	.733	.832	.753	.965	.946	.810	.797
SR-ICL (Chang et al., 2025)	✗	Swin-B (Liu et al., 2021)	.028	.751	.080	.747	.042	.179	.093	.091	.817	.733	.831	.758	.963	.944	.811	.801
TP-Seg (Xu et al., 2026b)	✗	Hiera-L (Ravi et al., 2024)	.027	.755	.075	.754	.041	.175	.090	.089	.828	.731	.830	.754	.967	.952	.822	.812
SAM3 (Carion et al., 2026)	✗	SAM3 (Carion et al., 2026)	.068	.596	.160	.567	.079	.197	.095	.091	.702	.685	.623	.594	.722	.715	.749	.693
Concept-R1 (Zhao et al., 2026)	✗	SAM3 (Carion et al., 2026)	.025	.758	.077	.757	.075	.195	.078	.077	.802	.711	.822	.739	.736	.723	.783	.761
EvoCDS	✗	ConvNeXt-B (Liu et al., 2022b)	.020	.772	.074	.760	.032	.153	**.071**	**.073**	**.840**	**.755**	.846	.783	.972	.953	.818	.809
EvoCDS	✗	ConvNeXt-L (Liu et al., 2022b)	.021	**.785**	**.071**	**.770**	.034	.189	**.071**	.076	.838	.750	.843	.776	**.974**	**.956**	**.860**	**.845**
EvoCDS	✗	Swin-B (Liu et al., 2021)	.020	.759	.079	.739	.031	**.148**	.077	.081	.832	.743	**.856**	**.789**	.970	.951	.841	.832
EvoCDS	✗	Swin-L (Liu et al., 2021)	**.017**	.771	.073	.760	**.029**	.160	.076	.082	.838	.750	.844	.777	.970	.951	.856	.840

independently sampled from corresponding training sets without temporal ordering. Historical discriminative information is disabled for the first 15 epochs and introduced from epoch 16 to support concept subspace evolution. For COVID-19 and liver tumor segmentation, 3D CT volumes are treated as ordered slice sequences, with adjacent slices ordered spatially rather than temporally.

Datasets and Metrics. As summarized in Tab. 1, our experiments cover eight CD concept segmentation tasks across four natural and four medical scenarios, including six video and two 3D medical datasets. These datasets vary in scale and use task-specific metrics, including MAE, S_m, BER, mDice, and mIoU. Beyond these datasets, we construct ***Video-CD*** as a benchmark for evaluating VCD robustness under concept drift. As shown in Fig. 4, an independent drift metric bank retrieves high-recall candidates from discriminative evidence, structural-topological, and temporal-context consistency shifts, avoiding model-dependent sampling bias. ***Video-CD*** contains 201 videos and 34,562 frames across eight concepts, with each video verified to contain substantial task-relevant concept drift. Candidates are reviewed by an MLLM and adjudicated by human annotators.

Quantitative Results. We compare ***EvoCDS*** with recent unified models and specialized methods in Tab. 2. ***EvoCDS*** achieves the best overall performance among unified models across all eight tasks. Without bells and whistles, ***EvoCDS*** also delivers dominant performance against specialized methods, achieving the best results on seven of the eight tasks despite using a single unified model. Notably, these specialized methods are individually designed for specific tasks and often exploit additional task-specific cues, whereas ***EvoCDS*** shares the same modeling framework across all scenarios. This highlights the effectiveness of learning transferable concept-discriminative structures across heterogeneous VCD tasks. Existing unified methods mainly focus on CD concept segmentation in image settings and therefore remain less effective when temporal concept drift emerges. For a focused evaluation under concept drift, we further collect drift-containing videos from the

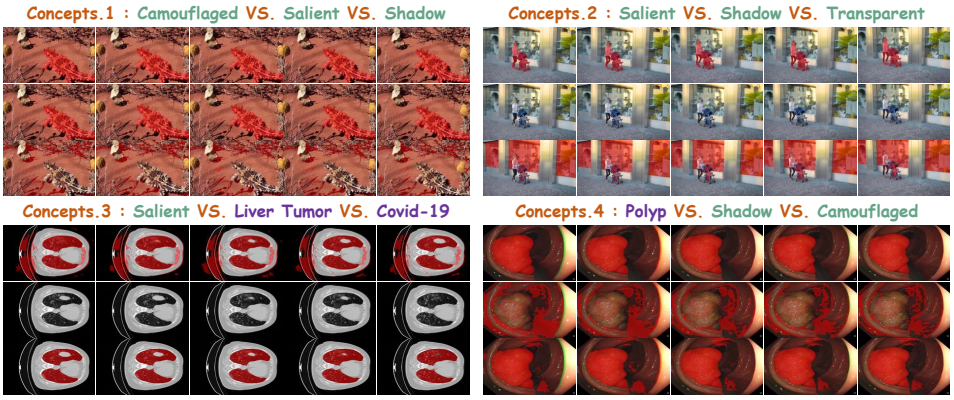

Figure 5: Some visual examples of ***EvoCDS***. More visual results can be found in § D.

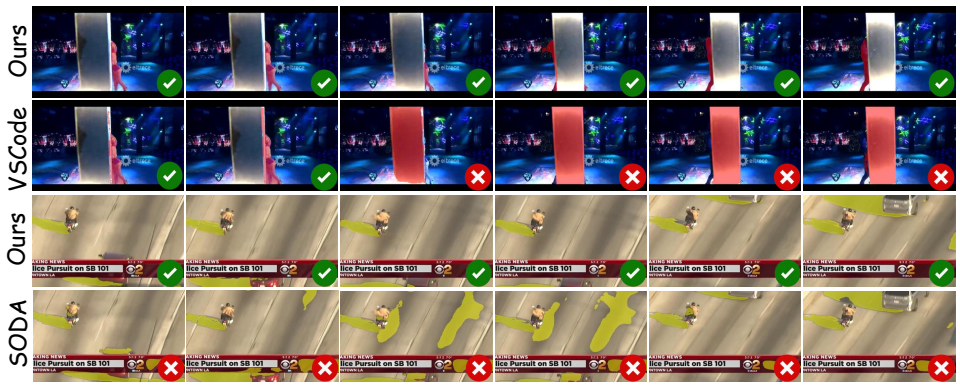

Figure 6: Visual comparison of ***EvoCDS*** with other SOTA methods under concept drift.

evaluation datasets of all eight tasks into ***Video-CD***. As shown in Tab. 3, ***EvoCDS*** consistently outperforms existing unified methods on these challenging sequences, further demonstrating the benefit of evolving concept representations under temporal drift.

Qualitative Results. Fig. 5 shows concept-conditioned predictions of ***EvoCDS*** under different concept definitions. For the same sequence, different concepts produce distinct responses, showing that the shared representation preserves concept-specific discrimination rather than collapsing tasks into a common foreground notion. This behavior holds across natural and medical scenes involving camouflaged, salient, shadow, transparent, and lesion-related concepts. Cross-domain concepts also reveal structural correlations; for example, "camouflaged" responses overlap with low-contrast polyp regions, while "shadow" highlights illumination-related structures in colonoscopy images. These results suggest that unified VCD learning captures transferable structures while retaining concept selectivity. Fig. 6 compares ***EvoCDS*** with specialized methods under concept drift. Competing methods progressively expand to distractors as appearance and context evolve, whereas ***EvoCDS*** maintains stable target discrimination, demonstrating stronger robustness to changing concept evidence. More qualitative results are provided in § D.

4.2 Ablation Study

Joint Training *vs.* Separate Training. We train the model separately for each task using the same architecture and number of iterations as joint training. In Tab. 4(a), joint training consistently outperforms separate training across all tasks, indicating that our unified framework benefits from cross-task knowledge while task-specific concept prompts preserve task-specific discrimination.

Effectiveness of Concept Discriminative Subspace (CDS). We evaluate CDS from three aspects: concept discrimination, scene understanding, and concept-drift adaptation. In Tab. 4(b), CDS consistently outperforms the baseline UNet, which uses a standard architecture without specialized designs, demonstrating improved concept discrimination. To evaluate scene understanding, we remove the local discriminative information, which leads to a clear performance drop. This indicates

Table 4: Ablation studies on the eight tasks. All models adopt the Swin-B as the backbone.

Methods	Camouflaged		Salient		Shadow		Transparent		Thyroid Nodule		Polyp		COVID-19		Liver Tumor	
	MAE ↓	S_m ↑	MAE ↓	S_m ↑	MAE ↓	BER ↓	MAE ↓	BER ↓	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑
(a) Joint Training *vs.* Separate Training																
Separate Training	.030	.799	.076	.730	.029	.135	.072	.073	.820	.727	.833	.765	.969	.950	.874	.869
Joint Training + Shared Prompt	.046	.749	.085	.717	.037	.147	.080	.082	.801	.705	.796	.702	.950	.932	.823	.806
Joint Training	.028	.802	.070	.742	.026	.103	.069	.069	.824	.731	.848	.783	.970	.951	.876	.870
(b) Concept Discriminative Subspace																
UNet	.089	.528	.139	.511	.101	.198	.147	.152	.672	.571	.701	.636	.714	.702	.695	.684
w/o Local Discriminative	.043	.756	.080	.727	.036	.145	.078	.077	.815	.720	.802	.713	.961	.944	.732	.716
w/o Historical Discriminative	.030	.800	.072	.739	.028	.105	.077	.074	.822	.728	.843	.778	.968	.950	.874	.868
w/o Current Discriminative	.030	.797	.074	.735	.028	.114	.074	.075	.822	.729	.820	.749	.966	.946	.860	.860
CDH ⇒ Distance-weighted	.033	.791	.079	.726	.040	.155	.082	.083	.821	.727	.845	.779	.965	.948	.869	.861
Evolving ⇒ Static	.033	.790	.077	.730	.031	.132	.075	.076	.816	.723	.840	.771	.965	.946	.865	.862
CDS	.028	.802	.070	.742	.026	.103	.069	.069	.824	.731	.848	.783	.970	.951	.876	.870
(c) Hyperparameters																
Low-rank ($r = 4$)	.031	.797	.073	.736	.030	.108	.071	.072	.816	.724	.838	.772	.968	.949	.873	.866
Low-rank ($r = 8$)	.028	.802	.070	.742	.026	.103	.069	.069	.824	.731	.848	.783	.970	.951	.876	.870
Low-rank ($r = 10$)	.029	.802	.070	.741	.027	.106	.071	.070	.823	.730	.847	.782	.969	.951	.875	.869
Concept Definition Samples = 6	.031	.796	.073	.737	.027	.095	.073	.071	.823	.729	.845	.781	.968	.949	.874	.869
Concept Definition Samples = 12	.028	.802	.070	.742	.026	.103	.069	.069	.824	.731	.848	.783	.970	.951	.876	.870
Concept Definition Samples = 18	.029	.801	.071	.742	.026	.102	.069	.069	.823	.731	.848	.782	.971	.951	.877	.871
(d) Training Strategies																
Frame-based Training	.036	.763	.083	.721	.041	.159	.079	.081	.820	.729	.837	.769	.962	.946	.871	.862
Sequential Task Optimization	.031	.800	.074	.732	.030	.129	.072	.073	.821	.728	.840	.774	.965	.947	.872	.864
Temporal Clip Balancing	.028	.802	.070	.742	.026	.103	.069	.069	.824	.731	.848	.783	.970	.951	.876	.870

the importance of modeling within-scene foreground-background relations, as foreground and background are relative concepts determined by contextual relationships rather than solely by semantic categories, textures, or appearances. To evaluate concept-drift adaptation, we separately remove the historical and current-frame discriminative information. Results reflect that historical information provides prior discriminative knowledge, while current-frame information adapts the representation to the evolving concept. Removing either degrades performance, particularly when current-frame information is absent. We further replace the concept discriminative head (CDH) with a distance-weighted prediction. While this alternative remains effective in relatively homogeneous medical scenes, its performance drops substantially in natural scenes. Finally, the comparison between static and evolving CDS validates the effectiveness of the proposed CDS evolution.

Hyperparameters. We study two key hyperparameters in Tab. 4(c), *i.e.*, the CDS rank r and the number of concept definition samples. Performance peaks at $r = 8$, suggesting that a compact CDS captures sufficient discriminative information. Increasing the number of concept definition samples improves concept prototype estimation, with gains saturating beyond 12 samples. These results indicate limited sensitivity to both hyperparameters. Further comparisons of concept definition sample usage are provided in § G.

Training Strategies. To evaluate the effectiveness of TCB, as shown in Tab. 4(d), to assess the role of temporal context, we adopt a frame-based training strategy by setting the query clip length to 1. Furthermore, to examine the importance of joint optimization, we employ sequential task optimization, where only one task is optimized per iteration in a round-robin manner, instead of optimizing all tasks simultaneously. TCB consistently outperforms both alternatives, demonstrating the effectiveness of TCB for stable joint optimization.

