UNICONVSEG: QUERY-FREE CONVOLUTIONAL UNIFIED SEGMENTATION IN IMAGES AND VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified segmentation models commonly rely on Transformer object queries and repeated global query–pixel interactions, raising the question of whether one efficient model can cover image, prompt, and video segmentation using convolution alone. We present UniConvSeg, a purely convolutional, query-free architecture that shares a ConvNeXt V2 feature pyramid and combines dense object-conditioned kernels, a convolutional panoptic readout, a prompt-conditioned mask decoder, bounded convolutional memory, and dataset-conditioned vocabulary selection. Training proceeds through task-specific prompt/VOS and panoptic stages followed by joint optimization over panoptic, prompted, VOS, and video-panoptic data with sampling probabilities $(0.40, 0.20, 0.20, 0.20)$. With 122M parameters, UniConvSeg reaches 130 FPS for image-panoptic inference on an A800, $5.9\times$ the 22 FPS of OMG-Seg under the same batch-1 end-to-end protocol. It achieves 52.7 COCO PQ, 43.2 instance AP, 48.3 VIPSeg VPQ, 53.2 YT-VIS AP, 88.7 DAVIS $\mathcal{J}\&\mathcal{F}$, and 68.1 promptable-segmentation mIoU. These results, together with controlled readout, memory, training, and association ablations, show that purely convolutional task conditioning can provide a strong accuracy–efficiency tradeoff across four segmentation interfaces without a Transformer query decoder.

1 INTRODUCTION

A unified segmentation model must discover objects, distinguish instances, respond to prompts, and preserve identities over time. These interfaces differ in what they are given: image panoptic segmentation starts from pixels alone, prompted segmentation receives a spatial cue, video object segmentation (VOS) receives an initial identity, and video panoptic segmentation must both discover and associate objects. Their common visual structure motivates one shared model, but does not imply an identical output rule for every task.

Query-based systems provide an effective abstraction for these requirements. Mask2Former represents masks through learned queries and masked attention; TarViS and OMG-Seg broaden unification across video and heterogeneous datasets (Cheng et al., 2022; Athar et al., 2023; Li et al., 2024). We ask a complementary question: *which parts of task unification require global query interaction, and which can be implemented with dense convolutional representations and explicit conditioning?* Conditional convolution already provides instance-specific masks (Tian et al., 2020). The challenge is to integrate this representation with semantic labeling, user prompts, and temporal state without obscuring their different dependencies.

UniConvSeg shares a ConvNeXt V2 encoder and a feature pyramid across four task families. Dense locations predict categories, object geometry, and small mask-generating kernels. A convolutional fusion module reconciles thing masks with stuff predictions. A separate prompt decoder reads the same visual features, and bounded convolutional memory extends it to VOS. Video panoptic inference reuses the panoptic readout with mask association. We use *query-free* to mean the absence of a Transformer object-query decoder; channel gates and learned temporal weighting remain. The sharing occurs in visual representation and optimization, while the readouts preserve task-specific conditioning.

This distinction determines what a meaningful ablation can show. Removing a panoptic fusion module at inference cannot directly change VOS, because VOS does not execute that module. Retraining shared features can change both. Similarly, a text classifier can influence category recognition while

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.

