TRACING OBJECT EVIDENCE ACROSS DEPTH: TEXT-MODULATED MEMORY FRAMEWORK FOR WEAKLY SUPERVISED SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Weakly supervised semantic segmentation (WSSS) learns pixel-level predictions from image-level category labels. While CLIP provides rich semantic priors for localization, recovering complete objects requires integrating complementary evidence across visual layers. Conventional multi-layer aggregation lacks an explicit mechanism for retaining and revising concept-specific evidence as semantic context deepens, leaving local structures weakly tied to category semantics and producing incomplete object coverage and spurious background responses. We propose a text-modulated cross-layer xConvLSTM framework that formulates visual-layer integration as concept-conditioned memory evolution across network depth. Treating selected CLIP layers as an ordered sequence, the framework uses text representations to control how spatial memory incorporates, retains, and suppresses visual evidence, progressively connecting local structures and object parts to category semantics. To adapt this process to each image while preserving CLIP's semantic priors, we introduce anchored adaptive prompting, which augments frozen text embeddings with bounded learnable and image-conditioned residuals. The resulting concept-specific memories calibrate visual tokens, guide decoding, and refine pixel relations, coupling evidence accumulation across depth with response propagation across space. Extensive experiments on WSSS benchmarks demonstrate more complete object localization, higher-quality pseudo-labels, and consistent segmentation gains.

1 INTRODUCTION

Weakly supervised semantic segmentation (WSSS) learns pixel-level predictions from image-level category labels. However, these labels provide no direct supervision of object location, extent, or boundaries. A common strategy generates class activation maps (CAMs) (Zhou et al., 2016) and refines them into segmentation pseudo-labels. Representative methods, including SEC (Kolesnikov & Lampert, 2016), AffinityNet (Ahn & Kwak, 2018), IRNet (Ahn et al., 2019), and SEAM (Wang et al., 2020), improve localization through expansion constraints, pixel relations, or consistency learning. Yet classification can rely on only a few discriminative parts, causing CAMs to overlook less distinctive regions and limiting complete object recovery.

Vision-language pre-training provides richer semantic priors for addressing this challenge. CLIP (Radford et al., 2021) learns visual-textual correspondences from large-scale image-text pairs, enabling category descriptions to serve as semantic queries for localization. CLIP-ES (Lin et al., 2023) demonstrates the effectiveness of these priors for generating and refining CAMs. Subsequent approaches improve different aspects of dense prediction: WeCLIP+ (Zhang et al., 2025) combines frozen CLIP and DINO representations; ExCEL (Yang et al., 2025a) enhances patch-text alignment through text semantic enrichment and visual calibration; and VDA (Qiu et al., 2026) constructs instance-adaptive visual queries with text guidance. These advances strengthen semantic responses and their spatial refinement, while leaving open how complementary evidence for the same object concept can be organized across the depth of the visual encoder.

In practice, object evidence is distributed across visual representations at different depths. Earlier features tend to retain local appearance and structural details, whereas deeper features provide more

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.

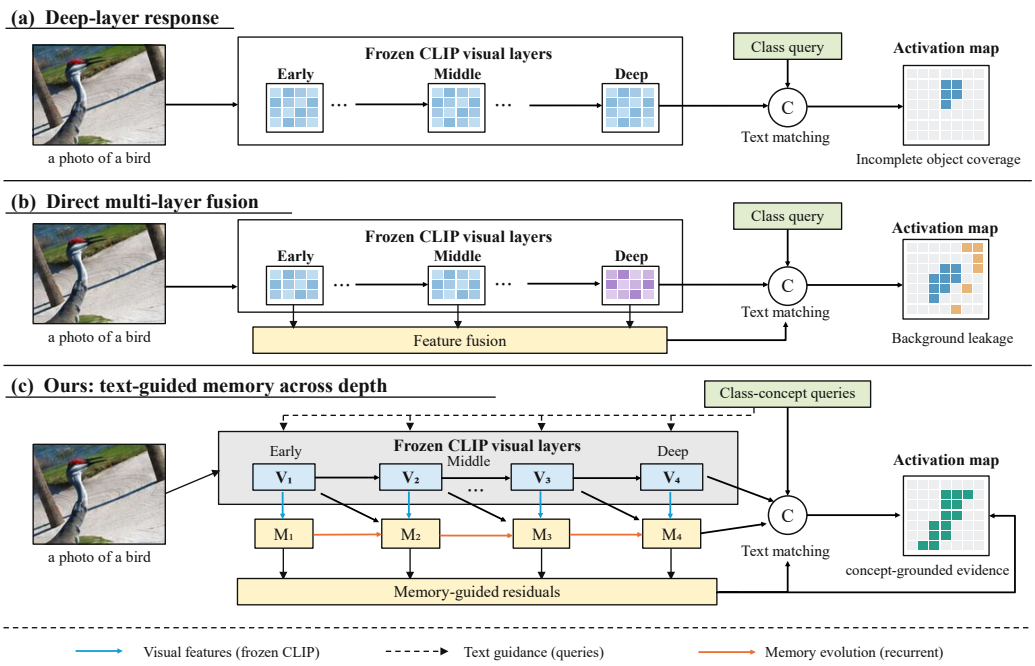


Figure 1: Motivation of the proposed cross-layer semantic memory.

abstract, category-discriminative information. These cues are complementary: a local contour may delineate a limb without identifying its category, while a strong category response may identify an animal without covering its limbs. Complete localization thus requires connecting local evidence with semantic context. A deep-layer response provides only one view of this distributed evidence. Spatial propagation can extend it to related pixels, but depends on the quality of semantic seeds and pixel relations: weak object activations or misleading background responses may lead to persistent omissions or amplified errors.

Fig. 1 illustrates our motivation: selectively accumulating and revising object evidence along network depth can connect local structural cues with deeper category semantics. Deep-layer readout offers limited access to this complementary evidence, while direct multi-layer fusion does not explicitly maintain a concept-specific state across layers. Concatenation, summation, and weighted aggregation can exploit multiple representations, but do not inherently specify how earlier evidence should be retained or revised as deeper features become available. Likewise, Transformer residual connections carry information between blocks without explicitly conditioning its retention on a queried object concept. This motivates treating depth as an ordered axis for concept-specific memory updates, preserving useful local evidence while suppressing responses inconsistent with the target concept. We therefore ask: *How can object evidence be selectively accumulated and preserved across network depth to support more complete localization under image-level supervision?*

To address this question, we propose a text-modulated cross-layer memory framework that traces object evidence through selected CLIP visual layers. We organize these layers in their original depth order and adapt xLSTM (Beck et al., 2024) to progressively update a spatial memory while preserving the two-dimensional arrangement of visual tokens. For each class-concept query, text representations modulate the integration of current visual features with previously accumulated evidence, guiding which information is incorporated, retained, or suppressed. This process allows local structures and object-part cues to remain available as deeper semantic context enters the memory, enabling text to guide evidence accumulation throughout network depth. To accommodate variations in object appearance, pose, and context, we further introduce *anchored adaptive prompting*, which augments frozen CLIP text embeddings with bounded learnable and image-conditioned residuals. These residuals provide task- and image-adaptive conditions for memory updates, while the frozen anchors and bounded adaptation constrain semantic drift. The resulting concept memories calibrate visual tokens and guide the segmentation decoder. We also retain the topology of the baseline refine-

ment module (RFM) and use the learned cross-layer representations to refine its pixel-relation estimates, connecting evidence accumulation across depth with response propagation across spatially related pixels. Experiments on standard WSSS benchmarks demonstrate more complete activation maps, improved pseudo-label quality, and consistent gains in segmentation performance.

Contributions. The main contributions of this work are summarized as follows: (**i**) We formulate object evidence aggregation for WSSS as an ordered process across network depth. A spatial cross-layer xLSTM maintains explicit memory over selected CLIP layers, allowing complementary visual evidence to be accumulated and revised before spatial response propagation. (**ii**) We introduce concept-level text modulation into the memory updates. Class-concept queries guide evidence selection, retention, and suppression throughout the layer sequence, producing concept-specific spatial representations for visual calibration and segmentation. (**iii**) We develop anchored adaptive prompting that combines frozen CLIP text anchors with bounded learnable and image-conditioned residuals. This mechanism supplies adaptive semantic conditions for cross-layer memory while constraining deviation from the pre-trained text representations.

2 Related Work

Weakly supervised semantic segmentation. Weakly supervised semantic segmentation (WSSS) learns pixel-level predictions from inexpensive annotations, with image-level labels being one of the most challenging settings. Most methods generate class activation maps (CAMs) (Zhou et al., 2016) and refine them into pseudo-labels. Early approaches improve incomplete CAMs through spatial constraints, affinity propagation, boundary modeling, and transformation consistency, including SEC (Kolesnikov & Lampert, 2016), AffinityNet and IRNet (Ahn et al., 2019), and SEAM (Wang et al., 2020). Transformer-based methods further exploit class tokens, self-attention, and token-level consistency, as demonstrated by MCTformer (Xu et al., 2022), AFA (Ru et al., 2022), and ToCo (Ru et al., 2023). Recent methods such as PCRE (Xu et al., 2025), FFR (Yang et al., 2025b), and DFA (Yang et al., 2026) improve region expansion and boundary preservation. Nevertheless, most existing approaches rely on a single deep representation or static multi-layer aggregation and do not explicitly model how semantic evidence evolves across network depth.

Vision-language models for segmentation. Vision-language pre-training provides rich semantic priors for segmentation. CLIP (Radford et al., 2021) establishes a shared image-text embedding space, while DenseCLIP (Rao et al., 2022), MaskCLIP (Zhou et al., 2022a), and CLIP-ES (Lin et al., 2023) extend its global alignment to dense localization and WSSS. DINO (Caron et al., 2021) and DINOv2 (Oquab et al., 2023) complement CLIP with locally consistent structural features, and WeCLIP+ (Zhang et al., 2025) combines their semantic and spatial advantages. More recent methods, including ExCEL (Yang et al., 2025a) and VDA (Qiu et al., 2026), improve patch-text alignment and adapt visual descriptions to the input. However, these methods mainly introduce textual or visual priors during CAM generation or at the final representation. In contrast, our method uses layer-adaptive category-concept semantics to continuously guide evidence selection and accumulation across multiple CLIP layers.