5 Conclusion

We present ***EvoCDS***, a unified framework for CD concept segmentation across natural videos, medical videos, and 3D medical scenarios. By modeling concept-specific discrimination in an evolving low-dimensional subspace, ***EvoCDS*** shares compact structures across heterogeneous concepts while adapting to concept drift. Together with temporal clip balancing, it enables stable joint learning across imbalanced VCD tasks. Extensive experiments across eight tasks show consistent improvements over existing unified models and competitive performance against specialized methods. We also construct ***Video-CD*** as a benchmark for evaluating robustness to concept drift. More broadly, our results suggest that sequential CD understanding should move beyond fixed representations toward modeling how discriminative evidence evolves with visual context. We hope this work provides a foundation for open-world, multi-modal, and long-term adaptive concept understanding.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experiments, assist with the interpretation of experimental results, and support translation and language refinement. We have not used generative AI tools to generate synthetic datasets, formulate mathematical claims, provide or write mathematical proofs, or implement the proposed method, and the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to assist with manuscript drafting and editing, improve readability and organization, and provide feedback on the presentation of figures and experimental results. We have reviewed all AI-assisted work. All technical claims, experimental results, numerical values, and cited literature were independently checked and verified by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, et al. Many-shot in-context learning. In *NeurIPS*, volume 37, pp. 76930–76966, 2024.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Lawrence W Barsalou. Context-independent and context-dependent information in concepts. *Memory & cognition*, 10(1):82–93, 1982.

Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *MIA*, 84:102680, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, volume 33, pp. 1877–1901, 2020.

Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, pp. 25–36, 2004.

Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Universeg: Universal medical image segmentation. In *ICCV*, pp. 21438–21451, 2023.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *ICLR*, 2026.

Shijie Chang, Xiaoqi Zhao, Lihe Zhang, and Tiancheng Wang. Unified medical lesion segmentation via self-referring indicator. In *CVPR*, pp. 10414–10424, 2025.

Guilian Chen, Huisi Wu, and Jing Qin. Stddnet: Harnessing mamba for video polyp segmentation via spatial-aligned temporal modeling and discriminative dynamic representation learning. In *ICCV*, pp. 21364–21373, 2025.

Zhihao Chen, Liang Wan, Lei Zhu, Jia Shen, Huazhu Fu, Wennan Liu, and Jing Qin. Triple-cooperative video shadow detection. In *CVPR*, pp. 2715–2724, 2021.

Runmin Cong, Anpeng Wang, Bin Wan, Cong Zhang, Xiaofei Zhou, and Wei Zhang. Divide-and-conquer decoupled network for cross-domain few-shot segmentation. In *AAAI*, volume 40, pp. 3416–3424, 2026a.

Runmin Cong, Hang Xiong, Jinpeng Chen, Feng Li, Huihui Bai, and Sam Kwong. Beyond simple extraction: Unleashing the potential of encoder interaction in few-shot segmentation. *IEEE TMM*, 2026b.

Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *CVPR*, pp. 8554–8564, 2019.

Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *MICCAI*, pp. 263–273, 2020.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys*, 46(4):1–37, 2014.

Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE TPAMI*, 43(2):652–662, 2019.

Joseph L Gastwirth. The estimation of the lorenz curve and gini index. *The review of economics and statistics*, pp. 306–316, 1972.

Chao Hao, Zitong Yu, Xin Liu, Jun Xu, Huanjing Yue, and Jingyu Yang. A simple yet effective network based on vision transformer for camouflaged object and salient object detection. *IEEE TIP*, 2025.

Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *CVPR*, pp. 22046–22055, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Jiesi Hu, Yanwu Yang, Xutao Guo, Bo Peng, Hua Huang, and Ting Ma. Decor-net: a covid-19 lung infection segmentation network improved by emphasizing low-level features and decorrelating features. In *2023 IEEE 20th International Symposium on Biomedical Imaging*, pp. 1–5, 2023.

Qiang Hu, Zhenyu Yi, Ying Zhou, Fang Peng, Mei Liu, Qiang Li, and Zhiwei Wang. Sali: Short-term alignment and long-term interaction network for colonoscopy video polyp segmentation. In *MICCAI*, pp. 531–541, 2024.

Ge-Peng Ji, Yu-Cheng Chou, Deng-Ping Fan, Geng Chen, Huazhu Fu, Debesh Jha, and Ling Shao. Progressively normalized self-attention network for video polyp segmentation. In *MICCAI*, pp. 142–152, 2021.

Ge-Peng Ji, Guobao Xiao, Yu-Cheng Chou, Deng-Ping Fan, Kai Zhao, Geng Chen, and Luc Van Gool. Video polyp segmentation: A deep learning perspective. *MIR*, 19(6):531–549, 2022.

Hefeng Ji, Jing Xiao, Jiefan Lin, Jimin Liu, and Haoyong Yu. Intelligent tumor synthesis based on medical image knowledge for liver tumor segmentation. *ACM TOMM*, 21(9):1–23, 2025.

Ma Jun, Ge Cheng, Wang Yixin, An Xingle, Gao Jiantao, Yu Ziqi, Zhang Minqing, Liu Xin, Deng Xueyuan, Cao Shucheng, et al. Covid-19 ct lung and infection segmentation dataset. *(No Title)*, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Virginia Klema and Alan Laub. The singular value decomposition: Its computation and some applications. *TAC*, 25(2):164–176, 1980.

Thomas Lachmann and Cees Van Leeuwen. Individual pattern representations are context independent, but their collectiverepresentation is context dependent. *The Quarterly Journal of Experimental Psychology Section A*, 58(7):1265–1294, 2005.

Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology*, 49(4):764–766, 2013.

Jingjing Li, Yue Feng, Yuchen Guo, Jincai Huang, Wei Ji, Qi Bi, Yongri Piao, Miao Zhang, Xiaoqi Zhao, Qiang Chen, et al. Sam3-i: Segment anything with instructions. *arXiv preprint arXiv:2512.04585*, 2025a.

Zhong-Yu Li, Ruoyi Du, Juncheng Yan, Le Zhuo, Zhen Li, Peng Gao, Zhanyu Ma, and Ming-Ming Cheng. Visualcloze: A universal image generation framework via visual in-context learning. In *ICCV*, pp. 18969–18979, 2025b.

Shuang Liang, Zeqing Wang, Yuxian Li, Xihui Liu, and Han Wang. From pixels to concepts: Do segmentation models understand what they segment? *arXiv preprint arXiv:2605.09591*, 2026.

Junhao Lin, Qian Dai, Lei Zhu, Huazhu Fu, Qiong Wang, Weibin Li, Wenhao Rao, Xiaoyang Huang, and Liansheng Wang. Shifting more attention to breast lesion segmentation in ultrasound videos. In *MICCAI*, pp. 497–507, 2023.

Fang Liu, Yuhao Liu, Jiaying Lin, Ke Xu, and Rynson WH Lau. Multi-view dynamic reflection prior for video glass surface detection. In *AAAI*, volume 38, pp. 3594–3602, 2024.

Lihao Liu, Jean Prost, Lei Zhu, Nicolas Papadakis, Pietro Liò, Carola-Bibiane Schönlieb, and Angelica I Aviles-Rivero. Scotch and soda: A transformer video shadow detection framework. In *CVPR*, pp. 10449–10458, 2023a.

Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *CVPR*, pp. 19434–19445, 2023b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pp. 10012–10022, 2021.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, pp. 11999–12009, 2022a.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pp. 11976–11986, 2022b.

Jie Lu, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang. Learning under concept drift: A review. *IEEE TKDE*, 31(12):2346–2363, 2018.

Lei Lu, Xinyu Xiong, Lianjun Liu, and Hao Tang. Bg-uisnet: A novel boundary-guided network for robust and precise underwater image segmentation. *IEEE Sensors Journal*, 2025.

Yiwei Lu, Hao Huang, and Tao Yan. Mvgd-net: A novel motion-aware video glass surface detection method. In *AAAI*, volume 40, pp. 7591–7599, 2026.

Ziyang Luo, Nian Liu, Xuguang Yang, Dingwen Zhang, Deng-Ping Fan, Fahad Shahbaz Khan, and Junwei Han. Vscode-v2: Dynamic prompt learning for general visual salient and camouflaged object detection with two-stage optimization. *IEEE TPAMI*, 2026.

Charlotte Martial, David Stawarczyk, and Arnaud D'Argembeau. Neural correlates of context-independent and context-dependent self-knowledge. *Brain and Cognition*, 125:23–31, 2018.

Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *CVPR*, pp. 2150–2160, 2022.

Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoomnext: A unified collaborative pyramid network for camouflaged object detection. *IEEE TPAMI*, 46(12):9205–9220, 2024.

Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. Iclr: In-context learning of representations. In *ICLR*, volume 2025, pp. 53258–53284, 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, volume 32, 2019.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, pp. 724–732, 2016.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pp. 234–241, 2015.

Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. UNETR++: delving into efficient and accurate 3d medical image segmentation. *IEEE TMI*, 43(9):3377–3390, 2024.

Zhuo Su, Li Liu, Matthias Müller, Jiehua Zhang, Diana Wofk, Ming-Ming Cheng, and Matti Pietikäinen. Rapid salient object detection with difference convolutional neural networks. *IEEE TPAMI*, 47(10):9061–9077, 2025.

Yanguang Sun, Jiawei Lian, Jian Yang, and Lei Luo. Controllable-lpmoe: Adapting to challenging object segmentation via dynamic local priors from mixture-of-experts. In *ICCV*, pp. 22327–22337, 2025a.

Yanpeng Sun, Qiang Chen, Jian Wang, Jingdong Wang, and Zechao Li. Exploring effective factors for improving visual in-context learning. *IEEE TIP*, 2025b.

Yecong Wan, Mingwen Shao, Jun Shu, Yuanshuo Cheng, Deyu Meng, and Wangmeng Zuo. Separating anything from image in context. *PR*, pp. 113839, 2026.

Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *ICCV*, pp. 9197–9206, 2019.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvtv2: Improved baselines with pyramid vision transformer. *CVM*, 8(3):1–10, 2022.

Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *CVPR*, pp. 6830–6839, 2023a.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. SegGPT: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023b.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, volume 34, pp. 12077–12090, 2021.

Zhaohu Xing, Tian Ye, Yijun Yang, Du Cai, Baowen Gai, Xiao-Jian Wu, Feng Gao, and Lei Zhu. Segmamba-v2: Long-range sequential modeling mamba for general 3d medical image segmentation. *IEEE TMI*, 2025.

Jiawei Xu, Qiangqiang Zhou, Jiacong Yu, Chen Liao, and Dandan Zhu. Semantic-orthogonal multimodal attention network for rgb-d salient object detection. *The Visual Computer*, pp. 1–13, 2025.

Jiawei Xu, Qiangqiang Zhou, Zhouping Li, Yanjiao Shi, Yugen Yi, and Jiacong Yu. Hvpnet: A bio-inspired network for general salient and camouflaged object detection. *Neural Networks*, pp. 109340, 2026a.

Jiawei Xu, Qiangqiang Zhou, Dandan Zhu, Yong Chen, Yugen Yi, and Xiaoqi Zhao. Tp-seg: Task-prototype framework for unified medical lesion segmentation. In *CVPR*, pp. 5452–5462, 2026b.

Yijun Yang, Zhaohu Xing, and Lei Zhu. Vivim: a video vision mamba for medical video object segmentation. *arXiv preprint arXiv:2401.14168*, 2024.

Siyuan Yao, Hao Sun, Ruiqi Yu, Xiwei Jiang, Wenqi Ren, and Xiaochun Cao. Camotion: A high-quality benchmark for camouflaged moving object detection in the wild. *arXiv preprint arXiv:2604.08287*, 2026.

Jin Zhang, Zhe Cao, Biwen Yang, and Ruiheng Zhang. Annotation-efficient coreset selection for context-dependent segmentation. In *CVPR*, pp. 20410–20420, 2026.

Xiaoqi Zhao, Shijie Chang, Youwei Pang, Jiaxing Yang, Lihe Zhang, and Huchuan Lu. Adaptive multi-source predictor for zero-shot video object segmentation. *IJCV*, 132(8):3232–3250, 2024a.

Xiaoqi Zhao, Youwei Pang, Shijie Chang, Yuan Zhao, Lihe Zhang, Chenyang Yu, Hanqi Liu, Jiaming Zuo, Jinsong Ouyang, Weisi Lin, et al. Inspiring the next generation of segment anything models: Comprehensively evaluate sam and sam 2 with diverse prompts towards context-dependent concepts under different scenes. *arXiv preprint arXiv:2412.01240*, 2024b.

Xiaoqi Zhao, Youwei Pang, Wei Ji, Baicheng Sheng, Jiaming Zuo, Lihe Zhang, and Huchuan Lu. Spider: A unified framework for context-dependent concept segmentation. In *ICML*, 2024c.

Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Towards diverse binary segmentation via a simple yet general gated network. *IJCV*, 132(10):4157–4234, 2024d.