Method	SS	IS	PS	VSS	VIS	VPS	VOS	Multi-DS	Interactive	Shared	Attn.-free
multicolumn12l*Task specialists*											
DeepLabV3+ (Chen et al., 2018)	✓										✓
Mask R-CNN (He et al., 2017)		✓									✓
Panoptic FPN (Kirillov et al., 2019)			✓								✓
DETR (Carion et al., 2020)			✓								
VisTR (Wang et al., 2021)					✓						
VPSNet (Kim et al., 2020)						✓					
STM (Oh et al., 2019)							✓				
Unified image or video models											
K-Net (Zhang et al., 2021)	✓	✓	✓								
Mask2Former (Cheng et al., 2022)	✓	✓	✓								
Video K-Net (Li et al., 2022)				✓	✓	✓					
Tube-Link (Li et al., 2023b)				✓	✓	✓					
OneFormer (Jain et al., 2023)	✓	✓	✓							✓	
TarViS (Athar et al., 2023)				✓	✓	✓	✓			✓	
Open-vocabulary and multi-dataset models											
MSeg (Lambert et al., 2020)	✓							✓			✓
UNINEXT (Yan et al., 2023)		✓			✓		✓	✓			
SAM (Kirillov et al., 2023)								✓	✓		
SAM 2 (Ravi et al., 2025)							✓	✓	✓	✓	
EfficientTAM (Xiong et al., 2025)							✓	✓	✓	✓	
Efficient-SAM2 (Zhang et al., 2026)							✓	✓	✓	✓	
Semantic-SAM (Li et al., 2023a)	✓	✓	✓					✓	✓	✓	
SEEM (Zou et al., 2023b)	✓	✓	✓					✓	✓	✓	
FreeSeg (Qin et al., 2023)	✓	✓	✓							✓	
OMG-Seg (Li et al., 2024)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
UniConvSeg (ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison of segmentation settings. SS/IS/PS denote image semantic/instance/panoptic segmentation; VSS/VIS/VPS are their video counterparts. Multi-DS denotes multi-dataset training, and Shared denotes one parameter set. Checkmarks describe task coverage; the four interfaces studied here are image panoptic, prompting, VOS, and video panoptic segmentation.

leaving a ground-truth-box-conditioned mask decoder unchanged. We therefore organize the study around the computational paths affected by each intervention, rather than attributing every score change to a component used by only one branch.

A second issue arises when datasets share concepts but disagree about annotation slots. A merged vocabulary can contain a thing and a stuff slot encoded by the same text vector. A fixed cosine classifier cannot distinguish identical prototypes, and selecting a label before dataset export can suppress a valid output. We characterize this failure and use dataset-conditioned selection. This analysis links the architecture to the output space: shared features alone do not guarantee compatible predictions.

The contributions are: (i) a four-task convolutional architecture with explicit sharing and conditioning boundaries; (ii) an operator-level cost analysis and a characterization of prototype collisions; and (iii) controlled experiments that separate branch-local interventions, temporal capacity, and training-induced transfer. Tables 1 and 2 summarize the comparison setting and measured performance. Appendix C details the additional controlled comparisons.

2 Related Work

Unified segmentation. MaskFormer and Mask2Former formulate segmentation through mask classification (Cheng et al., 2021; 2022). OneFormer adds task conditioning for image segmentation, while TarViS and OMG-Seg address broader video and multi-dataset settings (Jain et al., 2023; Athar et al., 2023; Li et al., 2024). UniConvSeg follows the goal of one parameter set serving several task interfaces, but uses dense spatial locations instead of an object-query decoder. We do not equate supporting four interfaces with matching the complete benchmark coverage of these systems.