Cross-layer feature and memory modeling. Features at different network depths provide complementary information: shallow layers preserve local structures, whereas deep layers encode abstract category semantics. Conventional segmentation models combine them through skip connections, concatenation, summation, or hierarchical decoding, but these operations generally lack a concept-specific state that evolves with network depth. ConvLSTM (Shi et al., 2015) introduces spatial recurrent gating, and xLSTM (Beck et al., 2024) improves long-range modeling through enhanced gates and stabilized memory updates. Existing recurrent visual models, however, primarily process temporal frames or iterative predictions. Our method instead treats selected CLIP layers as an ordered semantic sequence and maintains text-conditioned two-dimensional xConvLSTM states for category-concept queries. It therefore accumulates semantic evidence along network depth, while RFM subsequently propagates the resulting responses among spatially related pixels.

3 METHOD

3.1 OVERALL FRAMEWORK

Given an image I and its image-level label $\mathbf{y} \in \{0,1\}^K$, our goal is to learn pixel-level segmentation without dense annotations. Our framework is built upon frozen CLIP-DINO representations and the original RFM pipeline of WeCLIP+. It combines text-guided evidence accumulation across network depth with spatial response propagation. The framework consists of three main components: category-concept query construction and adaptation, text-modulated cross-layer memory, and memory-guided segmentation and refinement.

We first construct a category-concept text bank and adapt its queries to each image through *anchored adaptive prompting*, using bounded residuals while retaining frozen CLIP text embeddings as semantic anchors. A spatial xConvLSTM then processes selected CLIP layers $\mathcal{L} = \{2, 5, 8, 10\}$ in depth order, with the adapted queries guiding the incorporation, retention, and suppression of visual evidence during memory updates. Preserving the two-dimensional token layout, this process connects local structural cues with deeper category semantics to form concept-specific spatial memories. These memories calibrate visual tokens, guide the segmentation decoder, and refine pixel-relation estimates within the unchanged RFM topology. Together, depth-wise evidence accumulation and spatial response propagation support more complete object recovery.

3.2 ANCHORED ADAPTIVE CONCEPT REPRESENTATION

Category-concept text bank. For each foreground category c_k, the frozen CLIP text encoder E_T encodes its category templates to obtain a category anchor:

$$\mathbf{t}_k = \mathrm{Norm}\left[\frac{1}{M_c}\sum_{m=1}^{M_c} E_T\big(\mathcal{T}_m(c_k)\big)\right], \tag{1}$$

where $\mathcal{T}_m(c_k)$ is the m-th category template, M_c is the number of templates, and $\mathrm{Norm}(\cdot)$ denotes L_2 normalization.

To describe local object structures, each category is associated with a set of concepts, such as *head*, *arm*, and *leg* for *person*. The base representation of concept $p_{k,j}$ is

$$\mathbf{q}^0_{k,j} = \mathrm{Norm}\left[\frac{1}{M_p}\sum_{m=1}^{M_p} E_T\big(\mathcal{T}_m(c_k, p_{k,j})\big)\right], \tag{2}$$

where $\mathcal{T}_m(c_k, p_{k,j})$ is a category-conditioned concept template and M_p is the number of concept templates.

We further represent the semantic difference between the concept and its parent category as

$$\mathbf{r}^0_{k,j} = \mathrm{Norm}\left(\mathbf{q}^0_{k,j} - \mathbf{t}_k\right). \tag{3}$$

The category anchor retains global category semantics, whereas $\mathbf{r}^0_{k,j}$ captures concept-specific information. These concepts are textual queries and require no part annotations. The templates are detailed in Appendix B.1.

Anchored adaptive prompt. Fixed concept texts cannot fully accommodate image-specific variations. We therefore adapt the concept residual using a learnable semantic offset and a layer-wise image-conditioned offset:

$$\mathbf{r}^{I,l}_{k,j} = \mathrm{Norm}\left[\mathbf{r}^0_{k,j} + \alpha_s \tanh\left(\boldsymbol{\delta}_{k,j}\right) + \alpha_I \tanh\left(G(\mathbf{g}^l)\right)\right], \tag{4}$$

where $\boldsymbol{\delta}_{k,j}$ is a category-concept-specific learnable residual, $\mathbf{g}^l$ is the CLS token of the l-th selected CLIP layer, and $G(\cdot)$ is a shared image-conditioning network. The coefficients α_s and α_I bound the two offsets. A cosine anchoring constraint limits semantic drift from the original CLIP text representation.

Since not all concepts are visible in every image, an adaptive router estimates their reliability from $\mathbf{g}^l$, $\mathbf{t}_k$, and $\mathbf{r}^{I,l}_{k,j}$. A null concept and Top-K selection suppress irrelevant concepts. All valid concept states are still updated, while Top-K is applied only to semantic readout and feedback. The routing design is described in Appendix B.2.

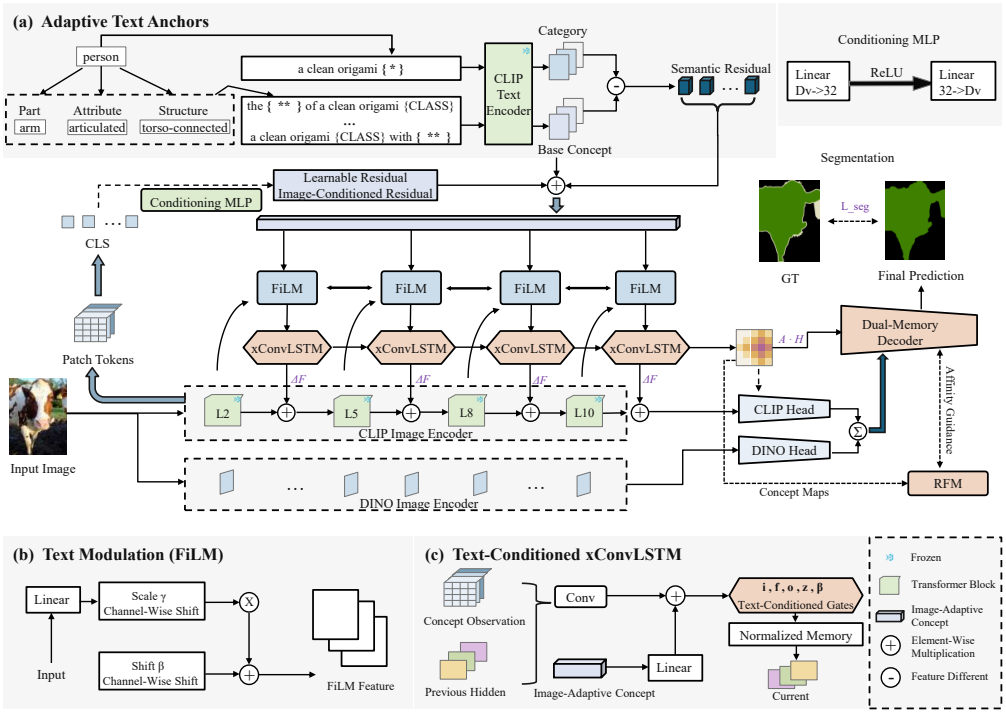


Figure 2: Overview of our weakly supervised semantic segmentation framework: (a) adaptive text anchors align frozen semantic embeddings with image content; (b) text modulation conditions visual features on semantic concepts; and (c) text-conditioned xConvLSTM aggregates spatial memory across visual depths. Together, they support segmentation with adaptive semantic guidance and cross-layer memory under CAM-derived pseudo-label supervision.

3.3 Text-Modulated Cross-Layer xLSTM

Layer-wise semantic observation. For each selected layer l, its patch tokens are reshaped into a spatial feature map and projected to a common dimension, producing $\mathbf{X}^l \in \mathbb{R}^{D\times H\times W}$. We first calculate the category support map:

$$\mathbf{A}_k^l(u,v) = \frac{\langle \mathbf{X}^l(u,v), \mathbf{t}_k \rangle}{\|\mathbf{X}^l(u,v)\|_2 \, \|\mathbf{t}_k\|_2}, \tag{5}$$

where (u,v) denotes a spatial location. To suppress concept responses that lack support from their parent category, we define

$$\mathbf{P}_k^l = \rho + (1-\rho)\sigma\left(\mathbf{A}_k^l\right), \tag{6}$$

where ρ is the minimum gate value. This lower bound preserves weak evidence in non-discriminative object regions.

The adaptive concept representation then modulates the visual feature through FiLM. Together with the local vision-text similarity, it forms the semantic observation:

$$\mathbf{Z}_{k,j}^l = \mathbf{P}_k^l \odot \left[\mathrm{FiLM}\left(\mathbf{X}^l; \mathbf{r}_{k,j}^{I,l}\right) + \lambda_d \,\mathrm{Sim}\left(\mathbf{X}^l, \mathbf{r}_{k,j}^{I,l}\right)\right], \tag{7}$$

where $\mathrm{FiLM}(\cdot)$ performs concept-conditioned channel scaling and shifting, while $\mathrm{Sim}(\cdot)$ produces the local concept-response map.

Cross-layer spatial memory. The selected layers are processed in the order $2 \to 5 \to 8 \to 10$. For each category-concept pair, the spatial state is updated as

$$\begin{aligned}&\left(\mathbf{H}_{k,j}^l, \mathbf{C}_{k,j}^l, \mathbf{N}_{k,j}^l, \mathbf{m}_{k,j}^l\right) \\ &\quad = \mathrm{xLSTM}_{2D}\left(\mathbf{Z}_{k,j}^l, \mathbf{H}_{k,j}^{l-1}, \mathbf{C}_{k,j}^{l-1}, \mathbf{N}_{k,j}^{l-1}, \mathbf{m}_{k,j}^{l-1}; \mathbf{r}_{k,j}^{I,l}\right),\end{aligned} \tag{8}$$

where $\mathbf{H}^l_{k,j}$ and $\mathbf{C}^l_{k,j}$ are the hidden and accumulated memory states, while $\mathbf{N}^l_{k,j}$ and $\mathbf{m}^l_{k,j}$ normalize and stabilize the cross-layer update. This memory progressively combines shallow structures with deeper semantic evidence. The text-conditioned gates and complete state update are given in Appendix B.3.