Yuan Zhao, Youwei Pang, Jiaming Zuo, Wei Ji, Kailai Zhou, Bin Fan, Yunkang Cao, Lihe Zhang, Xiaofeng Liu, Huchuan Lu, et al. ConceptSeg-R1: Segment any concept via meta-reinforcement learning. *arXiv preprint arXiv:2605.20385*, 2026.

Qiangqiang Zhou, Jiawei Xu, Yong Chen, Dandan Zhu, Yugen Yi, and Xiaoqi Zhao. Differseg: Towards diverse multimodal binary segmentation via differential perception and frequency guidance. *IEEE TCSVT*, 2026a.

Yuli Zhou, Yawei Li, Yuqian Fu, Luca Benini, Ender Konukoglu, and Guolei Sun. Camsam2: Segment anything accurately in camouflaged videos. In *NeurIPS*, volume 38, pp. 55489–55517, 2026b.

Indrė Žliobaitė, Mykola Pechenizkiy, and Joao Gama. An overview of concept drift applications. *Big data analysis: new algorithms for a new society*, pp. 91–114, 2015.

APPENDIX

A Definition of Different Video Context-Dependent Concept Segmentation Tasks **16**

A.1 Video Salient Object Detection 16

A.2 Video Camouflaged Object Detection 16

A.3 Video Shadow Detection 16

A.4 Video Transparent Object Segmentation 17

A.5 Video Polyp Segmentation 17

A.6 Video Thyroid Nodule Segmentation 18

A.7 3D Lung Infection Segmentation 18

A.8 3D Liver Tumor Segmentation 18

B Concept Drift: Definition and Interpretation **18**

C Empirical Analysis of Latent Concept Representations **20**

C.1 Channel-wise Discriminative Information Concentration 21

C.2 Low-rank Structure Consistency of Latent Concept Representations 22

C.3 Concept Discriminative Head 25

D Qualitative Visualization **25**

E Analysis of Concept Discriminative Subspace **29**

F Discussion **30**

F.1 Does Concept Discriminative Subspace Progressively Build Discriminative? 30

F.2 Does Concept Discriminative Subspace Evolution Effectively Mitigate Concept Drift? 31

G Comparative Analysis of Different Methods on Number of Prompts **31**

H The Reconstruction Error **32**

I *Video-CD* **34**

I.1 Stage 1: Independent Drift Metric Bank 34

I.2 Stage 2: MLLM-Assisted Pre-annotation 35

I.3 Stage 3: Human Arbitration 35

I.4 Video-level Concept Drift Annotation. 35

J Limitations and Future Work **35**

A Definition of Different Video Context-Dependent Concept Segmentation Tasks

A.1 Video Salient Object Detection



Figure 7: Illustration of the VSOD task (yellow regions indicate salient objects).

Video salient object detection (VSOD) (Zhao et al., 2024d; Xu et al., 2025) aims to automatically locate and segment visually prominent objects from consecutive video frames. These objects typically exhibit strong spatial contrast and maintain temporal motion consistency across frames (Fig. 7). Unlike image-based salient object detection, VSOD requires not only spatial saliency modeling but also temporal consistency modeling to enhance stability and robustness of the detected objects.

A.2 Video Camouflaged Object Detection

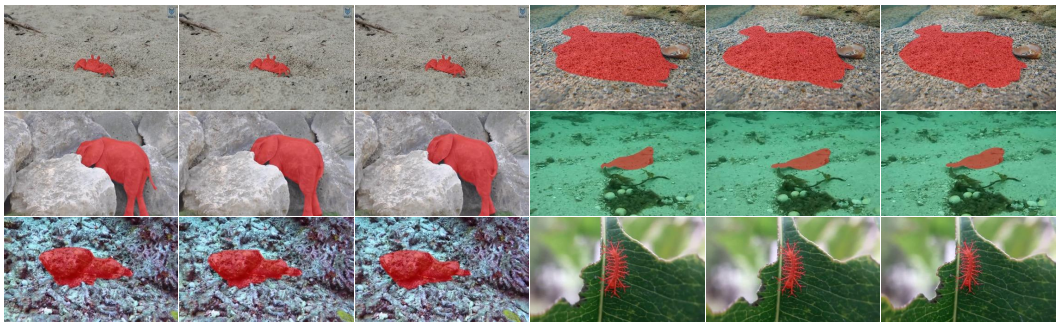

Figure 8: Illustration of the VCOD task (red regions indicate camouflaged objects).

Video camouflaged object detection (VCOD) (Pang et al., 2022; Sun et al., 2025a; Xu et al., 2026a) aims to identify and segment objects in video sequences that are highly similar to the surrounding background and exhibit strong concealment. Such objects often share similar color, texture, or shape with the background, making them difficult to distinguish using conventional methods. The key challenge of VCOD lies in jointly leveraging inter-frame motion cues and fine-grained appearance discrepancies to reveal truly camouflaged objects (Fig. 8).

A.3 Video Shadow Detection

Video shadow detection (VSD) (Chen et al., 2021; Liu et al., 2023a) aims to detect and segment shadow regions caused by illumination occlusion in video sequences (Fig. 9). Shadows contain rich geometric and lighting information but may interfere with object recognition and scene understanding. VSD requires not only accurate frame-wise shadow identification but also temporal consistency modeling to handle illumination variations and dynamic scene changes.

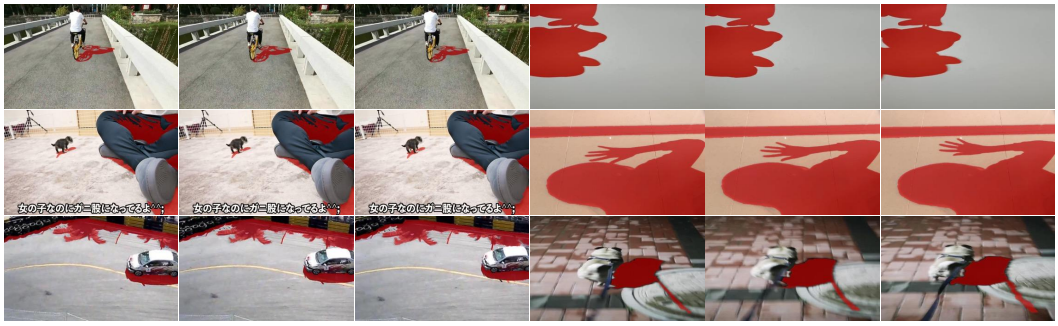

Figure 9: Illustration of the VSD task (red regions indicate shadow).



Figure 10: Illustration of the VTOS task (yellow regions indicate transparent objects).

A.4 Video Transparent Object Segmentation

Video transparent object segmentation (VTOS) (Liu et al., 2024; Lu et al., 2026) aims to identify objects with transparency, reflection, and refraction properties, such as glass or plastic materials, in video sequences (Fig. 10). Due to their strong dependency on background appearance, transparent objects suffer from severely degraded visual features, resulting in ambiguous boundaries and challenging segmentation. VTOS relies on multi-frame information and motion cues to mitigate visual uncertainty.

A.5 Video Polyp Segmentation

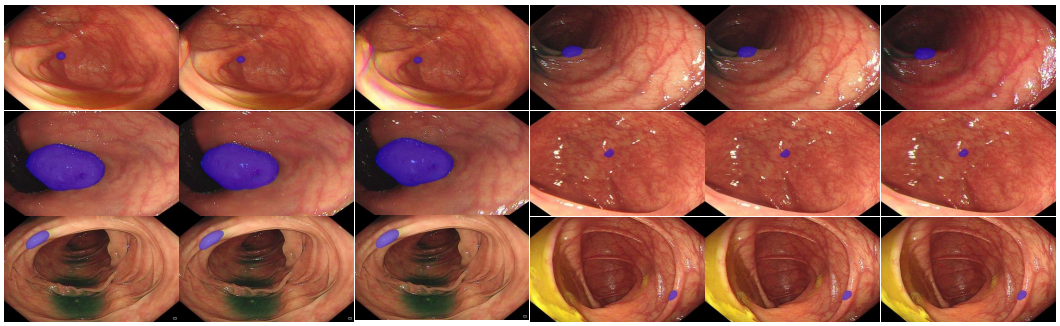

Figure 11: Illustration of the VPS task (blue regions indicate polyp lesion).

Video polyp segmentation (VPS) (Bernal et al., 2015; Hu et al., 2024) aims to automatically detect and segment polyp regions from endoscopic video sequences (Fig. 11). Polyps exhibit large variations in scale, shape, and color, and are often occluded by mucosal textures or specular reflections. VPS requires integration of spatial structural information and temporal continuity to improve robustness against small and occluded lesions.

A.6 Video Thyroid Nodule Segmentation

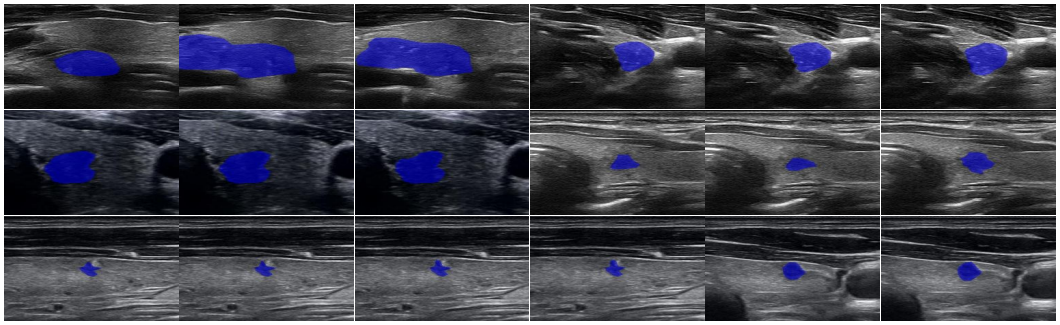

Figure 12: Illustration of the VTNS task (blue regions indicate thyroid nodule).

Video thyroid nodule segmentation (VTNS) (Yang et al., 2024; Lin et al., 2023) aims to identify and segment thyroid nodules from ultrasound video sequences (Fig. 12). Ultrasound images are often affected by speckle noise, low contrast, and complex tissue structures, making nodule boundaries ambiguous. VTNS leverages temporal redundancy in video sequences to enhance stability and improve boundary localization accuracy.

A.7 3D Lung Infection Segmentation

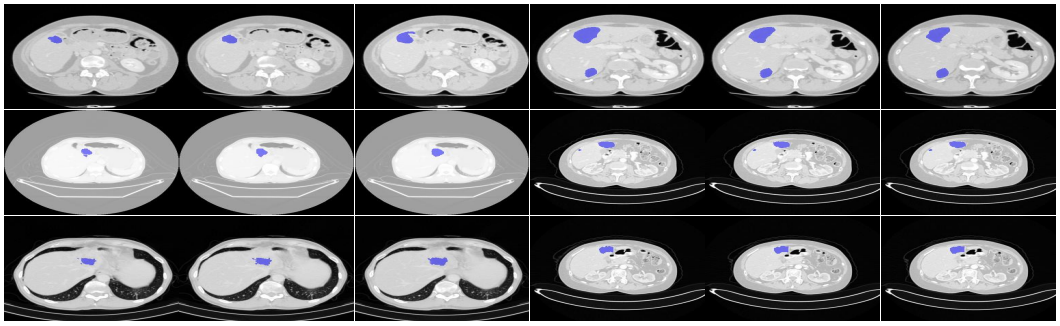

Figure 13: Illustration of the 3DLIS task (blue regions indicate COVID-19 lesion).

3D COVID-19 lung infection segmentation (3DLIS) (Jun et al., 2020; Shaker et al., 2024) aims to extract infected regions from volumetric CT data, including ground-glass opacities and consolidation areas (Fig. 13). This task requires modeling both 3D spatial structural information and lesion distribution patterns to address complex anatomical structures and multi-scale infection appearances.

A.8 3D Liver Tumor Segmentation

3D liver tumor segmentation (3DLTS) (Bilic et al., 2023; Ji et al., 2025) aims to accurately delineate tumor regions from volumetric medical images such as CT or MRI scans (Fig. 14). Liver tumors typically exhibit irregular shapes, fuzzy boundaries, and strong heterogeneity, and often share similar intensity distributions with surrounding liver tissues. This task relies on 3D contextual information to achieve precise localization and segmentation of multi-scale lesions.

B Concept Drift: Definition and Interpretation

Concept drift (Fig. 15) in video context-dependent (VCD) concept segmentation refers to the phenomenon where the semantic identity of a target remains temporally consistent, while the discriminative visual and contextual evidence associated with it continuously evolves due to dynamic scene changes, including occlusion, deformation, disappearance, and scene transitions. Accordingly, concept drift is formulated as a sequence-level phenomenon rather than an isolated frame-level event,

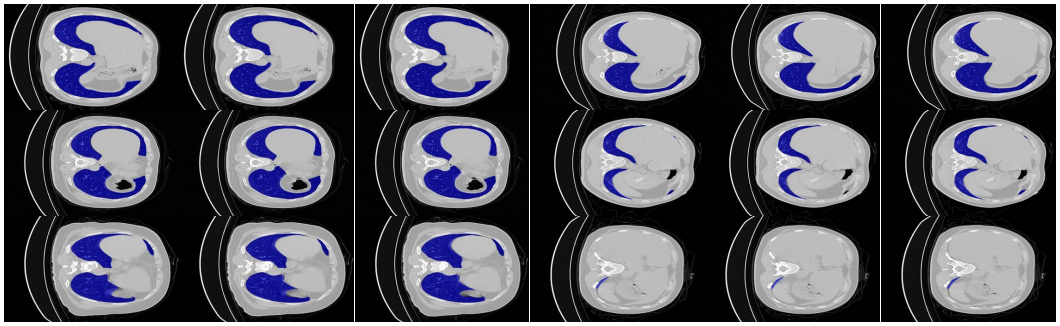

Figure 14: Illustration of the 3DLTS task (blue regions indicate liver tumor).