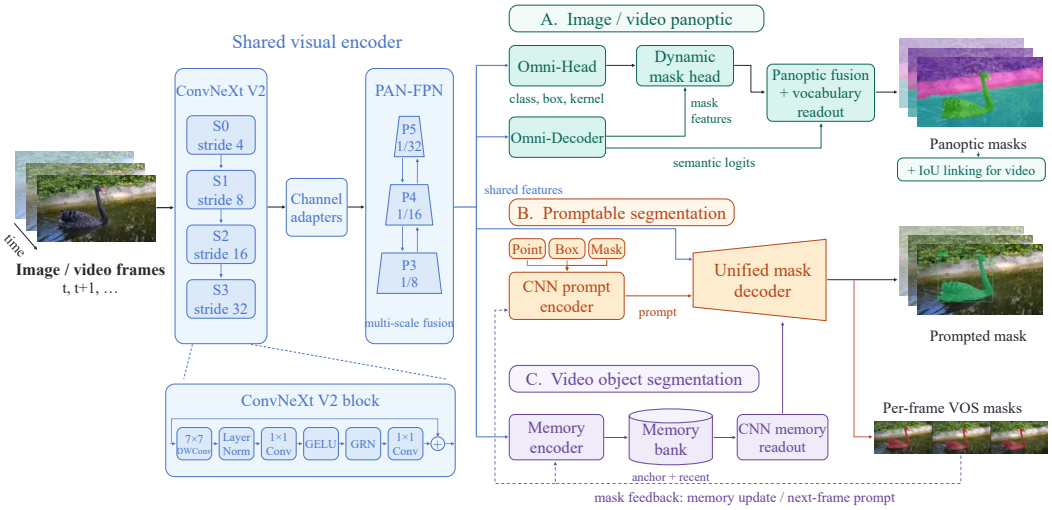

Figure 1: Architecture of UniConvSeg. The shared ConvNeXt V2 encoder and PAN-FPN feed (A) the image/video panoptic path, (B) the prompt-conditioned mask decoder, and (C) convolutional memory for VOS. OmniHead supplies instance kernels; OmniDecoder supplies mask features and semantic logits. Video panoptic inference adds IoU linking. Dashed feedback updates memory and the next-frame prompt; the inset shows a ConvNeXt V2 block. Output images illustrate the task interfaces.

Dense masks and convolutional representations. FCOS predicts objects at feature-grid locations (Tian et al., 2019). YOLACT represents masks through shared prototypes, and CondInst uses dynamically generated small convolutional networks (Bolya et al., 2019; Tian et al., 2020). ConvNeXt V2 supplies the convolutional backbone (Woo et al., 2023). Our instance head follows these established ideas; the contribution is not a new conditional-convolution primitive. We examine their compatibility with prompted decoding, memory, and a merged semantic space.

Prompting, memory, and text classification. SAM and SAM 2 establish prompt-conditioned image and video segmentation (Kirillov et al., 2023; Ravi et al., 2025); XMem studies memory for long videos (Cheng & Schwing, 2022). Our memory uses aligned-grid correlation and convolution rather than global space-time retrieval, imposing a different correspondence bias. CLIP and convolutional open-vocabulary segmentation motivate text-prototype classifiers (Radford et al., 2021; Yu et al., 2023). A text interface alone, however, does not establish unseen-category recognition. Our vocabulary experiment evaluates label compatibility within a known merged space; it is distinct from an open-vocabulary generalization benchmark.

3 Query-free architecture

3.1 Shared representation and dense panoptic readout

Given $x \in \mathbb{R}^{3\times H\times W}$, a ConvNeXt V2 backbone produces four scales. Channel adapters map them to widths $(160, 320, 640, 640)$ at strides $(4, 8, 16, 32)$; a PAN-FPN processes the last three. These adapters allow decoder interfaces to remain fixed across backbone sizes. Figure 1 distinguishes feature sharing from prediction-head sharing.

At each P3/P4/P5 location, the dense head predicts thing logits, objectness o, centerness c, nonnegative LTRB distances, a quality logit q, and a controller vector θ. Following FCOS, positive locations are selected using scale ranges and center sampling. In text mode the classification logit is

$$z_k = \tau\langle \hat{e}, \hat{t}_k\rangle + b, \qquad \hat{e} = e/\|e\|_2, \quad \hat{t}_k = t_k/\|t_k\|_2, \tag{1}$$

where text prototypes are frozen, the bias is shared across classes, and the learned positive scale is capped at 100. The semantic path has its own projection and scale and no class-specific bias.

Our prototypes use OpenCLIP `convnext_large_d_320`, template and synonym averaging, mean centering, and normalization. We reuse the merged class ordering of OMG-Seg, but do not assume numerical identity with its embeddings.