After the final selected layer, the routed concept states are projected and aggregated into a category-level concept map:

$$\mathbf{M}_k = \sum_{j \in \mathcal{K}_k} \bar{\pi}_{k,j}\, \sigma\left[\psi_m\left(\mathbf{H}^{L_{\max}}_{k,j}\right)\right], \tag{9}$$

where $\mathcal{K}_k$ is the Top-K concept set, $\bar{\pi}_{k,j}$ is the normalized routing weight, and $\psi_m(\cdot)$ projects the hidden state into a concept map.

The concept evidence is also injected into subsequent visual tokens through bounded token write-back, allowing later layers to receive previously accumulated semantic evidence without disrupting the frozen CLIP representation.

3.4 Segmentation and Pseudo-Label Learning

The cross-layer concept memory enhances the base segmentation prediction through a bounded residual:

$$\mathbf{S} = \mathbf{S}_{\text{base}} + \beta\, \mathbf{U}_{\text{base}} \odot \Psi_{\text{dec}}\left(\mathbf{M}, \overline{\mathbf{H}}, \overline{\mathbf{C}}\right), \tag{10}$$

where $\mathbf{S}_{\text{base}}$ is the base prediction, $\mathbf{U}_{\text{base}}$ measures its uncertainty, and $\overline{\mathbf{H}}$ and $\overline{\mathbf{C}}$ are the router-aggregated xLSTM states. The coefficient β bounds the feedback magnitude. This branch is zero-initialized and gradually activated during training.

We preserve the original RFM propagation topology. Initial CAMs are propagated using its pixel-relation matrix and refined by PAR:

$$\widehat{\mathbf{Y}} = \text{PAR}\left[\text{RFM}\left(\text{CAM}(I), \mathbf{R}\right)\right], \tag{11}$$

where $\mathbf{R}$ is the original RFM relation matrix and $\widehat{\mathbf{Y}}$ denotes the resulting pseudo-label. Thus, xLSTM integrates semantic evidence across network depth, whereas RFM propagates the resulting responses in the image space.

The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{mem}}\mathcal{L}_{\text{mem}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{aff}}\mathcal{L}_{\text{aff}}, \tag{12}$$

where $\mathcal{L}_{\text{seg}}$ and $\mathcal{L}_{\text{cls}}$ denote the segmentation and image-level classification losses. $\mathcal{L}_{\text{mem}}$ supervises cross-layer memory, $\mathcal{L}_{\text{con}}$ contains text anchoring and concept consistency constraints, and $\mathcal{L}_{\text{aff}}$ supervises the original RFM relation matrix.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on PASCAL VOC 2012 Everingham et al. (2010) and MS COCO 2014 Lin et al. (2014). Following the standard WSSS protocol, the augmented VOC training set is used for training, while results are reported on the validation and test sets. For MS COCO, we train on the training set and evaluate on the validation set. Only image-level category labels are used during training. Further details about the datasets and data splits are provided in Appendix C.

Implementation and evaluation. We evaluate our method under two independently trained configurations. The *controlled baseline configuration* is derived from WeCLIP+ by retaining the frozen CLIP ViT-B/16 and DINOv2 ViT-S/14 encoders while removing the additional RFM+ enhancement modules and using the original RFM propagation pipeline. The *full WeCLIP+ configuration* retains the complete WeCLIP+ pipeline, including its RFM+ modules, and integrates the proposed cross-layer components to evaluate their complementarity with a stronger system. In both configurations, the 3rd, 6th, 9th, and 11th CLIP visual blocks are used as the ordered cross-layer sequence. Models are optimized with AdamW and standard data augmentation. We report final segmentation, raw CAM, and RFM-refined pseudo-label mIoU, with all results given in percentage (%). Detailed architectural, training, inference, and evaluation settings are provided in Appendix C.

4.2 Comparison with Existing Methods

Table 1: Comparison with existing WSSS methods on PASCAL VOC 2012 and MS COCO 2014. Results are reported in mIoU (%).

Method	Sup.	Net.	VOC val	VOC test	COCO val
Multi-stage weakly supervised approaches					
AMN (Lee et al., 2022)	I	RN101	67.7	67.5	-
SIPE (Chen et al., 2022)	I	RN101	68.8	69.7	43.6
L2G (Jiang et al., 2022)	I+S	RN101	72.1	71.7	44.2
RCA (Zhou et al., 2022b)	I+S	RN38	72.2	72.8	36.8
CLIP-ES (Lin et al., 2023)	I+L	RN101	73.8	73.9	45.4
CPAL (Tang et al., 2024)	I+L	RN101	74.5	74.7	46.8
S2C (Kweon & Yoon, 2024)	I	SAM	78.2	77.5	49.8
Single-stage weakly supervised approaches					
AFA (Ru et al., 2022)	I	MiT-B1	66.0	66.3	38.9
ToCo (Ru et al., 2023)	I	ViT-B	71.1	72.2	42.3
SeCo (Yang et al., 2024)	I	ViT-B	74.0	73.8	46.7
DuPL (Wu et al., 2024)	I	ViT-B	73.3	72.8	44.6
WeCLIP (Zhang et al., 2024)	I+L	ViT-B	76.4	77.2	47.1
PCRE (Xu et al., 2025)	I	ViT-B	75.5	75.9	47.2
FFR (Yang et al., 2025b)	I	ViT-B	76.0	75.5	46.8
ExCEL (Yang et al., 2025a)	I+L	ViT-B	78.4	78.5	50.3
WeCLIP+ (Zhang et al., 2025)	I+L	ViT-B/16	84.0	83.9	**56.3**
DFA (Yang et al., 2026)	I+L	ViT-B	79.3	79.8	51.5
VDA (Qiu et al., 2026)	I+L	ViT-B	79.9	79.8	51.4
Ours 1 (Controlled baseline)	I+L	ViT-B/16	80.5	81.3	52.3
Ours 2 (Strong baseline)	I+L	ViT-B/16	**84.4**	**84.3**	**56.3**

Table 1 compares our method with representative multi-stage and single-stage WSSS approaches on PASCAL VOC 2012 and MS COCO 2014. We report both the controlled weak-baseline and full WeCLIP+ strong-baseline configurations to assess the proposed modules and their complementarity to a stronger system. Under the controlled configuration, our method achieves 80.5% and 81.3% mIoU on the VOC validation and test sets, respectively, exceeding VDA by 0.6% and 1.5% and DFA by 1.2% and 1.5%. These gains suggest that connecting local structural cues with deeper category semantics provides more complete object evidence for spatial refinement, even without the additional enhancements in the full WeCLIP+ pipeline. When integrated into full WeCLIP+, our method achieves 84.4% and 84.3% mIoU on the VOC validation and test sets, respectively, surpassing the original results of 84.0% and 83.9% by 0.4% on both splits. These additional gains suggest that concept-guided evidence accumulation across depth supplies useful cues that complement the existing WeCLIP+ pipeline. On the COCO validation set, the controlled configuration achieves 52.3% mIoU, outperforming VDA (51.4%) and DFA (51.5%). When integrated into full WeCLIP+, our method achieves 56.3% mIoU, matching the original WeCLIP+ result.

Figure 3 presents representative qualitative comparisons between our method, WeCLIP, and ExCEL on PASCAL VOC 2012 and MS COCO 2014. Compared with WeCLIP and ExCEL, our method recovers more complete foreground regions and better preserves the structures of thin, small, and partially occluded objects.

4.3 Component Ablation and Evidence Quality

We evaluate the proposed components through a cumulative construction under identical training and evaluation settings. L1 denotes the baseline model, L2 adds rich text anchors, L3 adds cross-layer xConvLSTM, L4 combines cross-layer xConvLSTM with text modulation, and L5 further adds rich text anchors to L4 and represents the complete model. The residual decoder and all other non-ablated components are enabled and kept fixed in all configurations. The progressive construction in the upper part of the table is evaluated without CRF, whereas the leave-one-out analysis in the lower part uses CRF for all configurations.

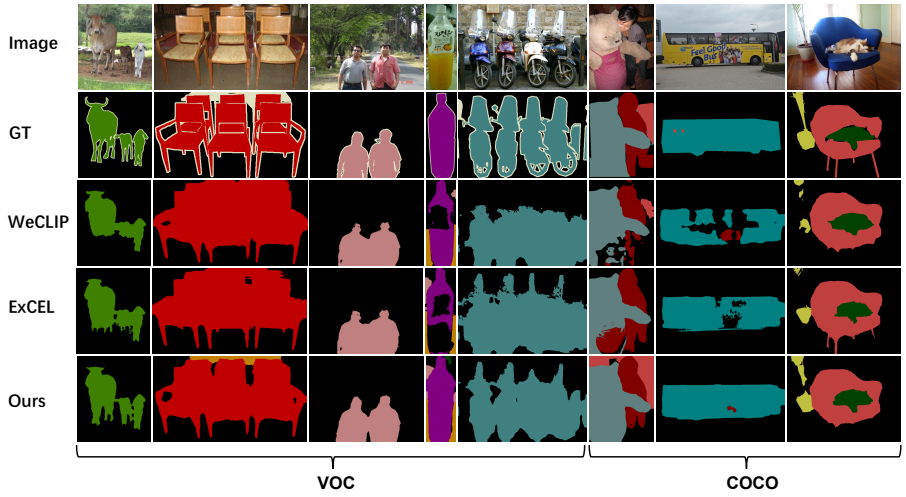


Figure 3: Qualitative comparison of final segmentation predictions among our method, WeCLIP, and ExCEL on PASCAL VOC 2012 and MS COCO 2014.