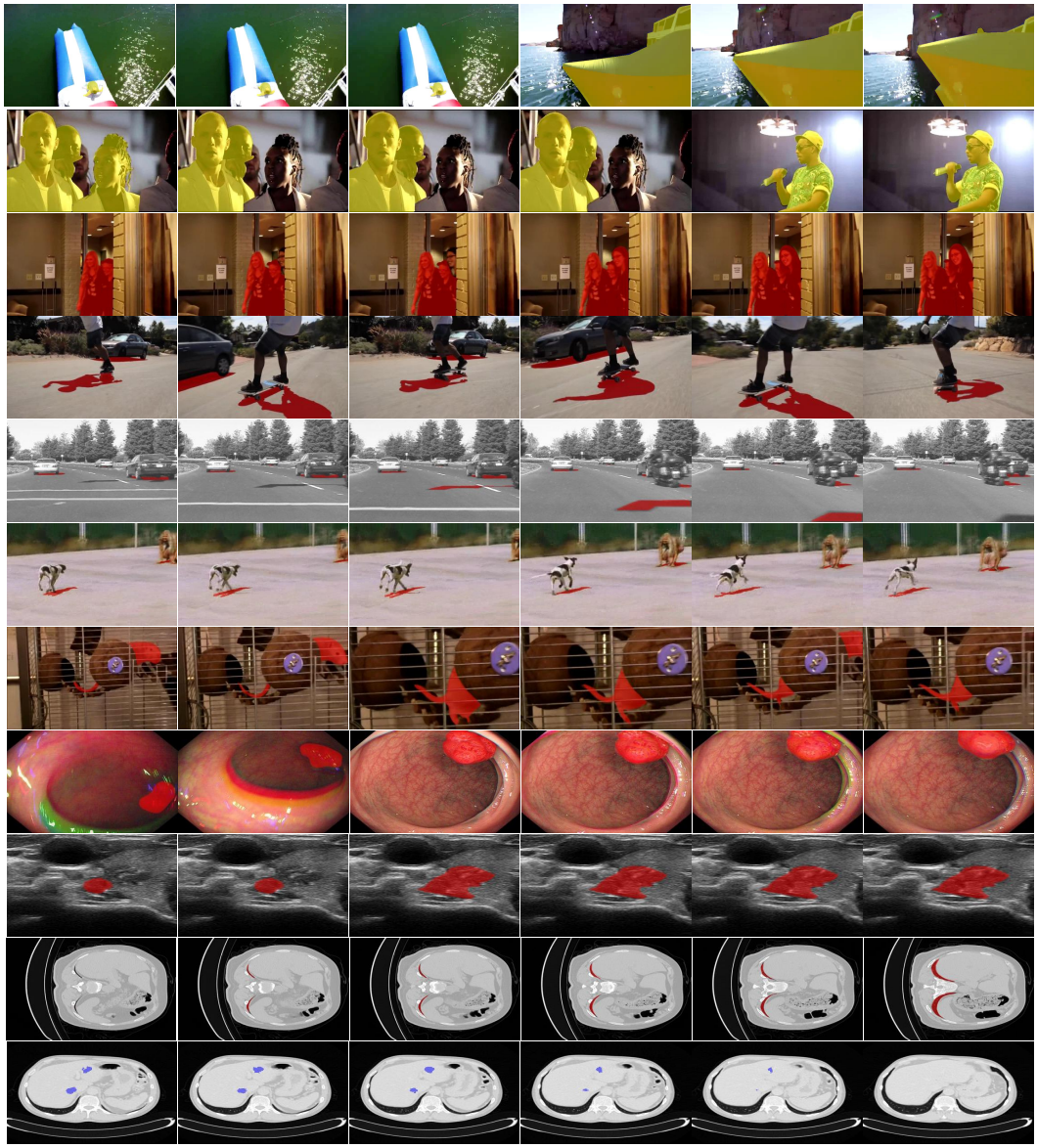

Figure 15: Illustrative examples of concept drift. The highlighted areas denote the segmentation targets or objects of interest.

as it reflects the temporal evolution of concept-discriminative evidence across consecutive observations. (**Note that our notion of concept drift differs from the conventional distributional concept drift studied (Lu et al., 2018; Gama et al., 2014; Žliobaitė et al., 2015) in online learning. Here, it specifically refers to the temporal evolution of task-relevant contextual and discriminative evidence for a semantically consistent target.**)

Different from ordinary temporal variation, which describes changes in visual observations while preserving a relatively stable recognition mechanism, concept drift characterizes changes in the underlying discriminative relationship between a concept and its supporting evidence. This distinction is particularly important in context-dependent concept segmentation, where the target identity is not determined solely by intrinsic appearance features, but jointly by its contextual relationship with surrounding environments. Therefore, temporal changes such as illumination variation or motion may introduce observation-level fluctuations without altering the concept definition, whereas context evolution, scene transition, target disappearance, or severe occlusion can fundamentally modify the evidence required to distinguish foreground and background concepts. In contrast to context-independent tasks, such as conventional video object segmentation where temporal consistency is often the primary objective, VCD concept segmentation requires modeling how the semantic interpretation of a target evolves with changing contexts.

From a unified perspective, this evolution is not merely caused by isolated appearance perturbations, but instead reflects a systematic shift in the underlying data-generating context over time. Specifically, although the target maintains a stable semantic category, the observable cues that support its recognition are progressively altered by both object-level state changes and scene-level structural dynamics. As a result, the mapping between semantic identity and its visual manifestation becomes inherently time-dependent.

This implies that the model does not face a static classification problem, but rather a continuously evolving representation problem, where the discriminative boundaries between concepts are dynamically reshaped across time steps. Consequently, the separability of semantic concepts in the feature space is no longer fixed, but fluctuates according to the degree of contextual consistency and visual stability. Under this formulation, concept drift can be interpreted as the core underlying mechanism that unifies multiple apparent challenges in VCD concept segmentation, such as occlusion-induced information loss, deformation-induced feature distortion, and scene transition-induced distribution shifts. All of these phenomena can be viewed as different manifestations of a single process: the temporal evolution of discriminative evidence under dynamic environments.

C Empirical Analysis of Latent Concept Representations

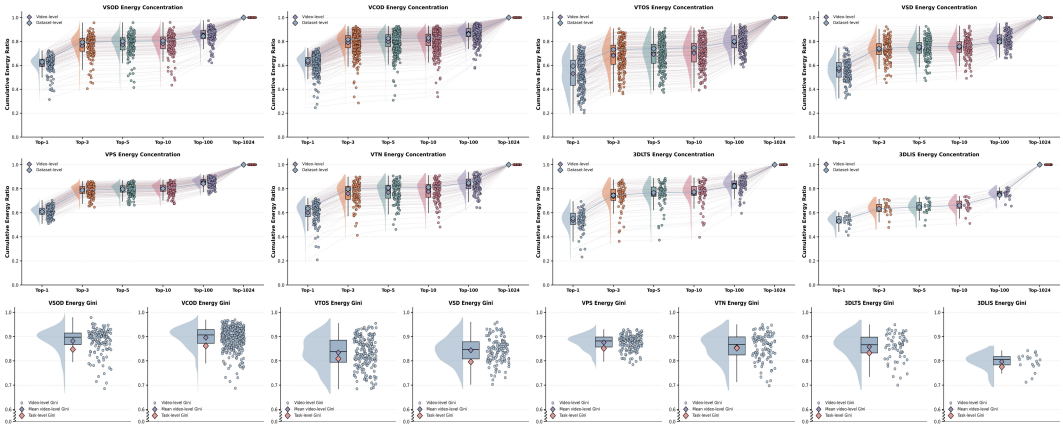

Figure 16: Analysis of energy distribution across feature dimensions.



Figure 17: Temporal consistency of latent concept representations measured by cosine similarity.

C.1 Channel-wise Discriminative Information Concentration

We hypothesize that latent conceptual representations exhibit compact low-rank structures embedded in high-dimensional feature spaces and maintain strong temporal consistency. To verify this hypothesis, we conduct experiments on eight diverse vision tasks, including video salient object detection (VSOD), video camouflaged object detection (VCOD), video shadow detection (VSD), video transparent object segmentation (VTOS), video polyp segmentation (VPS), video thyroid nodule segmentation (VTNS), 3D Lung infection segmentation (3DLIS), and 3D liver tumor segmentation (3DLTS).

We employ a frozen backbone network $f(\cdot)$ (Liu et al., 2021) (Fig. 16) to extract feature representations, and explicitly decouple foreground and background features using task-level ground truth (GT) annotations. Let the input feature be denoted as: $F \in \mathbb{R}^{C\times H\times W}$, where $C = 1024$ denotes the channel dimension. Given a binary GT mask $M \in \{0,1\}^{H\times W}$, the foreground and background features are defined as:

$$F_f = F \odot M, \quad F_b = F \odot (1 - M), \tag{11}$$

We then apply global average pooling over the spatial dimension to obtain channel-wise representations of foreground and background:

$$v_f = \text{GAP}(F_f) \in \mathbb{R}^C, \quad v_b = \text{GAP}(F_b) \in \mathbb{R}^C. \tag{12}$$

The foreground-background difference vector is defined as:

$$\Delta v = v_f - v_b. \tag{13}$$

To analyze the contribution of individual feature dimensions, we define the channel-wise discriminative energy as:

$$E_i = |\Delta v_i|^2, \quad i = 1, 2, \ldots, 1024, \tag{14}$$

and the total energy as:

$$E_{\text{total}} = \sum_{i=1}^{1024} E_i. \tag{15}$$

Empirically, we observe a highly skewed long-tailed distribution of discriminative energy across feature dimensions:

$$\sum_{i\in\mathcal{K}} E_i \gg \sum_{i\notin\mathcal{K}} E_i, \tag{16}$$

where $\mathcal{K}$ denotes a small subset of critical feature channels. Specifically, only a limited proportion of feature dimensions contribute the majority of foreground-background discriminative information, whereas the remaining dimensions contribute relatively weak signals. This observation indicates that discriminative information is unevenly distributed across the high-dimensional feature space.

To further quantify the concentration level of channel-wise discriminative energy, we introduce the Gini index (Gastwirth, 1972). Let $\{E_{(i)}\}_{i=1}^{C}$ denote the energies sorted in ascending order. The Gini index is defined as:

$$G = \frac{2\sum_{i=1}^{C} iE_{(i)}}{C\sum_{i=1}^{C} E_{(i)}} - \frac{C+1}{C}. \tag{17}$$

A higher Gini index indicates stronger inequality among feature dimensions, suggesting that a small subset of channels dominates the discriminative energy while most channels contribute marginally. These results demonstrate that foreground-background discriminative information is concentrated within a limited number of feature dimensions. However, channel-wise energy concentration alone does not directly characterize the intrinsic dimensionality of the representation space. Therefore, we further perform an SVD-based analysis to directly evaluate the low-rank structure of latent concept representations. Additionally, in Fig. 17 and Fig. 18, we show that latent concept representations exhibit strong temporal consistency, with variations mainly occurring when concept drift takes place.

C.2 Low-rank Structure Consistency of Latent Concept Representations

To directly quantify the intrinsic dimensionality of latent concept representations, we perform singular value decomposition (SVD) (Klema & Laub, 1980) on the concept representation matrix. Unlike the channel-wise energy analysis above, which measures the contribution of individual feature dimensions, SVD characterizes the intrinsic structure of the concept space by identifying dominant latent directions.

For each task, the foreground-background difference vectors from all sampled frames/volumes are collected to form the concept matrix:

$$D = [\Delta v_1, \Delta v_2, ..., \Delta v_N]^T \in \mathbb{R}^{N \times C}, \tag{18}$$

where N denotes the number of sampled concept representations and $C = 1024$ represents the original feature dimension.