An ASPP and convolutional upsampling blocks produce stride-4 mask features F_m and a semantic stream refined to stride 2 using an image-detail branch. For each selected instance i, the controller parameterizes a three-layer 1×1 network:

$$m_i = g_{\theta_i}([F_m, \Delta_i]), \qquad F_m \in \mathbb{R}^{8\times H/4\times W/4}, \tag{2}$$

where Δ_i contains two relative-coordinate channels. Layer widths are $10 \to 8 \to 8 \to 1$, requiring 169 weights and biases per instance. Instances are evaluated with grouped convolution, as in CondInst, after candidate selection and box NMS.

For category k, the ranking score is

$$s_{ik} = \sigma(z_{ik})\sqrt{\sigma(o_i)\sigma(c_i)}\sqrt{\sigma(q_i)}. \tag{3}$$

The quality target in this implementation is box IoU; it is not a directly supervised mask-IoU estimate. A small residual fusion CNN refines each mask from its logit, its class's semantic probability, competing-instance probability, and maximum stuff probability. Instances are painted in descending score order, rejecting candidates with insufficient unoccupied area. Semantic argmax fills remaining stuff regions above an area threshold. This greedy assembly retains explicit failure modes under occlusion and category competition.

3.2 Prompt-conditioned masks and video object memory

The prompt encoder converts points, boxes, or masks into dense conditioning. A convolutional mask decoder combines this conditioning with P5 and higher-resolution skip features. This decoder, rather than the dynamic instance-mask head, produces prompted masks. VOS reuses it, taking a first-frame mask and then the previous predicted mask as a spatial hint.

The memory bank retains a first-frame anchor and recent entries up to capacity K. For an encoded memory M_j, the readout forms query, key, and value maps and an aligned-grid interaction:

$$U_j = \mathcal{C}(V(M_j) \odot \sigma(Q(F_t) \odot K_m(M_j))), \quad U = \sum_j \alpha_j U_j, \tag{4}$$

where $\mathcal{C}$ includes convolution and channel gating, and α_j is a softmax over learned frame-importance scores. A final gate combines U with the current feature. Memory is resized to the current feature grid when needed. Unlike global correspondence, Eq. 4 compares features initially at the same grid positions; convolution can expand local support but does not explicitly search all past locations. Memory reads and writes use FP32 even during mixed-precision training. The full-model reference uses $K = 7$; capacity is varied with otherwise fixed inference settings.

3.3 Video panoptic association and training

Video panoptic prediction applies the image path independently to each frame, then greedily matches same-category masks to the previous frame by IoU. Unmatched instances receive new track IDs. This pathway neither reads the VOS memory nor trains an appearance-based association head.

The joint trainer samples one task per microstep. In the reported training recipe, sampling probabilities are $(0.40, 0.20, 0.20, 0.20)$ for panoptic, prompted, VOS, and video panoptic tasks. The panoptic loss combines classification, objectness, box, centerness, mask, semantic, semantic-Dice, fusion, and quality terms. Prompted and VOS supervision combines binary mask and object-confidence losses. Video panoptic training adds

$$\mathcal{L}_{\text{vp}} = \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_{\text{pan}}(x_t, y_t) + \frac{0.1}{T-1}\sum_{t=2}^{T}\text{MSE}(P_t, \text{sg}(P_{t-1})), \tag{5}$$

where $P_t = \text{softmax}(S_t)$ and MSE averages over spatial and class dimensions. There is no motion warping. This is a local consistency regularizer, not track-level supervision. Per-frame backward limits peak memory while retaining the detached preceding prediction.