As shown in Table 2, adding rich text anchors in L2 obtains 46.8% raw CAM mIoU, 73.9% pseudo-label mIoU, and 78.0% segmentation mIoU. Adding cross-layer xConvLSTM in L3 further improves these results to 52.1%, 76.3%, and 78.8%, respectively, demonstrating the benefit of integrating complementary evidence across CLIP layers. L4 adds text modulation, raising raw CAM mIoU to 55.9%, while pseudo-label and segmentation mIoU are 76.2% and 78.5%, respectively. L5 combines rich text anchors, cross-layer xConvLSTM, and text modulation, achieving 55.6% raw CAM mIoU, 76.3% pseudo-label mIoU, and 79.2% segmentation mIoU. The leave-one-out results reported in the lower part of the table further confirm the contribution of the three components.

Table 2: Component ablation and evidence quality on PASCAL VOC 2012. Results are reported in mIoU (%).

Group	Setting	xConv LSTM	Rich Text Anchors	Text Modulation	Raw CAM	Pseudo-Label	Seg.
	Progressive construction without CRF						
P	L1: Baseline				44.3	73.8	78.0
P	L2: + Rich text anchors		✓		46.8	73.9	78.0
P	L3: + Cross-layer xConvLSTM	✓			52.1	76.3	78.8
P	L4: + Text modulation	✓		✓	55.9	76.2	78.5
P	L5: Full model	✓	✓	✓	55.6	76.3	79.2
	Leave-one-out analysis with CRF						
L	Full model	✓	✓	✓	-	-	80.5
L	w/o Cross-layer xConvLSTM		✓	✓	-	-	79.6
L	w/o Rich text anchors	✓		✓	-	-	79.9
L	w/o Text modulation	✓	✓		-	-	80.3

The qualitative CAM visualizations in Fig. 4 show that the foreground regions become progressively more complete from L1 to L4, while irrelevant background responses are partially suppressed, consistent with the quantitative results.

4.4 Cross-Layer Aggregation Strategies

To determine whether the improvement originates from ordered recurrent aggregation rather than merely accessing multiple CLIP layers, we compare xConvLSTM with several alternative cross-layer strategies. This experiment is independent of the cumulative L1-L4 study above. All variants use the same selected CLIP layers, training protocol, CAM fusion coefficient, RFM topology, and

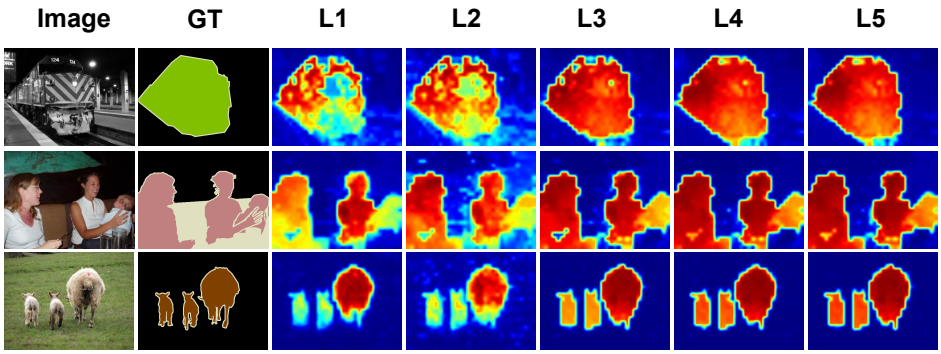


Figure 4: CAM comparison for the progressive ablation settings on PASCAL VOC 2012.

segmentation decoder. Rich concept text, anchored adaptive prompting, text modulation, and residual decoder feedback are disabled so that only the cross-layer aggregation mechanism is changed, as summarized in Table 3. Specifically, static averaging and learned weighted fusion yield only limited gains, indicating that simply combining multiple CLIP layers is insufficient. Although ConvLSTM improves raw CAM mIoU to 48.2%, its lower pseudo-label score suggests that conventional recurrence may accumulate unreliable responses. In contrast, full xConvLSTM achieves 52.1% raw CAM, 76.3% pseudo-label, and 78.8% segmentation mIoU, improving over the last-layer variant by 5.8, 1.8, and 1.7 points, respectively. It also exceeds the strongest alternative by 3.9, 1.8, and 0.9 points. These results demonstrate the benefit of ordered evidence accumulation, while the delta update prevents unreliable responses from overwriting the accumulated memory.

Table 3: Comparison of cross-layer aggregation strategies on the PASCAL VOC 2012. Results are reported in mIoU (%).

Aggregation method	Raw CAM	Pseudo-label	Seg.
Last layer only	46.3	74.5	77.1
Static average	47.5	74.1	77.5
Learned weighted fusion	47.7	74.5	77.2
ConvLSTM	48.2	73.1	77.9
xConvLSTM w/o delta update	48.1	74.1	77.2
Full xConvLSTM	**52.1**	**76.3**	**78.8**

5 Conclusion

In this work, we presented a text-modulated cross-layer xConvLSTM framework for weakly supervised semantic segmentation. By organizing selected CLIP layers as an ordered sequence, the proposed framework progressively connects local structural cues with deeper category semantics through concept-specific spatial memory. Text modulation guides the selection, retention, and suppression of visual evidence, while anchored adaptive prompting provides image-adaptive semantic conditions without substantially deviating from the frozen CLIP text space. The resulting memories enhance initial semantic responses, guide segmentation prediction, and complement the original RFM spatial propagation. Since the CLIP–DINO encoders and the original RFM topology remain frozen and unchanged, the improvements are obtained through a lightweight semantic-memory design rather than extensive backbone fine-tuning. Quantitative, qualitative, and ablation results demonstrate that explicitly organizing semantic evidence across network depth produces more complete activation regions, improves pseudo-label quality, and supports more accurate segmentation under image-level supervision. Overall, the proposed framework provides a simple and general strategy for combining semantic guidance with spatial evidence in weakly supervised segmentation.

AI USE STATEMENT

Generative AI tools were used only for language polishing, LaTeX formatting assistance, and debugging suggestions during the preparation of this manuscript and its implementation. All technical claims, experimental results, citations, code changes, and final text were reviewed and verified by the authors. The authors take full responsibility for the final content of this work, including all research conclusions and artifacts.

ETHICS STATEMENT

This work uses publicly available image datasets and does not involve human subjects, private personal data, user studies, or the collection of sensitive information. The proposed method is intended for research on weakly supervised semantic segmentation. We are not aware of any additional ethical risks beyond those inherent to the underlying public datasets.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we describe the datasets, data splits, training protocol, evaluation metrics, model configuration, and implementation details in the main paper and Appendix C. The appendix further provides the text templates, adaptive concept routing strategy, and the internal xConvLSTM update equations. The reported ablation settings specify the components enabled at each stage, and the segmentation and CAM evaluation procedures are described in the experimental section. Source code, configuration files, and trained models will be released in an anonymous repository during the review period or made publicly available upon publication.

REFERENCES

Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4981–4990. IEEE, 2018.

Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2204–2213. IEEE, 2019.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. *Advances in Neural Information Processing Systems*, 37:107547–107603, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Peng-Tao Jiang, Yuqi Yang, Qibin Hou, and Yunchao Wei. L2G: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European conference on computer vision*, pp. 695–711. Springer, 2016.

Hyeokjun Kweon and Kuk-Jin Yoon. From SAM to CAMs: Exploring segment anything model for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Minhyun Lee, Dongseob Kim, and Hyunjung Shim. Mining cross-image semantics for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15305–15314. IEEE, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Xianglin Qiu, Jian Wang, Xiaolei Wang, Zhen Zhang, and Jimin Xiao. Beyond text: Visual description assembly by probabilistic model for clip-based weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6346–6356, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 18061–18070. IEEE, 2022.

Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16825–16834. IEEE, 2022.

Lixiang Ru, Heliang Zheng, Yibing Zhan, and Bo Du. Token contrast for weakly-supervised semantic segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3093–3102. IEEE, 2023.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

Feilong Tang, Zhongxing Xu, Zhaojun Qu, Wei Feng, Xingjian Jiang, and Zongyuan Ge. Hunting attributes: Context prototype-aware learning for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 12272–12281. IEEE, 2020.

Yuanchen Wu, Xichen Ye, Kequan Yang, Jide Li, and Xiaoqiang Li. DuPL: Dual student with trustworthy progressive learning for robust weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4300–4309. IEEE, 2022.

Xiangfeng Xu, Pinyi Zhang, Wenxuan Huang, Yunhang Shen, Haosheng Chen, Jingzhong Lin, Wei Li, Gaoqi He, Jiao Xie, and Shaohui Lin. Weakly supervised semantic segmentation via progressive confidence region expansion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9829–9838. IEEE, 2025.

Zhiwei Yang, Kexue Fu, Minghong Duan, Linhao Qu, Shuo Wang, and Zhijian Song. Separate and conquer: Decoupling co-occurrence via decomposition and representation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Zhiwei Yang, Yucong Meng, Kexue Fu, Feilong Tang, Shuo Wang, and Zhijian Song. Exploring clip's dense knowledge for weakly supervised semantic segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20223–20232. IEEE, 2025a.

Ziqian Yang, Xinqiao Zhao, Xiaolei Wang, Quan Zhang, and Jimin Xiao. Ffr: Frequency feature rectification for weakly supervised semantic segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30261–30270. IEEE, 2025b.

Ziqian Yang, Xianglin Qiu, Xinqiao Zhao, Xiaolei Wang, Quan Zhang, and Jimin Xiao. Frequency-aware affinity for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20468–20477, 2026.

Bingfeng Zhang, Siyue Yu, Yunchao Wei, Yao Zhao, and Jimin Xiao. Frozen clip: A strong backbone for weakly supervised semantic segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3796–3806. IEEE, 2024.

Bingfeng Zhang, Siyue Yu, Jimin Xiao, Yunchao Wei, and Yao Zhao. Frozen clip-dino: A strong backbone for weakly supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):4198–4214, 2025.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, 2016.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European conference on computer vision*, pp. 696–712. Springer, 2022a.

Tianfei Zhou, Meijie Zhang, Fang Zhao, and Jianwu Li. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022b.

A APPENDIX

We organize the appendix as follows.

Appendix B presents additional method details, including:

- Appendix B.1: construction of text templates and the category-specific concept vocabulary;
- Appendix B.2: adaptive concept routing and its qualitative visualization;
- Appendix B.3: the internal update of the two-dimensional xLSTM spatial memory.