The singular value decomposition is formulated as:

$$D = U \Sigma V^T, \tag{19}$$

where $\Sigma = \{\sigma_1, \sigma_2, ..., \sigma_r\}$ contains the singular values corresponding to different latent directions. The normalized energy contribution of each singular direction is defined as:

$$P_i = \frac{\sigma_i^2}{\sum_{j=1}^{r} \sigma_j^2}. \tag{20}$$

The cumulative explained energy of the first k singular directions is defined as:

$$R_k = \frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{j=1}^{r} \sigma_j^2}. \tag{21}$$

As shown in Fig. 19, the intrinsic ranks required to preserve different levels of discriminative energy are substantially smaller than the original feature dimension. Across all eight tasks, only 2–6 singular directions are required to preserve 90% of the discriminative energy, while 4–8 singular directions are sufficient to retain more than 95% of the information. Even under the stricter 99% energy criterion, the required ranks remain below 10 for all tasks. These results demonstrate that latent concept representations occupy compact low-dimensional subspaces rather than uniformly distributed high-dimensional spaces.

The singular-value energy distributions in Fig. 20 further reveal a strong spectral concentration phenomenon. The first several singular directions dominate the total representation energy, whereas subsequent directions contribute only marginal information. This rapid spectral decay indicates that latent concepts are primarily governed by a small number of dominant semantic factors.

Furthermore, the cumulative explained energy curves in Fig. 21 quantitatively verify the compactness of the latent concept space. The first eight singular directions preserve 95.70%—99.61% of the total discriminative energy across all evaluated tasks. Together with the channel-wise energy analysis, these results provide direct evidence that latent concept representations exhibit low intrinsic dimensionality and can be accurately approximated using low-rank subspaces.

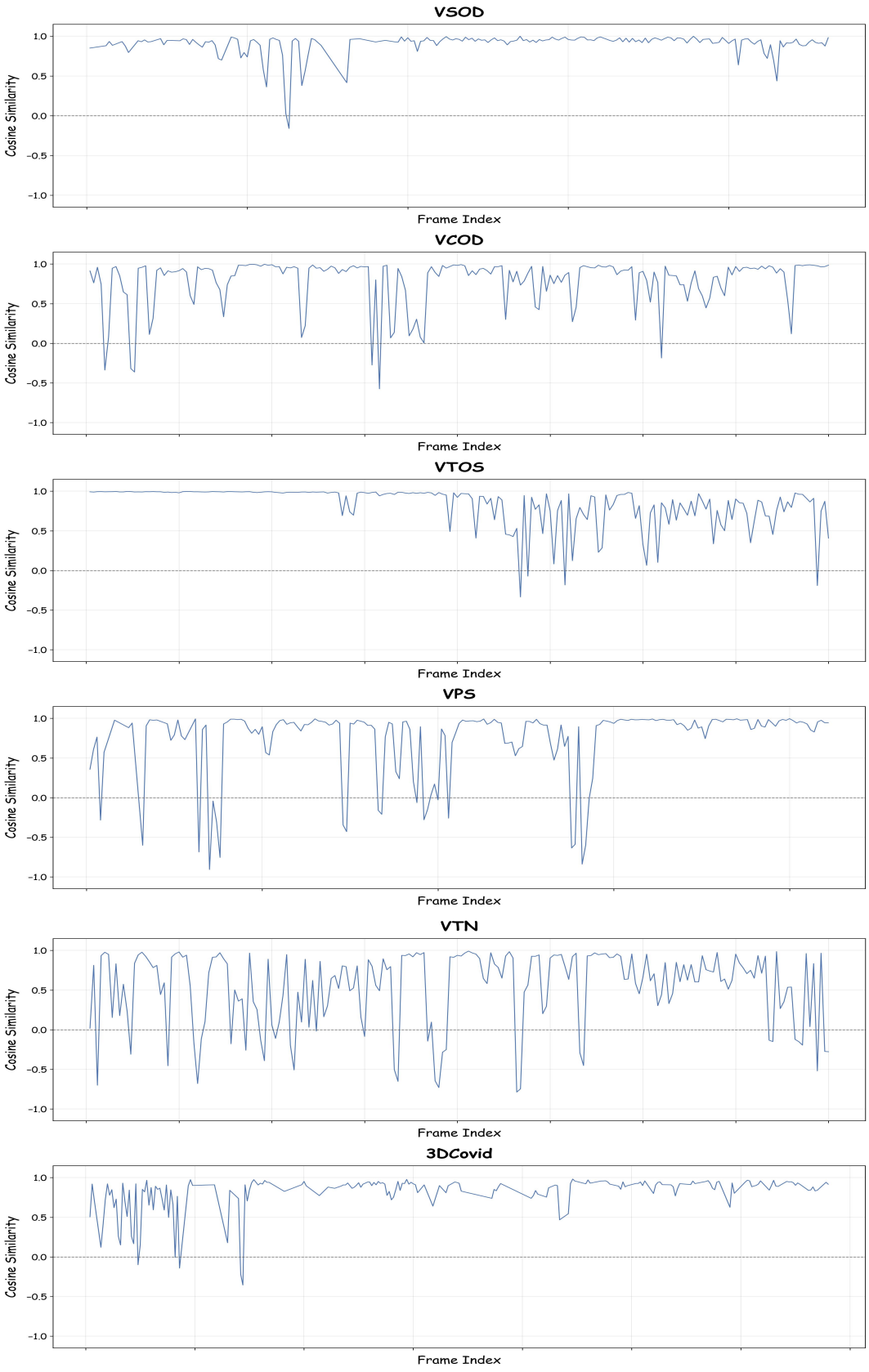

Figure 18: Variation of discriminative direction cosine similarity over time.

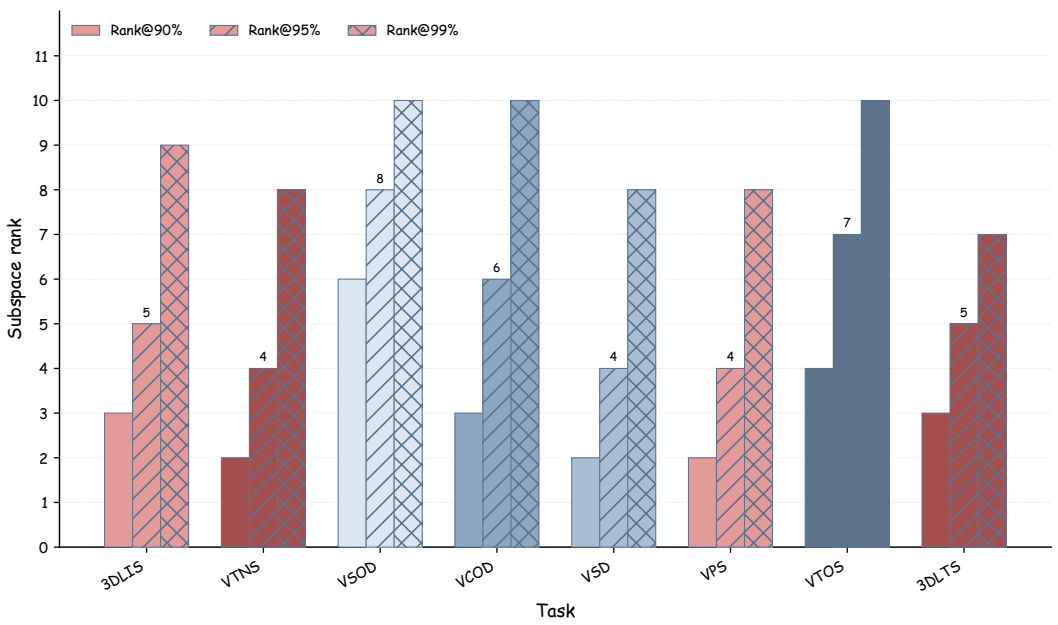

Figure 19: Intrinsic dimensionality of latent concept representations across different tasks. The required subspace ranks for preserving 90%, 95%, and 99% discriminative energy are reported. The results demonstrate that latent concepts can be accurately represented using only a small number of latent directions despite the original 1024-dimensional feature space.

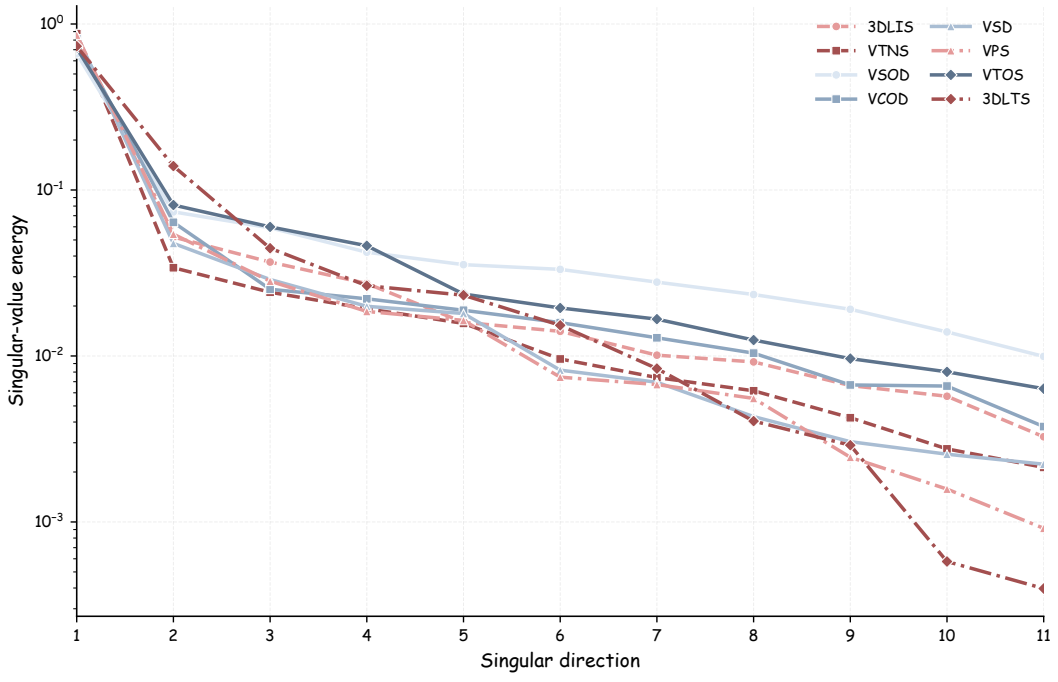

Figure 20: Singular-value energy distribution of latent concept representations. The normalized energy contribution of each singular direction is shown for different tasks. The rapidly decreasing energy curves indicate that only a few dominant directions capture most discriminative information.

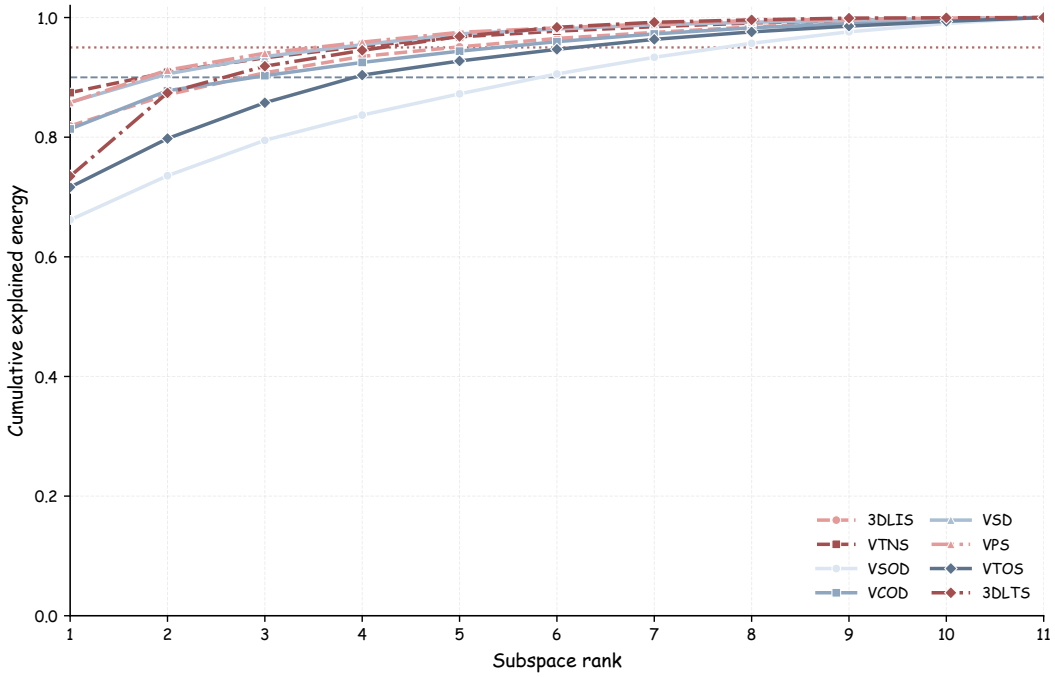

Figure 21: Cumulative explained discriminative energy with respect to subspace rank. The curves show the proportion of foreground-background discriminative information preserved by the first k singular directions.

Table 5: Performance comparison of distance-weighted and concept discriminative head (CDH) on different CD concepts.