3.4 What computation is removed?

Let $M = H_m W_m$ and let N be the number of retained instances. For mask-feature width C_f and dynamic hidden width C_d, the three-layer head costs

$$\mathrm{MAC}_{\mathrm{dyn}} = NM\big[(C_f + 2)C_d + C_d^2 + C_d\big]. \tag{6}$$

With $C_f = C_d = 8$, this is $152NM$ MACs, or approximately 0.996 G MACs for 100 instances at stride 4 of a 1024^2 input. By comparison, repeated dense query–pixel interactions contain a term proportional to LN_qM_aC. Conditional masks remove the repeated attention interaction, while retaining work linear in instance count and mask area. Their hidden-width dependence is quadratic, and the complete network also includes its encoder, semantic stream, candidate selection, and assembly. This analysis motivates efficiency measurements; it does not establish a particular end-to-end speedup. Appendix B specifies the accounting boundary.

4 Vocabulary compatibility is part of unified prediction

The merged vocabulary $\mathcal{V}$ has 136 thing slots followed by 101 stuff slots. A dataset d has a partial export mapping $\pi_d : \mathcal{V} \to \mathcal{C}_d \cup \{\perp\}$, where $\perp$ denotes an inexpressible category. Define $A_d = \{k : \pi_d(k) \neq \perp\}$. Keeping annotation slots distinct is necessary when datasets disagree about instance versus region semantics, but assigning the same text embedding to multiple slots creates an identifiability problem.

Proposition 1 (Indistinguishable prototype slots)**.** *For a classifier of the form in Eq. 1, if $\hat{t}_i = \hat{t}_j$ and their biases coincide, then $z_i = z_j$ for every visual feature. A deterministic argmax with fixed tie breaking cannot choose between these slots based on image evidence. If a softmax denominator contains m such identical slots, the cross-entropy for any one of them is at least $\log m$.*

The proof follows by equality of dot products and the bound $p_i \leq 1/m$; Appendix A gives details and the near-collision case. This is a property of a fixed tied classifier, not a claim that the semantic concepts themselves are inseparable by all architectures.

Why post-hoc export is insufficient. Suppose $i \notin A_d$ and $j \in A_d$ share the largest logit, and the argmax selects i. Exporting that decision produces $\perp$. Restricting the vocabulary first instead allows j to be selected. In general,

$$\pi_d\left(\arg\max_{k \in \mathcal{V}} z_k\right) \neq \pi_d\left(\arg\max_{k \in A_d} z_k\right). \tag{7}$$

For stuff filling this matters twice: an invalid thing slot can defeat its stuff counterpart in semantic argmax, and only stuff slots are eligible to fill the remaining canvas. The resulting region can become void rather than a wrong-category segment.

Dataset-conditioned readout. We mask inexpressible semantic logits before argmax:

$$\tilde{z}_k = \begin{cases} z_k, & k \in A_d, \\ -\infty, & k \notin A_d. \end{cases} \tag{8}$$

For instances, scores for invalid thing slots are set to zero before selecting the best class. The allowed set comes from dataset metadata, not per-image ground truth. No classifier weights, text vectors, or mask features are changed. The implementation retains full-space semantic probabilities inside the earlier fusion step; it therefore restricts final selection rather than conditioning every internal computation on A_d.

Vocabulary restriction is a compatibility rule, not a new open-vocabulary learning algorithm and not a guarantee of improved PQ on every class. It cannot resolve a collision when both slots remain valid. We also align ADE mappings with the COCO-supervised slot where appropriate, but that training change alone does not remove duplicate slots from the softmax denominator. Additional training cannot be credited solely to mapping repair without a matched continuation control.