Appendix C provides detailed experimental settings, including:

- Appendix C.1: datasets and data splits;
- Appendix C.2: implementation and training details;
- Appendix C.3: the evaluation protocol and reported metrics.

Appendix D contains additional xConvLSTM analysis, including:

- Appendix D.1: selection of CLIP visual layers;
- Appendix D.2: layer-wise evolution of semantic evidence;
- Appendix D.3: qualitative cross-layer evidence accumulation;
- Appendix D.4: recurrent-gate behavior across visual depth;
- Appendix D.5: text modulation and semantic stability;
- Appendix D.6: cross-layer evolution of pixel representations;
- Appendix D.7: computational efficiency.

Appendix E discusses the limitations of the proposed framework and briefly outlines possible future extensions.

B ADDITIONAL METHOD DETAILS

This section provides additional design details omitted from the main paper, including the construction of text templates, adaptive concept routing, and the internal update of the two-dimensional xLSTM. These details supplement the compact descriptions in Section 3.2 and Section 3.3.

B.1 CONSTRUCTION OF TEXT TEMPLATES AND CONCEPT VOCABULARY

Equations equation 1 and equation 2 construct category and concept anchors using the frozen CLIP text encoder. This section provides the exact templates and category-specific concept vocabulary used in our PASCAL VOC experiments.

Category template. For category-level text, we use a single template:

```
a clean origami [category].
```

For example, the category *person* is represented as

```
a clean origami person.
```

Accordingly, the current implementation uses $M_c = 1$. We retain the general template-ensemble formulation in Equation equation 1 for compatibility with alternative category templates.

Category-conditioned concept templates. Each valid concept term is instantiated using the following four category-conditioned templates:

```
the [concept] of a clean origami [category].
a clean origami [category] with a visible [concept].
the local shape and appearance of the [concept] on a
[category].
the [concept], a distinctive component of a [category].
```

For example, the *head* concept of the *person* category produces

```
the head of a clean origami person.
a clean origami person with a visible head.
the local shape and appearance of the head on a person.
the head, a distinctive component of a person.
```

The four CLIP embeddings are averaged and L_2-normalized to form the frozen textual anchor associated with that concept.

Rich concept-bank construction. For each category, we first define a set of physical or structural parent concepts. Each parent concept is associated with its canonical name, visual attributes, structural descriptions, and commonly used aliases. The rich concept bank is constructed in a round-robin manner: we first include the canonical name of every parent concept and then progressively add its attributes, descriptions, and aliases until at most $P_{\max} = 12$ valid queries are obtained. This ordering prevents categories with many aliases from suppressing the coverage of other physical components.

Tables 4 report the exact valid concept terms used in our experiments. Categories with fewer than 12 valid terms are padded internally, while the padded slots are excluded from concept routing, memory updates, and auxiliary losses by a binary validity mask. At each selected CLIP layer, the adaptive router scores all valid terms and retains at most four concepts for the image-dependent memory update.

The entries in the tables define the frozen semantic anchors rather than spatial part annotations. No part masks, keypoints, or part-level supervision are used. Anchored adaptive prompting adds a bounded learnable residual and an image-conditioned residual to these anchors, while the original frozen representation remains available as a semantic reference. Consequently, the router can adapt concept relevance to the current image without freely replacing the predefined category semantics.

B.2 Adaptive Concept Routing

Not every fine-grained concept is visible in each image. The adaptive router therefore evaluates concept reliability using the layer-wise CLS token, parent-category text, and adaptive concept representation.

For concept $p_{k,j}$ at the selected visual layer l, the joint routing feature is defined as

$$\mathbf{u}_{k,j}^{l} = \tanh\left(W_g \mathbf{g}^{l} + W_t \mathbf{t}_k + W_r \mathbf{r}_{k,j}^{I,l}\right), \tag{B.1}$$

where W_g, W_t, and W_r project the three inputs into a shared routing space.

The router predicts the concept score and an additional null-concept score:

$$\begin{aligned} a_{k,j}^{l} &= W_s \mathbf{u}_{k,j}^{l}, \\ a_{k,\emptyset}^{l} &= W_\emptyset \tanh\left(W_g \mathbf{g}^{l} + W_t \mathbf{t}_k\right), \end{aligned} \tag{B.2}$$

where $a_{k,\emptyset}^{l}$ represents the absence of reliable fine-grained evidence for category c_k.

The routing probabilities are obtained by jointly normalizing the null node and all valid concepts:

$$\boldsymbol{\pi}_k^{l} = \mathrm{Softmax}\left(\frac{\left[a_{k,\emptyset}^{l}, a_{k,1}^{l}, \ldots, a_{k,P_k}^{l}\right]}{\tau_r}\right), \tag{B.3}$$

Table 4: Exact category-specific concept terms used for PASCAL VOC 2012. N_k denotes the number of valid queries for category k. Each listed term is instantiated using the four concept templates given above.

Category	N_k	Valid routed concept terms
aeroplane	12	fuselage; wing; tail; engine; landing gear; long streamlined central body; broad flat lateral wing; upright rear stabilizing fin; cylindrical engine under or beside a wing; small wheels below the fuselage; central body connecting the wings and tail; wing extending sideways from the fuselage.
bicycle	12	wheel; frame; handlebar; saddle; drivetrain; thin circular wheel with spokes; triangular metal frame connecting the wheels; narrow steering bar above the front wheel; small dark seat above the frame; toothed circular chain mechanism; front wheel; bicycle frame.
bird	12	head; beak; wing; tail; leg and foot; small rounded feathered head; pointed hard beak at the front of the head; broad feather-covered wing; fan-shaped group of tail feathers; thin legs with small grasping feet; eye; layered flight feathers on the side of the body.
boat	6	hull; deck; cabin; long curved body floating on water; flat upper surface above the hull; raised enclosed cabin with windows.
bottle	7	body; neck and cap; upright hollow container with a smooth surface; narrow neck and small top closure; bottle body; neck; cap.
bus	12	body; window; wheel; door; front; large elongated rectangular vehicle body; repeated row of large rectangular windows; large dark circular wheel below the body; tall rectangular passenger door; broad flat front with a large windshield; bus body; front side.
car	12	body; wheel; window; door; headlight; roof; compact curved vehicle body; dark circular rubber wheel; transparent dark glass window; rectangular side door integrated into the body; small bright lamp on the front; curved upper surface above the windows.
cat	12	head; ear; torso; leg and paw; tail; rounded furry head with two ears; small pointed triangular ear; soft fur-covered flexible body; slender furry leg ending in a rounded paw; long narrow furry tail; face; left ear.
chair	9	seat; backrest; leg; armrest; flat horizontal sitting surface; upright support behind the seat; narrow vertical support below the seat; horizontal side support beside the seat; chair leg.
cow	12	head; horn and ear; torso; leg and hoof; tail; large animal head with a broad muzzle; short horn or lateral ear above the head; large heavy four-legged torso; long sturdy leg ending in a hoof; thin rear tail with a tuft; muzzle; horn.
diningtable	7	tabletop; base and legs; large flat horizontal top surface; vertical supports below the tabletop; table edge; table leg; support frame.
dog	12	head; ear; torso; leg and paw; tail; furry head with a projecting muzzle; upright or hanging furry ear; fur-covered four-legged torso; furry leg ending in a compact paw; narrow furry tail extending from the rear; muzzle; left ear.
horse	12	head and neck; ear; torso; leg and hoof; tail; long head connected to an elongated neck; small upright pointed ear; large muscular four-legged torso; long slender leg ending in a hard hoof; long flowing hair tail; head; body.
motorbike	12	wheel; body and engine; handlebar; saddle; thick circular wheel below the frame; compact mechanical body between the wheels; steering bar above the front wheel; long narrow seat above the body; front wheel; engine; seat; back wheel.
person	12	head; torso; arm; hand; leg; foot; rounded head above the torso; central upright upper body; articulated upper limb attached to the torso; small hand at the end of an arm; long lower limb below the torso; small foot at the end of a leg.
pottedplant	7	plant; pot; cluster of green leaves and thin stems; solid container below the plant; leaf; flower pot; stem.
sheep	12	head; ear; torso; leg and hoof; tail; small animal head with a short muzzle; short lateral ear; rounded body covered with dense wool; thin leg below the woolly body; short rear tail; muzzle; body.
sofa	11	seat cushion; backrest; armrest; base and leg; broad soft horizontal cushion; wide padded upright back support; thick padded support on either side; low support beneath the cushions; seat; base; sofa leg.
train	12	coach; front; window; roof; wheel and undercarriage; long connected rectangular carriage; large leading face of the train; repeated row of rectangular windows; long upper surface above the coaches; dark mechanical structure below the coach; carriage; train front.
tvmonitor	7	screen; frame and stand; large flat rectangular display surface; thin border and support around or below the screen; display; frame; stand.

where P_k is the number of concepts associated with category c_k and τ_r is the routing temperature.

We retain the K highest-scoring non-null concepts and renormalize their weights. The resulting weights are further scaled by the non-null probability $1 - \pi^l_{k,\emptyset}$. Top-K routing is applied to concept-map aggregation, token writeback, and decoder feedback. All valid concepts still maintain and update their cross-layer xLSTM states.