Methods	Camouflaged		Salient		Shadow		Transparent		Thyroid Nodule		Polyp		COVID-19		Liver Tumor	
	MAE ↓	S_m ↑	MAE ↓	S_m ↑	MAE ↓	BER ↓	MAE ↓	BER ↓	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑
Distance-weighted	.033	.791	.079	.726	.040	.155	.082	.083	.821	.727	.845	.779	.965	.948	.869	.861
CDH	.028	.802	.070	.742	.026	.103	.069	.069	.824	.731	.848	.783	.970	.951	.876	.870

C.3 Concept Discriminative Head

Taken together, the above analyses reveal that latent concept representations exhibit two important properties: compact intrinsic dimensionality and temporal stability. The former indicates that discriminative concepts can be effectively represented within a small number of latent directions, while the latter demonstrates that these representations maintain consistent semantic structures over time. Furthermore, as shown in Tab. 5, explicitly incorporating such structured concept representations through the proposed concept discriminative head consistently improves performance across diverse natural and medical scenarios. These findings collectively validate the effectiveness of modeling latent concepts as compact and stable semantic representations.

D Qualitative Visualization

We conduct a comprehensive qualitative evaluation from three complementary perspectives: (1) comparison with unified modeling approaches (Fig. 22); (2) comparison with several task-specific methods (Fig. 23); and (3) qualitative evaluation of ***EvoCDS*** across different segmentation tasks and scenarios (Fig. 24). The visual comparisons show that ***EvoCDS*** produces more accurate and stable segmentation masks than the compared methods across a variety of challenging cases. First, the proposed method better preserves fine-grained target structures and delineates object boundaries more precisely. It can identify small regions, thin structures, and irregular contours while reducing boundary fragmentation, excessive smoothing, and the loss of local details commonly observed in

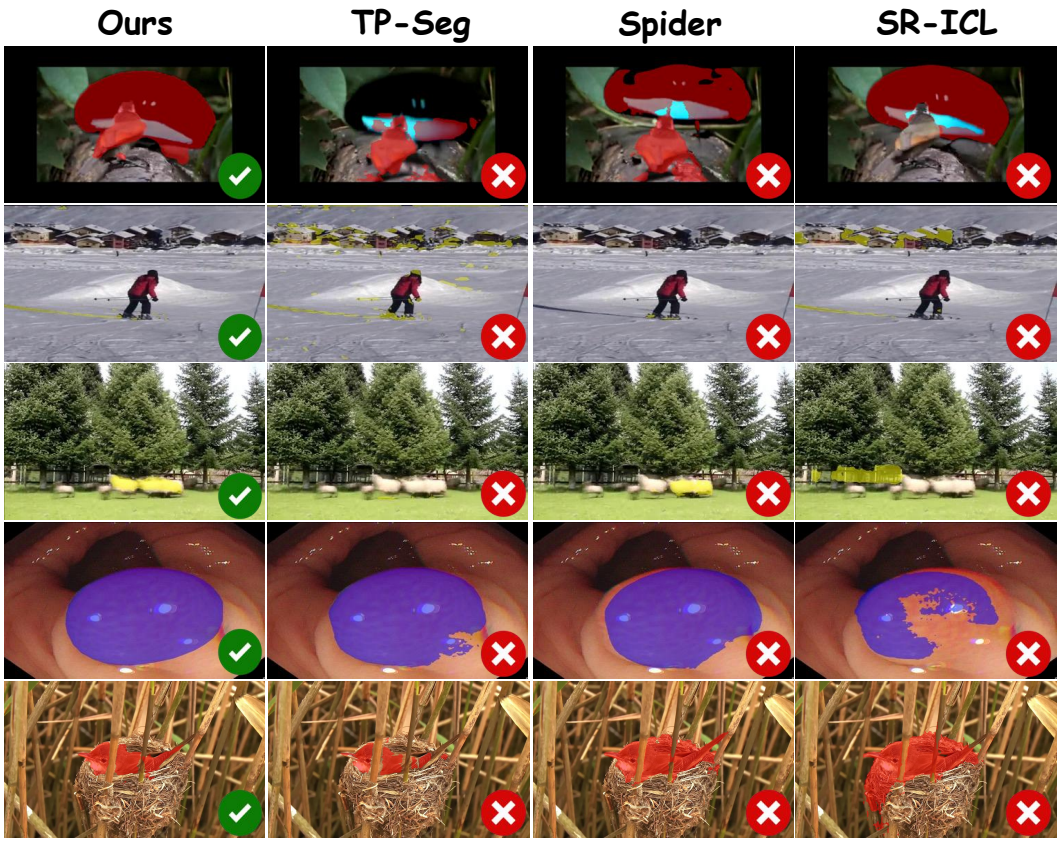

Figure 22: Qualitative comparison on diverse unified methods.

competing methods. Consequently, the predicted masks exhibit greater spatial agreement with the corresponding ground-truth annotations.

Second, ***EvoCDS*** demonstrates a stronger ability to maintain the structural completeness of target regions. In cases involving complex backgrounds, low foreground–background contrast, partial occlusion, or multiple adjacent objects, the proposed method more effectively separates foreground regions from surrounding interference. It also reduces false-positive responses in background areas and false-negative predictions within target regions, yielding more complete and coherent segmentation results.

Furthermore, under complex scenes and multi-task settings, ***EvoCDS*** shows favorable adaptability and robustness. Compared with task-specific methods optimized for individual tasks, the proposed framework maintains consistently competitive segmentation quality across different task categories without requiring separate model designs. Compared with unified modeling approaches, it more effectively accommodates diverse target appearances, spatial structures, and task-specific segmentation requirements. In particular, the predicted masks contain fewer isolated false-positive regions, missing components, boundary deviations, and topological inconsistencies.

Overall, the qualitative results demonstrate the effectiveness of ***EvoCDS*** in boundary delineation, target-region preservation, background suppression, and cross-task generalization. Across different datasets and segmentation scenarios, the proposed method consistently produces masks that more closely match the ground-truth annotations, particularly for small targets, irregular boundaries, and structurally complex regions. These observations provide qualitative evidence that the proposed design and overall framework improve segmentation accuracy, prediction consistency, and robustness across diverse tasks.

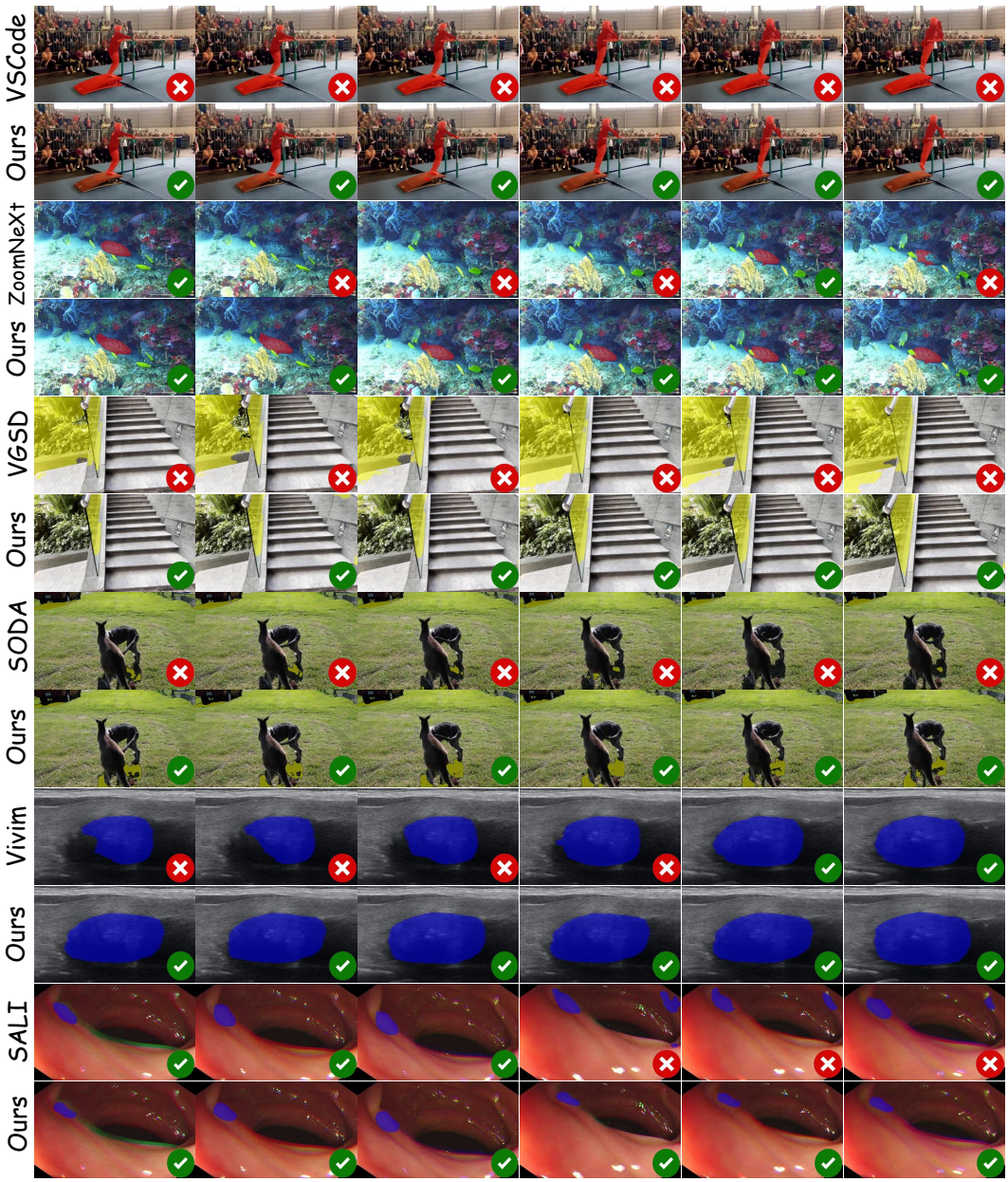

Figure 23: Qualitative comparison on diverse specialized methods.



Figure 24: Qualitative comparison on diverse concept tasks.

E Analysis of Concept Discriminative Subspace

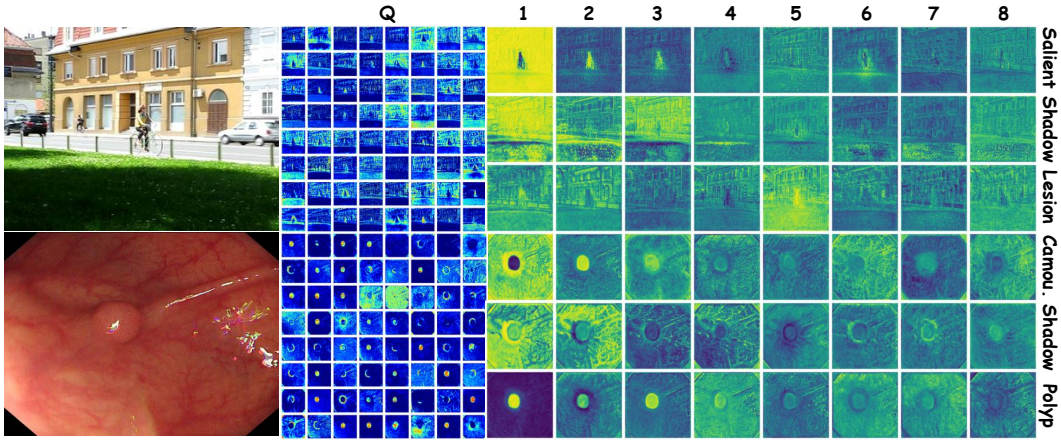

Figure 25: Visualization of the 8-dimensional concept discriminative subspace.

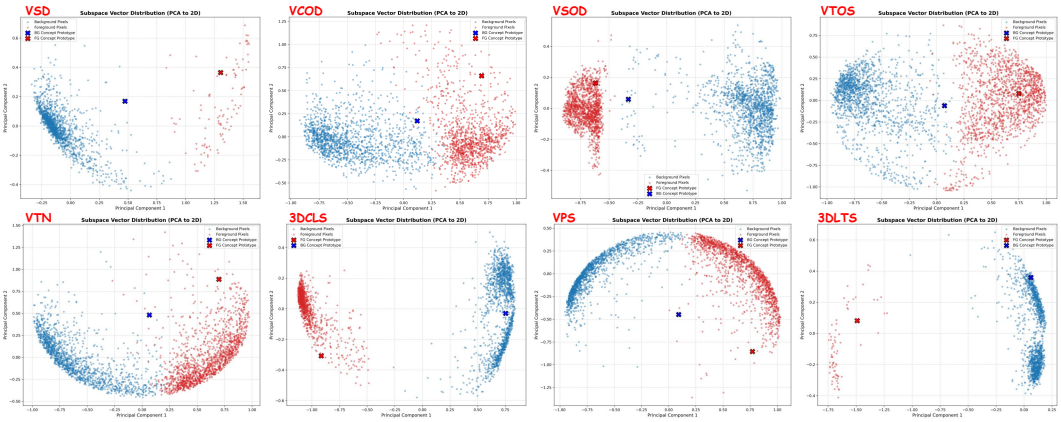

Figure 26: Clustering visualization of foreground and background features in task-specific concept discriminative subspace.