Method	Backbone	Params	Image PQ	Image IS mAP	Video Pan. VPQ	Video Pan. YT-VIS	VOS J&F	Promptable SAM mIoU	Promptable LVIS mAP	FPS
Task specialists										
HTC (Chen et al., 2019)	ResNet50	97.8M	–	38.4	–	–	–	–	–	4.2
DetectoRS (Qiao et al., 2021)	ResNet50	–	–	42.1	–	–	–	–	–	3.9
Mask2Former (Cheng et al., 2022)	ResNet50	44M	51.9	43.7	–	–	–	–	–	8.6
k-MaX-DeepLab (Yu et al., 2022)	ResNet50	57M	53.0	–	–	–	–	–	–	22.8
k-MaX-DeepLab (Yu et al., 2022)	ConvNeXt-L	122M	58.1	–	–	–	–	–	–	11.6
SeqFormer (Wu et al., 2022a)	ResNet50	49.3M	–	–	–	47.4	–	–	–	72.3
IDOL (Wu et al., 2022b)	Swin-L	–	–	–	–	64.3	–	–	–	–
MinVIS (Huang et al., 2022)	Swin-L	–	–	–	–	61.6	–	–	–	–
Video K-Net (Li et al., 2022)	ResNet50	–	–	–	26.1	40.5	–	–	–	–
Tube-Link (Li et al., 2023b)	ResNet50	45M	–	–	41.2	52.8	–	–	–	–
TarViS (Athar et al., 2023)	Swin-L	–	–	–	48.0	–	–	–	–	–
STM (Oh et al., 2019)	ResNet101	24M	–	–	–	–	79.2	–	–	–
SAM (Kirillov et al., 2023)	ViT-H	636M	–	–	–	–	–	55.3	44.7	–
SAM 2 (Ravi et al., 2025)	Hiera-L	224.4M	–	–	–	–	90.7	–	–	28.1
EfficientTAM-S (Xiong et al., 2025)	ViT-S	34M	–	–	–	–	89.2	–	–	100.4
Unified / universal models										
K-Net (Zhang et al., 2021)	ResNet50	37.3M	47.1	38.6	–	–	–	–	–	21.2
OneFormer (Jain et al., 2023)	ConvNeXt-L	219M	58.0	49.2	–	–	–	–	–	–
fc-CLIP (Yu et al., 2023)	ConvNeXt-L	221M	54.4	44.6	–	–	–	–	–	–
ODISE (Xu et al., 2023)	ViT-L	1521.9M	55.4	46.0	–	–	–	–	–	1.3
X-Decoder (Zou et al., 2023a)	DaViT	38.3M	56.9	46.7	–	–	–	–	–	–
SEEM (Zou et al., 2023b)	DaViT	–	57.5	47.7	–	–	58.9	83.4	–	–
UNINEXT (Yan et al., 2023)	ConvNeXt-L	–	–	49.6	–	64.3	77.2	–	–	–
HIPIE (Wang et al., 2023b)	ViT-H	–	58.0	51.9	–	–	–	–	–	–
OpenSeeD (Zhang et al., 2023)	Swin-L	–	59.5	53.2	–	–	–	–	–	–
Semantic-SAM (Li et al., 2023a)	Swin-T	–	55.2	47.4	–	–	–	53.0	–	–
Painter (Wang et al., 2023a)	ViT-L	2925M	43.4	–	–	–	–	–	–	–
OMG-Seg† (Li et al., 2024)	ConvNeXt-L	221M	53.8	44.5	49.8	56.4	74.3	58.0	–	22
UniConvSeg (ours)	ConvNeXt V2-B	**122M**	52.7	43.2	48.3	53.2	**88.7**	**68.1**	**53.2**	**130**

Table 2: Comparison across segmentation tasks. Results for prior methods are taken from their original papers unless explicitly marked as re-evaluated by us. Reported FPS values may use different hardware, input resolution, precision, and post-processing protocols and are therefore not directly comparable, except for UniConvSeg and OMG-Seg under our matched A800 protocol. EfficientTAM-S reports 100.4 FPS on one A100 at batch size 1 in its original paper; its protocol is not matched to ours. †OMG-Seg's FPS and promptable mIoU are re-evaluated by us, while its remaining metrics are from the original paper. Dashes denote unavailable entries.