Visualization of adaptive concept routing. To verify that concept routing performs image- and layer-dependent selection rather than using a fixed concept subset, we visualize the routing weights

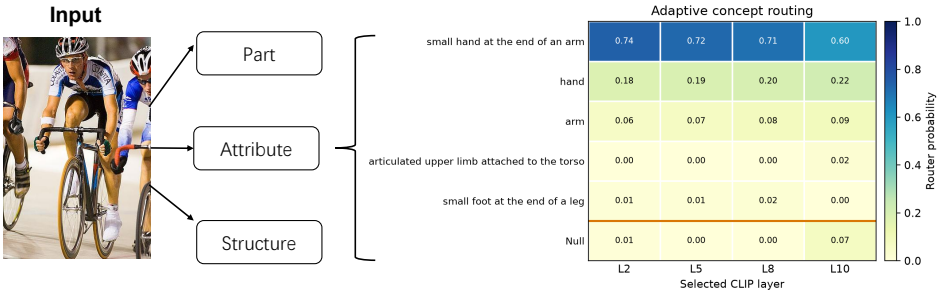


Figure 5: Image-adaptive concept routing across selected CLIP layers. Each row contains an input image and the routing weights for its labeled category. The heatmaps show the union of the top-four concepts selected at layers L2, L5, L8, and L10. Darker colors denote larger routing weights, and the *Null* row denotes the probability assigned to the null concept at each layer.

at the selected CLIP layers. For each image and its image-level category label, we collect the union of the top-four concepts selected at layers $\{2, 5, 8, 10\}$. Figure 5 shows their routing weights together with the null-concept probability. A darker cell represents a larger routing weight, while an empty cell indicates that the corresponding concept is not selected among the top-four concepts at that layer.

The visualization reveals that concept selection changes as visual evidence evolves across depth. Earlier layers may assign larger weights to locally supported appearance or structural concepts, whereas later layers can emphasize concepts that are more consistent with the category-level context. Different images of the same category also produce different routing distributions according to the visible object regions. Concepts with weak support are suppressed by top-K routing or absorbed into the null branch, preventing all textual concepts from contributing equally to the cross-layer memory.

B.3 Two-Dimensional xLSTM Spatial Memory

Equation equation 8 presents the cross-layer xLSTM as a compact state transition. This section specifies its internal update. For clarity, the category-concept indices (k, j) are omitted below.

The current semantic observation and previous hidden state are first concatenated. A depth-wise 3×3 convolution captures local spatial context, while a 1×1 convolution and a textual projection produce the recurrent gates:

$$\begin{aligned}&\left[\widehat{\mathbf{i}}^l, \widehat{\mathbf{f}}^l, \widehat{\mathbf{o}}^l, \widetilde{\mathbf{C}}^l, \widehat{\boldsymbol{\eta}}^l\right] \\ &\quad = \mathrm{Conv}_{1\times1}\left(\mathrm{DWConv}_{3\times3}\left[\mathbf{Z}^l \| \mathbf{H}^{l-1}\right]\right) + B\left(\mathbf{r}^{I,l}\right),\end{aligned} \tag{B.4}$$

where $\|$ denotes channel-wise concatenation and $B(\cdot)$ maps the adaptive concept representation to text-conditioned gate biases. The five outputs correspond to the raw input gate, forget gate, output gate, candidate memory, and delta gate, respectively.

The stabilizer state is used to control the numerical range of the exponential input and forget gates:

$$\begin{aligned}\mathbf{m}^l &= \max\left(\widehat{\mathbf{i}}^l, \log\sigma\left(\widehat{\mathbf{f}}^l\right) + \mathbf{m}^{l-1}\right), \\ \mathbf{i}^l &= \exp\left(\widehat{\mathbf{i}}^l - \mathbf{m}^l\right), \\ \mathbf{f}^l &= \exp\left(\log\sigma\left(\widehat{\mathbf{f}}^l\right) + \mathbf{m}^{l-1} - \mathbf{m}^l\right).\end{aligned} \tag{B.5}$$

For the first selected layer, the stabilizer is initialized from the current input response because no valid memory history is available.

A delta gate controls how far the normalized previous memory moves toward the current candidate:

$$\begin{aligned} \overline{\mathbf{C}}^{l-1} &= \frac{\mathbf{C}^{l-1}}{\mathbf{N}^{l-1}+\epsilon}, \\ \boldsymbol{\eta}^{l} &= \sigma\left(\widehat{\boldsymbol{\eta}}^{l}\right), \\ \mathbf{V}^{l} &= \overline{\mathbf{C}}^{l-1} + \boldsymbol{\eta}^{l} \odot \left[\tanh\left(\widetilde{\mathbf{C}}^{l}\right) - \overline{\mathbf{C}}^{l-1}\right], \end{aligned} \tag{B.6}$$

where ϵ is a small numerical constant and $\mathbf{V}^l$ is the value written into memory. At the first selected layer, the candidate memory is used directly for initialization.

Finally, the accumulated memory, normalization state, and hidden state are updated as

$$\begin{aligned} \mathbf{C}^{l} &= \mathbf{f}^{l} \odot \mathbf{C}^{l-1} + \mathbf{i}^{l} \odot \mathbf{V}^{l}, \\ \mathbf{N}^{l} &= \mathbf{f}^{l} \odot \mathbf{N}^{l-1} + \mathbf{i}^{l}, \\ \mathbf{H}^{l} &= \sigma\left(\widehat{\mathbf{o}}^{l}\right) \odot \frac{\mathbf{C}^{l}}{\mathbf{N}^{l}+\epsilon}. \end{aligned} \tag{B.7}$$

The accumulated state $\mathbf{C}^l$ preserves cross-layer evidence, $\mathbf{N}^l$ normalizes the memory magnitude, and $\mathbf{m}^l$ stabilizes the exponential gates. The hidden state $\mathbf{H}^l$ provides the concept-specific spatial representation at the current visual depth.

The spatial convolution preserves local structures, whereas textual conditioning determines which evidence should be written, retained, or suppressed. All category-concept pairs share the xLSTM parameters but maintain independent spatial states.

C Detailed Experimental Settings

C.1 Datasets and Data Splits

PASCAL VOC 2012 contains 20 foreground categories and one background category. Following the standard WSSS protocol, we use the augmented training set containing 10,582 images. The validation and test sets contain 1,449 and 1,456 images, respectively.

MS COCO 2014 contains 80 foreground categories and includes more complex scenes, frequent category co-occurrence, small objects, and large variations in object scale. We train the model on the COCO training set and report results on its validation set. No pixel-level annotations, bounding boxes, saliency maps, or part masks are used during training.

C.2 Implementation Details

Our model is built upon WeCLIP and retains its original RFM propagation topology. A frozen CLIP ViT-B/16 is used to provide category-semantic features, while a frozen DINOv2 ViT-S/14 provides spatially consistent visual features. The outputs of the 2nd, 5th, 8th, and 10th CLIP visual layers are arranged according to network depth and processed by the cross-layer memory module.

Each category contains at most 12 fine-grained concepts, from which the four concepts with the highest routing weights are retained. The hidden dimension of each concept state is set to 8. A 3×3 convolution is used for local gated mapping, followed by a lightweight 1×1 convolutional readout head.

During training, images are randomly rescaled by a factor between 0.5 and 2.0 and cropped to 320×320. We use AdamW with an initial learning rate of 2×10^{-5} and a weight decay of 0.01. The models are trained for 30,000 iterations on PASCAL VOC and 80,000 iterations on MS COCO. Horizontal flipping and multi-scale inference are used during evaluation.

C.3 Evaluation Protocol

Mean intersection over union (mIoU) is used to evaluate the overlap between the predicted and ground-truth regions. For class c, the intersection over union is defined as

Table 5: Sensitivity to the choice of CLIP visual layers on the PASCAL VOC 2012 validation set. All settings use the same model, training schedule, and checkpoint budget. The reported values are computed before CRF.

Layer selection	Block indices	Raw CAM	Pseudo-label	Seg. mIoU
Six-layer selection	$\{1, 2, 4, 6, 8, 10\}$	57.9	76.3	78.1
All-layer selection	$\{0, \ldots, 11\}$	**61.8**	**76.4**	77.4
Current selection	$\{2, 5, 8, 10\}$	55.6	76.3	**79.2**

$$\mathrm{IoU}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}, \tag{C.1}$$

where TP_c, FP_c, and FN_c denote the numbers of true-positive, false-positive, and false-negative pixels, respectively. The final mIoU is obtained by averaging the IoU values over all evaluated categories.

We report three types of results. *Raw CAM* measures the initial semantic response before RFM propagation; *Pseudo-label* measures the response refined by RFM/PAR; and *Seg.* denotes the final segmentation prediction. These metrics allow us to separately analyze the quality of the initial semantic evidence, spatial propagation, and final dense prediction.

In the comparison tables, "I" denotes image-level category supervision, and "L" denotes linguistic priors obtained from vision-language pre-training. The linguistic priors contain no pixel-level locations or part annotations.

D Additional xConvLSTM Analysis

This section provides additional analysis of cross-layer xConvLSTM. We first evaluate the selection of CLIP visual layers and then examine how semantic evidence evolves across the selected layers. We further analyze qualitative memory accumulation, recurrent-gate behavior, text modulation, pixel-feature separability, and computational efficiency.

D.1 Selection of CLIP Visual Layers

For the CLIP ViT-B/16 visual encoder, the transformer blocks are indexed from 0 to 11. We select the block indices $\mathcal{L} = \{2, 5, 8, 10\}$, corresponding to the 3rd, 6th, 9th, and 11th visual blocks, respectively. The selection is designed to cover different stages of representation formation while avoiding excessive sampling of neighboring blocks.

The selected sequence spans early, intermediate, and late visual representations. The early selected block retains local appearance, contours, and object-part structures, whereas the intermediate blocks progressively incorporate category-level context. The late selected block provides broader category-related responses. Arranging these representations according to network depth gives the recurrent memory an ordered semantic trajectory:

local structure $\rightarrow$ part-level evidence $\rightarrow$ category-aware localization $\rightarrow$ high-level semantic context.

The cross-layer memory can therefore preserve useful intermediate evidence while selectively revising it using representations from deeper layers.

Table 5 shows that using more CLIP layers does not necessarily improve the final segmentation performance. Although the all-layer selection obtains the highest raw CAM and pseudo-label mIoU, its segmentation mIoU decreases to 77.4%, suggesting that densely sampled neighboring layers may introduce redundant or noisy evidence. In contrast, the current selection achieves the best segmentation mIoU of 79.2%, outperforming the six-layer and all-layer selections by 1.1% and 1.8%, respectively. These results support using a compact set of complementary layers for cross-layer evidence accumulation.