To validate the interpretability and discriminative capability of the proposed concept discriminative subspace across different VCD tasks, we perform concept discriminative subspace projection on intermediate feature representations and visualize the responses across different dimensions.

Specifically, given the feature representation of an input frame $f \in \mathbb{R}^{64\times H\times W}$, we project it into task-specific concept discriminative subspaces to obtain subspace responses $z \in \mathbb{R}^{8\times H\times W}$. Each channel of the resulting response is then normalized using min-max normalization and visualized, enabling analysis of the variations in model attention regions across different tasks.

As shown in Fig. 25, in the video salient object detection task, the concept discriminative subspace exhibits strong responses in the primary object regions, indicating that the model effectively focuses on globally salient areas. In the video shadow detection task, the concept discriminative subspace tends to respond more strongly to low-illumination regions or boundary transition areas, reflecting the model's sensitivity to illumination variations and structural boundaries.

For the video polyp segmentation task, the corresponding concept discriminative subspace responses show a high degree of similarity to those observed in the camouflaged object detection task, suggesting that these two implicit visual tasks share certain representational structures in the feature space. This observation further indicates that the model may rely on similar semantic discrimination mechanisms when handling low-contrast objects.

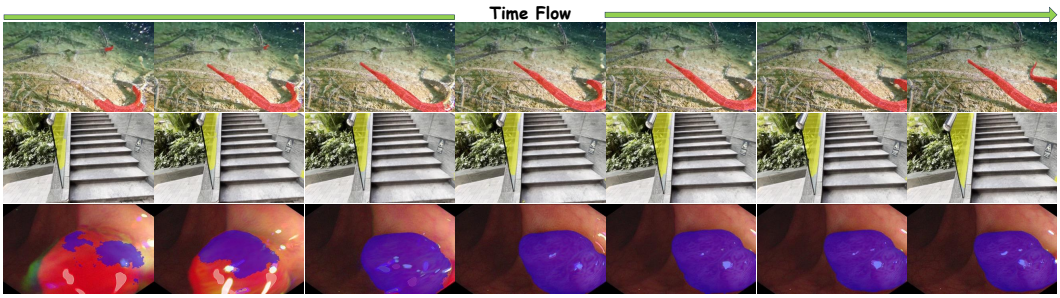

Figure 27: Evolution of concept discriminability over time within a single stable scene.

Overall, the concept discriminative subspace exhibits both consistency and diversity across different vision tasks. On the one hand, some concept discriminative subspaces demonstrate task-invariant semantic response patterns shared across tasks; on the other hand, there also exist task-specific subspaces that capture unique structural characteristics of different visual tasks.

To examine the foreground–background discriminability of the learned task-specific concept subspace, we visualize the distributions of query features and concept prototypes in the concept discriminative subspace. Specifically, we stack the reference-induced concept directions as columns to construct the concept matrix:

$$D = [d_1, \ldots, d_S] \in \mathbb{R}^{C \times S}. \tag{22}$$

We then perform singular value decomposition (SVD), $D = U\Sigma V^\top$, and retain the r dominant left singular vectors to form the concept discriminative subspace:

$$U_r = [u_1, \ldots, u_r] \in \mathbb{R}^{C \times r}. \tag{23}$$

The flattened pixel-wise query features $Q \in \mathbb{R}^{N \times C}$, are subsequently projected into this subspace and L_2-normalized:

$$z_i = \frac{Q_i U_r}{\|Q_i U_r\|_2}, \tag{24}$$

where Q_i denotes the original C-dimensional feature of the i-th pixel and $z_i \in \mathbb{R}^r$ is its normalized subspace representation. Based on the predicted mask, we sample foreground and background query representations and jointly project them, together with the foreground and background concept prototypes μ_{fg} and μ_{bg}, onto a two-dimensional plane using Principal Component Analysis (PCA).

As shown in Fig. 26, the visualization reveals two core advantages of our module. ***First***, the subspace features exhibit high linear separability. The foreground features (**red** dots) and background features (**blue** dots) are distinctly separated with a clear margin, proving that the U_r projection effectively suppresses task-irrelevant noise and maximizes inter-class variance. ***Second***, the model effectively preserves intra-class variance. It can be observed that the pixel features do not suffer from severe feature collapse towards their respective prototypes μ_{fg} or μ_{bg} (marked as crosses). Instead, they form a well-structured hyperspherical distribution around the prototypes. This indicates that while decoupling the foreground and background, the model retains the inherent fine-grained texture variations of different object parts and complex backgrounds, which is crucial for the spatial granularity required in dense prediction tasks.

F Discussion

F.1 Does Concept Discriminative Subspace Progressively Build Discriminative?

In our online concept discriminative subspace evolution, the concept discriminative subspace U_r is updated incrementally. When the current discriminative direction becomes insufficiently represented by the accumulated historical subspace, the subspace-reset mechanism discards outdated historical information. The online evolution process can therefore be divided into several subspace evolution intervals separated by reset events. To verify that the subspace within each interval becomes progressively more comprehensive, we introduce a piecewise "offline" segmentation setting. Specifically,

Table 6: Performance comparison of online and offline on different CD concepts.

Methods	Camouflaged		Salient		Shadow		Transparent		Thyroid Nodule		Polyp		COVID-19		Liver Tumor	
	MAE ↓	S_m ↑	MAE ↓	S_m ↑	MAE ↓	BER ↓	MAE ↓	BER ↓	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑
offline	.026	.807	.067	.751	.025	.100	.067	.066	.827	.731	.852	.786	.971	.951	.878	.872
online	.028	.802	.070	.742	.026	.103	.069	.069	.824	.731	.848	.783	.970	.951	.876	.870

for each subspace evolution interval, we extract the finalized concept subspace obtained just before the subspace reset is triggered (or at the end of the sequence) and utilize this fully-evolved subspace to re-segment all preceding frames within that specific interval. As shown in Tab. 6, this piecewise offline strategy yields consistently better performance than the standard online process. This confirms our intuition: throughout the continuous update process, the concept subspace successfully accumulates and integrates newly observed visual cues. Consequently, the finalized subspace provides a more robust and complete representation of the target object compared to the early-stage, partially evolved subspace. As shown in Fig. 27, our concept discriminability gradually adjusts to an optimal state over time within the same scene, clearly illustrating the dynamic adaptation process of subspace evolution.

F.2 Does Concept Discriminative Subspace Evolution Effectively Mitigate Concept Drift?

In dynamic scenes, contextual information often undergoes significant changes, leading to the problem of concept drift. To address this challenge, we propose a concept discriminative subspace evolution mechanism, which continuously updates the concept discriminative subspace (CDS) by integrating both historical information and current-frame observations, enabling the model to adapt to evolving scene semantics. In our ablation experiments, we validated the effectiveness of this design: removing either historical information or current-frame guidance leads to a notable drop in performance, indicating that both components play a crucial role in mitigating concept drift.

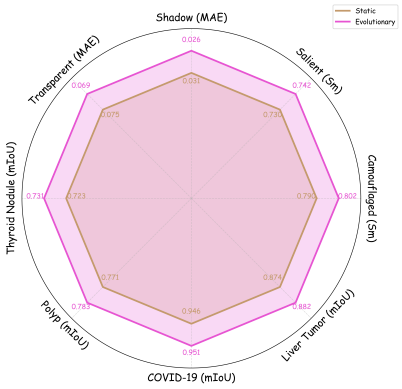

Figure 28: Quantitative comparison of static *vs.* evolving subspaces.

To further demonstrate the necessity of concept discriminative subspace evolution, we conducted both quantitative and qualitative analyses. As shown in Fig. 27, our concept discriminability gradually adjusts to an optimal state over time within the same scene, clearly illustrating the dynamic adaptation process of subspace evolution. Fig. 28 shows that the evolving subspace consistently outperforms the static subspace, validating the effectiveness of continuous subspace adaptation. To further assess its robustness to concept drift, Tab. 7 compares the two variants on ***Video-CD***, where the evolving CDS consistently achieves better performance across the evaluated concepts. Moreover, Fig. 29 provides qualitative visualizations: even in the presence of substantial concept drift, ***EvoCDS*** is still able to accurately segment the target object, while the static subspace suffers from severe prediction errors due to its inability to adapt to semantic shifts.

In summary, ***EvoCDS*** continuously adjusts concept representations, allowing it to more effectively handle concept drift in dynamic scenes.

G Comparative Analysis of Different Methods on Number of Prompts

In Tab. 8, we compare three methods across eight segmentation concepts, with the number of prompts as the key variable. SR-ICL and ***EvoCDS*** both use 12 prompts, while Spider uses 64 prompts. Despite having over five times more prompts, Spider does not consistently outperform the other methods. In contrast, ***EvoCDS*** achieves the best performance on most tasks and evaluation metrics using only 12 prompts. For example, on the camouflaged concept, our MAE drops to 0.024

Table 7: Comparison of static and evolving concept discriminative subspaces on ***Video-CD***.

Methods	Camouflaged		Salient		Shadow		Transparent		Thyroid Nodule		Polyp		COVID-19		Liver Tumor	
	MAE ↓	S_m ↑	MAE ↓	S_m ↑	MAE ↓	BER ↓	MAE ↓	BER ↓	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑
Static	.024	.750	.081	.735	.033	.159	.079	.083	.826	.737	.849	.787	.965	.946	.836	.826
Evolving	.020	.759	.079	.739	.031	.148	.077	.081	.832	.743	.856	.789	.970	.951	.841	.832

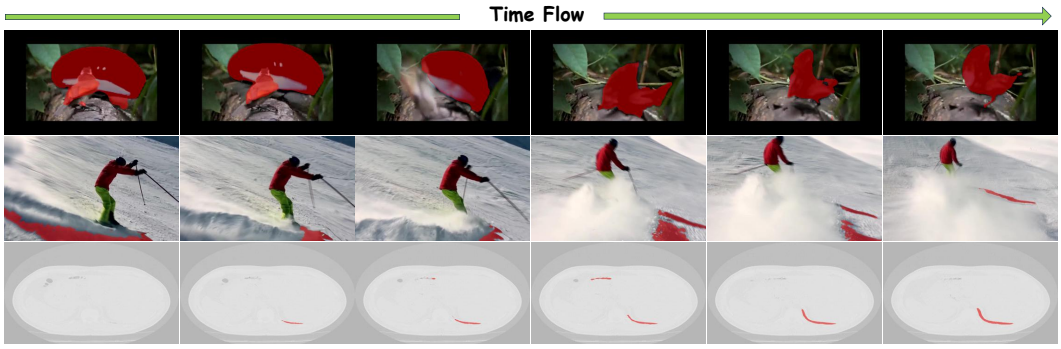

Figure 29: Qualitative results under significant concept drift.

(*vs.* 0.034 for SR-ICL and 0.034 for Spider), and our S_m rises to 0.837 (*vs.* 0.781 and 0.776). Similar improvements are observed across all eight concepts, including salient, shadow, transparent, nodule, polyp, COVID-19, and liver tumor concepts. These results demonstrate that increasing the number of prompts does not guarantee better performance; instead, the quality and design of prompts are the decisive factors.

The strong performance of ***EvoCDS*** with only a small number of prompts can be attributed to the carefully designed concept discriminative subspace (CDS), which extracts compact and task-relevant discriminative information from the provided concept-definition samples. Specifically, CDS improves prompt efficiency through three complementary aspects: (1) Concept discriminative capability: CDS models foreground-background discriminative directions, enabling the model to focus on task-relevant concept information while suppressing less informative variations; (2) Scene understanding capability: CDS incorporates local discriminative information to model contextual relationships within individual scenes, allowing a limited number of prompts to provide effective scene-specific concept cues. Experiments show that removing local discriminative information consistently degrades performance; (3) Concept drift adaptation capability: CDS further integrates discriminative information from both historical and current frames, allowing the concept representation to adapt as the visual context evolves over time. Ablation results show that removing either historical or current discriminative information leads to performance degradation, supporting their complementary roles in dynamic scenarios. Compared with Spider, which uses 64 prompts, ***EvoCDS*** achieves stronger overall performance with only 12 prompts. Moreover, increasing the number of concept-definition samples beyond 12 yields only marginal gains, indicating that CDS can capture sufficient concept-discriminative information from a relatively small sample set. These results suggest that effectively extracting and organizing discriminative concept information can be more important than simply increasing the number of prompts, demonstrating the sample efficiency of the proposed CDS for unified context-dependent segmentation.

H The Reconstruction Error

The reconstruction error threshold controls the sensitivity of the drift-aware subspace reset rather than the regular frame-by-frame update of the concept representation. For each valid frame, the current difference vector is incorporated into the concept matrix to support gradual subspace evolution. The threshold determines whether the accumulated historical difference vectors should be retained or discarded. An excessively low threshold may trigger frequent subspace resets due to minor feature fluctuations, whereas an excessively high threshold may retain outdated historical directions and prevent timely adaptation to substantial temporal variations. We therefore conduct an ablation study in Tab. 9 to investigate the influence of different reset thresholds. Specifically, the reconstruction

Table 8: Comparative analysis of different methods on number of prompts.