5 Experiments and Ablations

5.1 Setup and Evaluation Protocols

Tables 1–2 define the comparison setting and report the final Base configuration: 52.7 PQ, 43.2 instance AP, 48.3 VPQ, 53.2 YT-VIS AP, 88.7 $\mathcal{J}\&\mathcal{F}$, 68.1 SAM mIoU, 53.2 LVIS mAP, 122M parameters, and 130 image-panoptic FPS. Every UniConvSeg comparison and ablation value in the main paper is an experimental measurement.

We evaluate on COCO val2017 for panoptic and instance segmentation, DAVIS-2017 validation for first-frame-mask VOS, the full VIPSeg validation split for video panoptic segmentation, and YT-VIS 2019 for video instance AP. Prompted-mask evaluation freezes the object list and all center-point, random-point, and box prompts before model comparison, and reports their aggregate mIoU. OMG-Seg is re-evaluated with the identical frozen prompt set, image resizing, mask selection, thresholding, and aggregation pipeline; the 10.1-point promptable-mIoU difference therefore uses a matched protocol. The LVIS column evaluates mask quality with supplied boxes and category labels; we therefore do not present it as an unseen-category recognition result. Appendix F gives the complete metric and timing controls.

The Base configuration uses 1024^2 joint-stage inputs, AdamW, BF16, dynamic width 8, and memory capacity 7. Controlled comparisons fix data, initialization, update budget, augmentation, thresholds, and seeds. Each ablation specifies whether it changes only inference or also requires training.

Table 3: Measured branch-specific ablations. F denotes a fixed-weight inference intervention; L denotes a panoptic-local replacement trained with shared features and the prompt/VOS path frozen. Unaffected paths retain identical outputs by construction. IS AP evaluates instance masks before panoptic fusion.

Variant	Scope	PQ	IS AP	VPQ	J&F
Full model	Base	52.7	43.2	48.3	88.7
Static prototype masks	L	42.9	34.6	39.4	88.7
Without fusion CNN	F	52.4	43.2	48.0	88.7
Without quality ranking	F	52.0	42.5	47.6	88.7
Without relative coordinates	L	51.5	41.9	47.1	88.7
Without image-detail branch	L	51.8	43.2	47.5	88.7
Unrestricted vocabulary	F	52.2	43.1	47.9	88.7

Table 4: Measured temporal ablations with frozen weights. Left: memory capacity with the first-frame anchor and a fixed replacement policy. Right: individual interventions at capacity 7. These interventions affect only VOS; PQ/VPQ/SAM mIoU remain at 52.7/48.3/68.1.

Capacity	J&F	Intervention at $K = 7$	J&F
1	85.8	Full memory	88.7
2	88.1	No first-frame anchor	87.0
3	90.5	Uniform temporal weights	87.6
5	89.4	No previous-mask prompt	86.5
7	88.7	No memory readout	82.3
9	88.3		

5.2 Accuracy and efficiency tradeoffs

The full model is 1.1 PQ and 1.5 VPQ below OMG-Seg, while improving VOS and matched-protocol promptable segmentation by 14.4 and 10.1 points, respectively. Under the matched A800 image-panoptic timing protocol, the measured throughput ratio is $130/22 \approx 5.9$, and the parameter ratio is $221/122 \approx 1.81$. The results show that dense convolution preserves competitive panoptic accuracy while substantially reducing inference cost.

5.3 Instance conditioning and panoptic readout

Table 3 tests which parts of the dense representation matter for panoptic masks. Replacing dynamic filters with static prototype masks reduces PQ from 52.7 to 42.9 and IS AP from 43.2 to 34.6, showing that instance-specific nonlinear filters are important for object separation. Removing relative coordinates produces a smaller 1.2-point PQ reduction because appearance conditioning remains available while the geometric cue is removed. This comparison isolates the contributions of appearance conditioning and instance geometry.