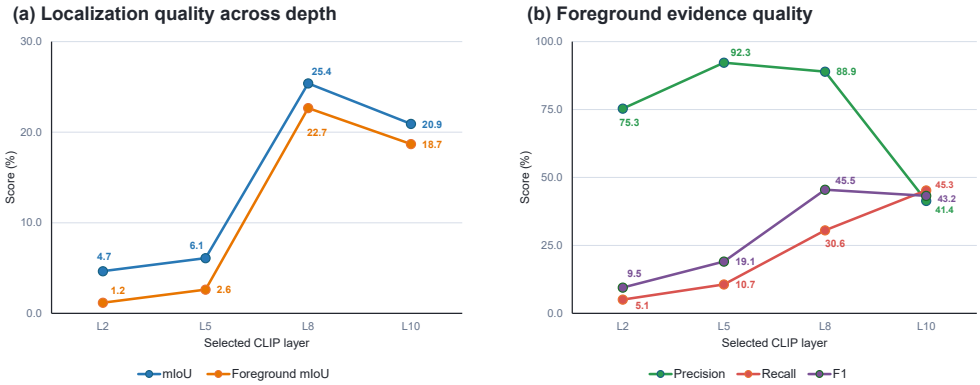


Figure 6: Layer-wise evolution of semantic evidence on the PASCAL VOC 2012 validation set. **(a)** CAM mIoU and foreground IoU measure localization quality. **(b)** Foreground precision, recall, and F1 measure the reliability and completeness of the responses. The intermediate block provides the strongest localization and F1, while the deepest block increases recall at the cost of precision.

D.2 Layer-Wise Evolution of Semantic Evidence

We next examine the class-specific responses produced by the selected CLIP visual blocks to understand how localization evidence changes with network depth. In addition to CAM mIoU, we report foreground IoU, precision, recall, and F1 score. The selected zero-based block indices $\{2, 5, 8, 10\}$ correspond to the 3rd, 6th, 9th, and 11th visual blocks, respectively.

Figure 6 shows that semantic evidence does not evolve monotonically across CLIP depth. Localization quality improves substantially from the early blocks and reaches its highest level at the 9th visual block, where CAM mIoU, foreground IoU, and F1 reach 25.4%, 22.7%, and 45.5%, respectively. This result indicates that intermediate representations provide the most spatially concentrated category evidence.

At the 11th visual block, recall increases from 30.6% to 45.3%, whereas precision decreases from 88.9% to 41.4%. Consequently, CAM mIoU and F1 decrease to 20.9% and 43.2%, respectively. The deepest representation therefore covers more category-related regions but also introduces additional contextual or background responses. Together, these results show that the selected layers provide complementary localization characteristics and motivate preserving and selectively revising intermediate evidence through cross-layer memory.

D.3 Qualitative Cross-Layer Evidence Accumulation

Figure 7 compares the independent class-specific responses extracted from the selected CLIP visual blocks with the spatial states produced by the cross-layer xConvLSTM. The ground-truth masks are shown only as visual references and are not used as model inputs during training or inference.

The visualization shows that the independently extracted CLIP responses change substantially across visual depth. Early responses tend to preserve local structures and object-part details, but they may be fragmented or incomplete. Deeper responses provide stronger category semantics and broader object coverage, while also introducing more contextual or background activations. This depth-dependent variation indicates that no single visual block consistently provides the most complete evidence.

In contrast, the xConvLSTM states progressively integrate information from the ordered layer sequence. Spatially concentrated evidence from intermediate blocks can be retained when later layers become more diffuse, while deeper semantic responses can supplement regions that are weakly activated at earlier layers. This behavior is particularly clear for small or partially occluded objects, whose evidence may be incomplete in an individual CLIP response but becomes more coherent after

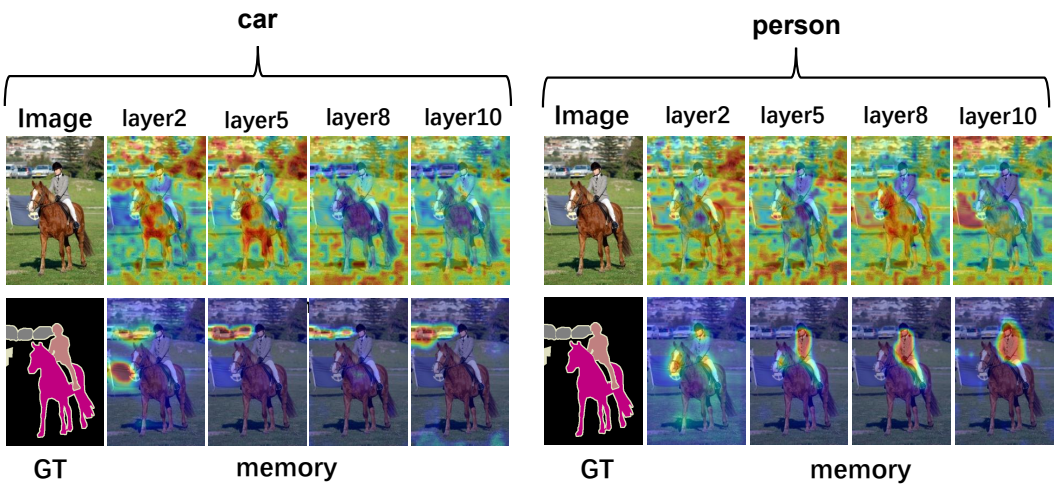


Figure 7: Qualitative evolution of class-specific evidence across CLIP depth. The figure compares independent CLIP responses with the corresponding xConvLSTM memory states.

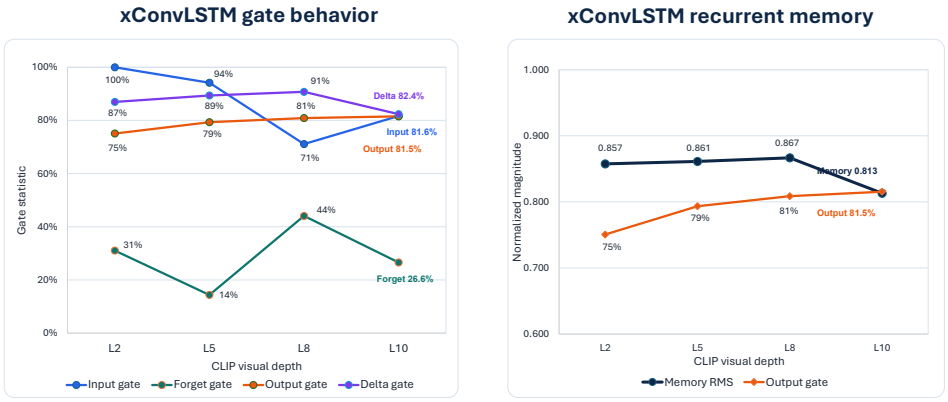


Figure 8: Mean recurrent-gate responses and memory magnitude across the selected CLIP visual blocks.

recurrent updates. The memory states therefore preserve useful localization cues while incorporating the category-level context of deeper representations.

These observations are consistent with the quantitative layer-wise results in Fig. 6. Rather than replacing all intermediate evidence with the deepest representation, the proposed cross-layer memory maintains an evolving spatial record and selectively revises it at each visual depth. This provides a qualitative explanation for why ordered evidence accumulation improves the completeness and spatial consistency of the final representation.

D.4 Gate Behavior Across Visual Depth

To examine how xConvLSTM updates and preserves spatial evidence across visual depth, we report the mean input, forget, output, and delta gates at the selected CLIP visual blocks. Statistics are averaged over the PASCAL VOC 2012 validation set, active categories, and routed concepts.

Figure 8 shows that the recurrent gates vary substantially across visual depth, indicating that the xConvLSTM does not apply the same update rule at every layer. The input gate is highest at L2, reaching 100%, and decreases at the intermediate layers before partially recovering at L10. This

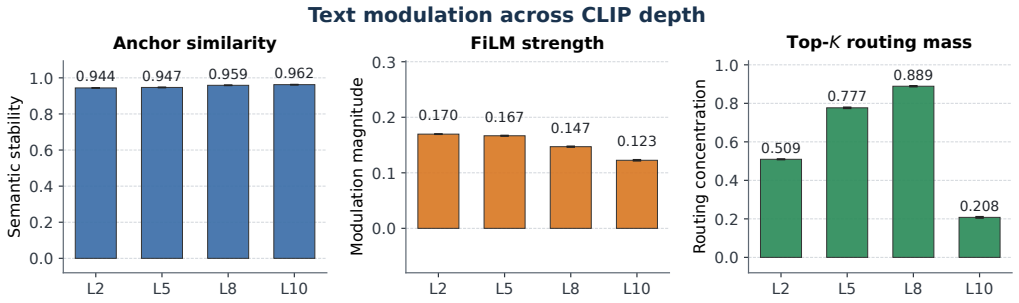


Figure 9: Bar-chart visualization of text-modulation statistics across the selected CLIP layers. Anchor similarity measures the semantic stability of the adaptive concept representation relative to its frozen text anchor. FiLM strength measures the magnitude of the visual-channel modulation, while Top-K routing mass measures the concentration of image-adaptive concept selection.

suggests that the memory relies more strongly on the current visual observation at the beginning of the sequence, while later updates increasingly incorporate previously accumulated evidence.

The forget gate changes non-monotonically, decreasing from 31% at L2 to 14% at L5, increasing to 44% at L8, and then decreasing to 26.6% at L10. Its relatively large response at L8 indicates stronger retention of previously accumulated evidence at the intermediate visual depth. The output gate gradually increases from 75% to approximately 81.5%, showing that the recurrent state remains available for spatial readout throughout the layer sequence.

The delta gate reaches its highest value at L8, increasing from 87% at L2 to 91%, before decreasing to 82.4% at L10. Thus, the intermediate layer performs a relatively strong adjustment of the normalized memory toward the current candidate, whereas the deepest layer applies a more conservative update. This behavior is consistent with the role of intermediate representations in providing spatially concentrated localization evidence.

The recurrent-memory magnitude remains relatively stable from L2 to L8, increasing slightly from 0.857 to 0.867, and then decreases to 0.813 at L10. The memory therefore does not simply accumulate activation magnitude with depth. Instead, it preserves a stable representation through the intermediate layers and selectively normalizes or suppresses part of the accumulated response at the deepest layer. Overall, the gate statistics indicate that xConvLSTM balances new evidence, memory retention, and state readout differently across depth, allowing useful intermediate localization cues to be retained while deeper semantic evidence is incorporated.