Methods	Camouflaged		Salient		Shadow		Transparent		Thyroid Nodule		Polyp		COVID-19		Liver Tumor	
	MAE ↓	S_m ↑	MAE ↓	S_m ↑	MAE ↓	BER ↓	MAE ↓	BER ↓	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑
Prompt = 12 (SR-ICL)	.034	.781	.077	.753	.033	.152	.088	.086	.812	.731	.825	.754	.964	.945	.871	.857
Prompt = 64 (Spider)	.034	.776	.079	.763	.036	.151	.086	.091	.810	.722	.825	.742	.965	.946	.871	.849
Prompt = 12 (Ours)	.024	.837	.064	.783	.032	.154	.073	.073	.819	.727	.831	.764	.974	.956	.897	.886

Table 9: Performance comparison of τ_{rec} on different CD concepts.

Methods	Camouflaged		Salient		Shadow		Transparent		Thyroid Nodule		Polyp		COVID-19		Liver Tumor	
	MAE ↓	S_m ↑	MAE ↓	S_m ↑	MAE ↓	BER ↓	MAE ↓	BER ↓	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑
τ^1_{rec}	.029	.800	.071	.741	.028	.106	.070	.071	.822	.730	.846	.782	.969	.950	.874	.867
τ^2_{rec}	.028	.802	.070	.742	.026	.103	.069	.069	.824	.731	.848	.783	.970	.951	.876	.870
τ^3_{rec}	.030	.798	.072	.739	.027	.100	.070	.070	.821	.729	.845	.780	.967	.948	.873	.865

error of the current frame is defined as:

$$E^t_{\text{rec}} = \left\| d^t_{\text{now}} - d^t_{\text{recon}} \right\|_2, \tag{25}$$

where d^t_{now} denotes the current foreground-background concept difference vector at frame t, and d^t_{recon} represents its reconstruction from the historical concept discriminative subspace. A larger reconstruction error indicates that the current discriminative direction is insufficiently represented by the historical concept subspace, suggesting that retaining the accumulated historical directions may no longer be appropriate.

To adaptively determine whether the historical discriminative information should be retained or reset, the threshold is calculated from the statistical distribution of historical reconstruction errors:

$$\tau_{\text{rec}} = \mu_{\text{rec}} + \lambda \sigma_{\text{rec}}, \tag{26}$$

where μ_{rec} and σ_{rec} denote the mean and standard deviation of historical reconstruction errors, respectively. They are calculated as:

$$\mu_{\text{rec}} = \frac{1}{N} \sum_{i=1}^{N} E^i_{\text{rec}}, \tag{27}$$

and

$$\sigma_{\text{rec}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(E^i_{\text{rec}} - \mu_{\text{rec}} \right)^2}, \tag{28}$$

where N represents the number of historical frames used for threshold estimation. The dynamic concept update is activated when:

$$E^t_{\text{rec}} > \tau_{\text{rec}}, \tag{29}$$

where λ controls the sensitivity of the update strategy. To evaluate the robustness of the proposed adaptive threshold mechanism, three threshold configurations are compared:

$$\begin{cases} \tau^1_{\text{rec}} = \mu_{\text{rec}} + \sigma_{\text{rec}}, \\ \tau^2_{\text{rec}} = \mu_{\text{rec}} + 2\sigma_{\text{rec}}, \\ \tau^3_{\text{rec}} = \mu_{\text{rec}} + 3\sigma_{\text{rec}}. \end{cases} \tag{30}$$

These three configurations correspond to $\lambda = 1$, $\lambda = 2$, and $\lambda = 3$, respectively. The first setting provides a more sensitive reset strategy, in which deviations from the historical concept subspace trigger reconstruction more frequently. The second setting introduces a moderate reset criterion, balancing adaptation to substantial subspace deviations with the preservation of useful historical information. The third setting adopts a more conservative strategy, resetting the historical subspace only when substantial deviations occur. These threshold configurations correspond to different sensitivities of the history-reset mechanism. A smaller threshold increases the frequency of subspace reconstruction and enables a faster response to abrupt changes, but may discard useful historical information because of minor fluctuations. Conversely, a larger threshold preserves historical directions more conservatively, but may delay adaptation when the current discriminative evidence

Table 10: Effect of different confidence thresholds for foreground and background pixel selection.

Methods	Camouflaged		Salient		Shadow		Transparent		Thyroid Nodule		Polyp		COVID-19		Liver Tumor	
	MAE ↓	S_m ↑	MAE ↓	S_m ↑	MAE ↓	BER ↓	MAE ↓	BER ↓	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑
$y(f) > 0.5$ and $y(b) < 0.5$	.029	.800	.071	.741	.028	.106	.071	.072	.822	.730	.846	.782	.968	.949	.874	.868
$y(f) > 0.7$ and $y(b) < 0.3$	.028	.802	.070	.742	.026	.103	.069	.069	.824	.731	.848	.783	.970	.951	.876	.870

changes substantially. The best performance achieved with $\lambda = 2$ indicates that the proposed criterion provides an effective balance between temporal adaptability and representational stability. Since the reconstruction error functions as an internal subspace-management criterion rather than a standalone drift classifier, its effectiveness is evaluated through downstream segmentation performance. We investigate the confidence thresholds used to select foreground and background pixels for concept construction. Specifically, we compare the conventional threshold setting of $y(f) > 0.5$ and $y(b) < 0.5$ with the stricter setting of $y(f) > 0.7$ and $y(b) < 0.3$. As shown in Tab. 10, the stricter thresholds achieve better overall performance. This indicates that filtering out uncertain pixels produces more reliable foreground and background concept representations, thereby reducing noisy concept updates and improving the robustness of the evolving discriminative subspace. We emphasize that the reconstruction error criterion is used only as an internal signal for updating the concept subspace and is not intended to provide supervised or frame-level concept-drift localization.

I *Video-CD*

To systematically evaluate ***EvoCDS***, we constructed ***Video-CD***, a benchmark for concept drift. To eliminate self-referential bias (i.e., relying on the failure cases of specific tracking or segmentation models to mine hard samples, which inherently biases the dataset toward the drift types those models are sensitive to) and ensure absolute neutrality in evaluation, we propose a three-stage construction pipeline: (1) candidate retrieval based on an Independent Drift Metric Bank; (2) MLLM-Assisted Pre-annotation; and (3) Human Arbitration. It is worth emphasizing that ***Video-CD*** is built upon existing large-scale, fully-annotated video datasets, enabling us to directly utilize ground-truth (GT) masks as reliable references for measuring drift-sensitive structural and discriminative changes.

I.1 Stage 1: Independent Drift Metric Bank

Discriminative Evidence Shift (DES). Concept drift frequently manifests as the evolution of visual evidence that distinguishes the foreground from the background (e.g., appearance variations, background clutter), rather than a direct change in the target's categorical identity. To measure this evolution while avoiding task-specific encoder bias, we employ a frozen backbone (Liu et al., 2021) for feature extraction. Given the GT mask, we compute the foreground-background discriminative directional vector d_t for frame t:

$$d_t = \mu_t^{fg} - \mu_t^{bg}, \tag{31}$$

where μ_t^{fg} and μ_t^{bg} denote the average features of the target foreground and its corresponding background region, respectively. The DES score quantifies the frame-to-frame change in the discriminative direction by calculating the cosine deviation between the discriminative vector of the current frame and a previous-frame d_{t-1}:

$$S_{DES}(t) = 1 - \cos(d_t, d_{t-1}). \tag{32}$$

Structural-Topological Shift (STS). To capture severe geometric deformation, target fragmentation, or topological collapse, we propose the STS metric. This metric relies entirely on GT masks, thereby providing the strongest defense against feature-level bias. We first normalize the current mask M_t and the reference mask M_{t-1} to achieve translation and scale invariance. The structural shift is calculated as follows:

$$S_{STS}(t) = (1 - \text{IoU}(M_t, M_{t-1})), \tag{33}$$

Temporal-Context Shift (TCS). Drift is also frequently triggered by scene transitions, abrupt motion, or contextual changes. We quantify this by evaluating the consistency of temporal deformation:

$$S_{TCS}(t) = 1 - \left(\text{IoU}(W_{t-1\to t}(M_{t-1}), M_t)\right), \tag{34}$$

where W represents the optical flow warping operator (Brox et al., 2004). Since substantial drift-related changes may appear as local deviations from the typical temporal dynamics of a sequence, we employ the median absolute deviation (MAD) (Leys et al., 2013) for robust metric normalization:

$$z_k(t) = \frac{S_k(t) - \text{median}(S_k)}{1.4826 \cdot \text{MAD}(S_k)}, \quad k \in \{DES, STS, TCS\}. \tag{35}$$

To prioritize high recall during candidate retrieval, we adopt a union-based trigger mechanism. If any of the three normalized metrics exhibits a significant deviation, the corresponding temporal sequence is extracted as a candidate clip for subsequent review.

I.2 Stage 2: MLLM-Assisted Pre-annotation

Since the candidate clips retrieved by the metric bank include both drift-containing sequences and hard-but-stable segments (false positives), we introduce a Multimodal Large Language Model (MLLM) (Wang et al., 2024) as a drafting agent rather than a hard filter. This accelerates human review without increasing the risk of false-negative filtering. For each candidate clip, the MLLM generates a structured pre-annotation report based on uniformly sampled frames and their temporal context, including: (1) the estimated probability that the candidate clip contains substantial concept drift; (2) a preliminary taxonomy label; and (3) a textual rationale. The MLLM does not assign frame-level drift labels or determine a specific drift onset. Crucially, it does not have the authority to directly discard clips; all candidate clips and their corresponding MLLM reports are submitted entirely to the final human review stage.

I.3 Stage 3: Human Arbitration

In the final stage, three domain experts serve as the ultimate adjudicators. To ensure maximum reliability and reduce individual annotation bias, each candidate video sequence is independently reviewed by all three experts through an exhaustive, end-to-end manual inspection process. Although Stages 1 and 2 are provided as auxiliary references, the experts strictly examine each clip from beginning to end on a frame-by-frame basis. The final decision is determined through a majority voting mechanism among the three experts. The core task of human annotation is to comprehensively verify whether each candidate video clip contains substantial task-relevant concept drift.

I.4 Video-level Concept Drift Annotation.

Concept drift in real-world sequences is often gradual and temporally extended rather than occurring at a single well-defined frame. Changes in appearance, visibility, shape, target-context relationships, or scene structure may accumulate over multiple consecutive frames, making an exact frame-level drift boundary inherently ambiguous and potentially subjective. Therefore, ***Video-CD*** adopts a video-level annotation protocol rather than assigning binary drift labels to individual frames. Specifically, each selected video is verified to contain substantial evolution in the task-relevant discriminative evidence of a semantically consistent target, while the precise onset and duration of the drift are not explicitly annotated. Accordingly, ***Video-CD*** is designed to evaluate segmentation robustness over drift-containing sequences, rather than frame-level drift detection or localization. This protocol avoids introducing potentially unreliable frame-wise drift labels and better reflects the continuous nature of concept evolution in realistic dynamic scenarios.

Since ***Video-CD*** does not assign binary drift labels to individual frames, all segmentation metrics are computed over the complete drift-containing sequences. The benchmark therefore measures whether a segmentation model can maintain reliable concept discrimination throughout videos in which task-relevant evidence evolves, rather than measuring the accuracy of drift detection at particular timestamps. Consequently, improvements on Video-CD should be interpreted as stronger robustness to drift-containing sequential scenarios, rather than as evidence of precise frame-level drift localization.

J Limitations and Future Work

As an initial step toward video context-dependent (VCD) concept segmentation, our ***EvoCDS*** still presents several limitations and opportunities for future improvement. ***First***, the current concept

evolution mechanism is mainly reactive, where concept updates rely on high-confidence predictions from previous frames. When severe concept drift leads to inaccurate predictions, erroneous updates may potentially affect subsequent evolution. Future work will investigate more proactive evolution strategies by incorporating uncertainty estimation, predictive modeling, and longer-term structured memory mechanisms beyond the lightweight historical difference-vector memory used in the current framework. ***Second***, ***EvoCDS*** requires a small number of concept definition samples to initialize task-specific discriminative subspaces. Although this requirement is much lighter than conventional fully supervised training, obtaining reliable concept definitions may still involve human annotation. Future research could explore automatic concept discovery and vision-language guided concept initialization for more flexible zero-shot or open-world scenarios. ***Third***, our framework primarily models concept evolution in the visual feature space. However, real-world video understanding involves richer temporal dynamics, object interactions, and multimodal information. Extending evolving concept representations with motion-aware, language-guided, and long-term memory mechanisms represents an important direction for future research. Furthermore, we plan to extend ***EvoCDS*** to additional VCD concept segmentation tasks, such as video industrial defect detection, video inharmonious region localization, and video defocus blur detection. Our future work will also explore how this subspace evolution can facilitate zero-shot adaptation and open-world concept discovery in dynamic video streams.