D.5 Text Modulation and Semantic Stability

We analyze text modulation using three statistics. First, we measure the cosine similarity between the adaptive concept representation and its frozen text anchor to quantify semantic stability:

$$s^l_{\text{anchor}} = \cos\left(\mathbf{r}^{I,l}_{k,j}, \mathbf{r}^{0}_{k,j}\right). \tag{D.1}$$

Second, we measure the strength of the FiLM modulation applied to the visual feature:

$$s^l_{\text{FiLM}} = \frac{1}{D}\left(\left\|\boldsymbol{\gamma}^l_{k,j} - \mathbf{1}\right\|_2 + \left\|\boldsymbol{\beta}^l_{k,j}\right\|_2\right), \tag{D.2}$$

where $\boldsymbol{\gamma}^l_{k,j}$ and $\boldsymbol{\beta}^l_{k,j}$ denote the concept-conditioned channel scale and shift. Finally, we measure the total routing mass assigned to the selected Top-K concepts:

$$s^l_{\text{route}} = \sum_{j \in \mathcal{K}_k} \pi^l_{k,j}. \tag{D.3}$$

All statistics are averaged over validation images, active categories, and routed concepts at the selected CLIP layers $L2$, $L5$, $L8$, and $L10$. Ground-truth masks are not used for routing, memory updating, or statistical computation. Fig. 9 visualizes the three statistics with bar charts, providing

a compact view of how text conditions the visual evidence before cross-layer memory updates. The statistics reveal a depth-dependent modulation pattern. Anchor similarity increases from 0.944 at L2 to 0.962 at L10, indicating that the adaptive concepts remain close to their frozen semantic anchors and become increasingly stable at deeper layers. In contrast, FiLM strength decreases from 0.170 to 0.123, suggesting that stronger visual modulation is mainly required at earlier layers, while deeper features already provide more reliable semantic evidence. The routing mass increases from 0.509 at L2 to 0.889 at L8, showing that the selected concepts become more concentrated during intermediate-layer aggregation. Its decrease to 0.208 at L10 indicates a more distributed use of concepts in the final layer, where broader semantic evidence may be integrated. Overall, these results support a complementary role of text modulation across depth: early layers receive stronger adaptive conditioning, intermediate layers form concentrated concept evidence, and late layers preserve stable semantics while integrating broader context.

D.6 Cross-Layer Evolution of Pixel Representations

We further investigate how pixel-level representations evolve across CLIP depth and how this evolution is affected by the proposed cross-layer semantic memory. Instead of comparing only the final decoder features, we visualize representations at the selected CLIP layers $\mathcal{L} = \{2, 5, 8, 10\}$. The first row of Fig. 10 shows the original CLIP visual tokens, while the second row shows the corresponding representations after text-modulated cross-layer memory accumulation.

For a fair comparison, all panels use exactly the same PASCAL VOC 2012 validation images and pixel coordinates. We sample pixels uniformly from the foreground categories and background, with at most 100 pixels per category from each image. Ground-truth masks are used only to assign labels to sampled validation pixels and are never provided as model inputs.

All representations are ℓ_2-normalized before dimensionality reduction. To preserve comparability across panels, the features from all layers and both representation types are jointly projected using one shared PCA and t-SNE transformation. We first reduce the features to 50 dimensions using PCA and then apply t-SNE with a perplexity of 50, PCA initialization, 2,000 optimization iterations, and a fixed random seed of 2027. The same category colors and plotting ranges are used in every panel.

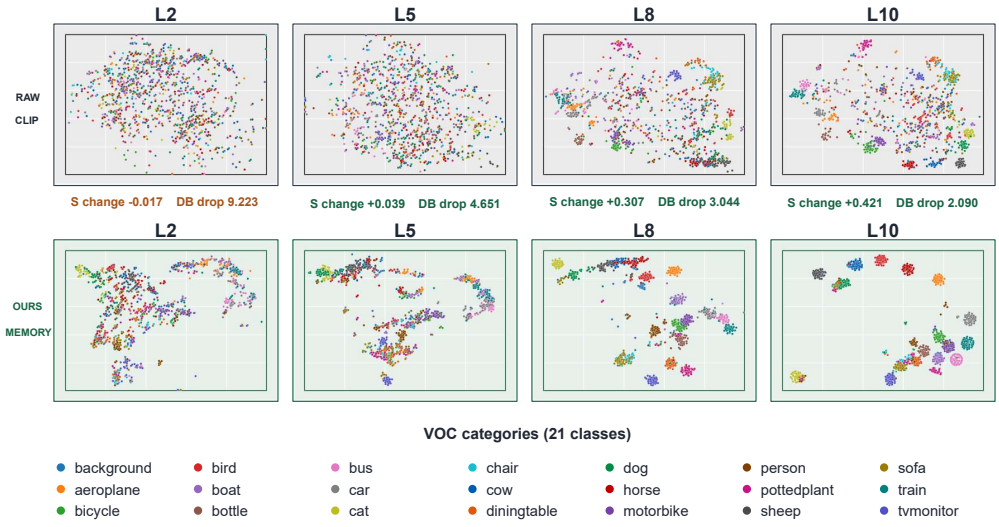


Figure 10: Cross-layer evolution of pixel-level representations on the PASCAL VOC 2012 validation set. The top row shows the original CLIP pixel features extracted from layers L2, L5, L8, and L10. The bottom row shows the corresponding representations after text-modulated cross-layer memory accumulation.

As shown in Fig. 10, the original CLIP representations become more semantically discriminative with depth, but still contain considerable inter-class overlap. The proposed memory progressively reorganizes these representations by accumulating complementary evidence across layers. This trend

Table 6: Pixel-feature separability across selected CLIP layers. Higher Silhouette scores and lower Davies–Bouldin indices indicate better feature separability. Metrics are computed in the original normalized feature space rather than in the two-dimensional t-SNE space.

	L2		L5		L8		L10	
Representation	Sil.↑	DB↓	Sil.↑	DB↓	Sil.↑	DB↓	Sil.↑	DB↓
Original CLIP	-0.055	15.021	-0.053	9.353	-0.007	4.967	0.058	3.716
Ours	-0.072	5.798	-0.015	4.702	0.301	1.923	0.480	1.626

Table 7: Computational comparison of different cross-layer aggregation modules. FLOPs count the approximate operations of convolutional and linear layers. Latency and FPS are measured with a batch size of 1 and an input resolution of 320×320 on the same GPU.

Method	Params (M)	FLOPs (G)	Memory (GB)	Latency (ms)	FPS
Baseline	13.49	69.71	1.24	195.70	5.11
ConvLSTM	15.47	70.92	1.26	262.81	3.81
xConvLSTM	15.48	71.06	1.26	265.69	3.76

is quantitatively confirmed by Table 6. At L2, the Davies–Bouldin index decreases from 15.021 to 5.798, although the Silhouette score changes only slightly from -0.055 to -0.072, suggesting that the early memory update mainly suppresses broad inter-class confusion. At L5, the Silhouette score improves from -0.053 to -0.015 and the Davies–Bouldin index decreases from 9.353 to 4.702. The improvement becomes more pronounced at L8 and L10: the Silhouette score increases from -0.007 to 0.301 and from 0.058 to 0.480, while the Davies–Bouldin index decreases from 4.967 to 1.923 and from 3.716 to 1.626, respectively. These results indicate that the benefit of cross-layer memory accumulates gradually and becomes strongest in the deeper layers, where local structural evidence has been combined with more reliable category semantics. Thus, the proposed method improves the feature space throughout the network rather than merely altering the final decoder representation.

D.7 Computational Efficiency

We measure trainable parameters, approximate convolutional and linear FLOPs, peak GPU memory, inference latency, and throughput. The baseline, standard ConvLSTM, and xConvLSTM variants are evaluated using the same input resolution, batch size, device, and inference protocol.

Table 7 compares the computational cost of the different cross-layer aggregation mechanisms. Introducing ConvLSTM increases the trainable parameters from 13.49M to 15.47M and the approximate FLOPs from 69.71G to 70.92G. The proposed xConvLSTM requires 15.48M parameters and 71.06G FLOPs, indicating that its additional memory stabilization operations introduce only a negligible overhead in model size and computational complexity.

Compared with the baseline, xConvLSTM adds approximately 1.99M parameters and 1.35G FLOPs, while increasing the latency from 195.70ms to 265.69ms. The corresponding throughput decreases from 5.11 FPS to 3.76 FPS. This overhead is mainly caused by the sequential processing of visual features across the selected CLIP layers and the additional state-update operations.

More importantly, the comparison with standard ConvLSTM isolates the cost of the proposed memory update. xConvLSTM adds only approximately 0.01M parameters and 0.14G FLOPs over ConvLSTM, with an additional latency of 2.88ms and a throughput decrease of only 0.05 FPS. Peak GPU memory remains unchanged at 1.26GB for both recurrent variants. Thus, xConvLSTM preserves nearly the same computational profile as ConvLSTM while providing a more stable mechanism for accumulating semantic evidence across network depth. The measured overhead is therefore small relative to the representational benefit observed in the cross-layer aggregation experiments.

E LIMITATIONS AND FUTURE EXTENSIONS

The proposed framework remain has several limitations. Since the model is trained with image-level supervision, the generated CAMs and pseudo-labels may contain incomplete object regions, inaccurate boundaries, or responses from visually similar background regions. Moreover, the performance depends on the selected CLIP layers, text templates, and routed concepts, which are currently determined using a fixed configuration. The cross-layer memory and multi-scale inference also introduce additional computational overhead compared with a single-layer baseline.

Future work will explore adaptive layer selection and category-aware concept routing so that the model can dynamically select the most informative visual and textual evidence. We will also investigate learnable prompts, lightweight memory updates, and more efficient boundary-refinement strategies. Extending the evaluation to fine-grained and cross-domain datasets is another important direction for assessing the generalization ability of the proposed method.