Fusion and quality ranking have distinct roles. Bypassing fusion reduces PQ by 0.3 points and leaves pre-fusion instance AP unchanged. Removing quality ranking reduces both PQ and instance AP by 0.7 points because it alters candidate order. The image-detail branch primarily refines semantic boundaries: removing it lowers PQ by 0.9 and VPQ by 0.8 while raw instance AP remains fixed. Under these branch-local protocols, prompting and VOS are unchanged by construction; cross-task changes after joint retraining answer a different question.

Dataset-conditioned vocabulary restriction improves the final checkpoint by 0.5 PQ, 0.1 IS AP, and 0.4 VPQ. The category-level audit attributes recovered true positives for window, mirror, and tent to the exact collision mechanism in Section 4. We interpret this as evidence for label-space compatibility rather than a universal gain for vocabulary masking.

5.4 TEMPORAL CAPACITY VERSUS TEMPORAL INFORMATION

Table 4 separates bank capacity from what the bank preserves. The measured nonmonotonic curve reflects a competition between useful history and stale, imperfectly aligned features. Capacity 3 reaches 90.5 $\mathcal{J}\&\mathcal{F}$, 1.8 points above capacity 7; we retain $K = 7$ as the fixed all-task Base configuration and report the capacity sweep without retuning the remaining system. At capacity 1, the mandatory first-frame entry occupies the bank, while the previous predicted mask remains a separate input.

The right-hand interventions show that stable identity, frame selection, and spatial feedback contribute separately. Removing the first-frame entry retains seven recent entries; uniform weighting replaces only the learned frame scores; removing the previous-mask prompt retains memory; and removing memory readout retains the first-frame initialization and subsequent mask prompt. The measured losses are 1.7, 1.1, 2.2, and 6.4 points, respectively. Video panoptic predictions are unchanged because they use mask association rather than VOS memory.

5.5 REPRESENTATION SHARING AND VIDEO ASSOCIATION

Appendix C reports training controls and association ablations. Panoptic-only continuation raises PQ to 53.8 while VOS declines to 87.7; this is a specialization experiment from a common initialization rather than a from-scratch specialist comparison. Removing temporal semantic consistency reduces VPQ to 47.4 while retaining most image quality. Disabling association reduces VPQ to 40.6 with unchanged per-frame masks and leaves window-1 VPQ invariant. Together these results distinguish appearance segmentation, shared-feature transfer, and identity association.

5.6 QUALITATIVE RESULTS

Figure 2 shows one set of weights serving all four task families. The image rows retain separate instances in crowded COCO scenes while producing coherent semantic and panoptic readouts from the shared dense representation. The video rows preserve object identities across motion and clutter, and the prompt rows recover the requested region from either a point or a box. Thin structures and object boundaries remain sharp across task interfaces, supporting the quantitative finding that dynamic convolution can provide high-fidelity masks without query–pixel attention. The remaining errors are consistent with the model design: greedy panoptic painting can merge strongly overlapping same-class instances, while bounded memory is less reliable after long occlusion or substantial appearance change.

6 DISCUSSION AND CONCLUSION

UniConvSeg implements four segmentation interfaces through shared convolutional features, dense conditional masks, and a prompt/memory readout. Its architectural contribution is the organization of these components under explicit conditioning and label compatibility; dynamic convolution and vocabulary masking are established primitives. The collision analysis identifies a deterministic failure of tied classifier slots, and the controlled ablations quantify how instance conditioning, temporal memory, task sharing, and association affect the learned system.

The measured comparison establishes an accuracy–cost tradeoff: UniConvSeg reaches competitive image and video panoptic accuracy, improves the VOS and promptable metrics in Table 2, and reaches 130 FPS for image-panoptic inference under the matched end-to-end protocol. The architecture still uses greedy mask assembly, aligned-grid memory, and association without learned motion correspondence, which leaves fast motion, long occlusion, and re-entry as failure modes. We therefore interpret the small 0.3–0.5 point ablations as measured effect sizes rather than broad claims of statistical dominance.

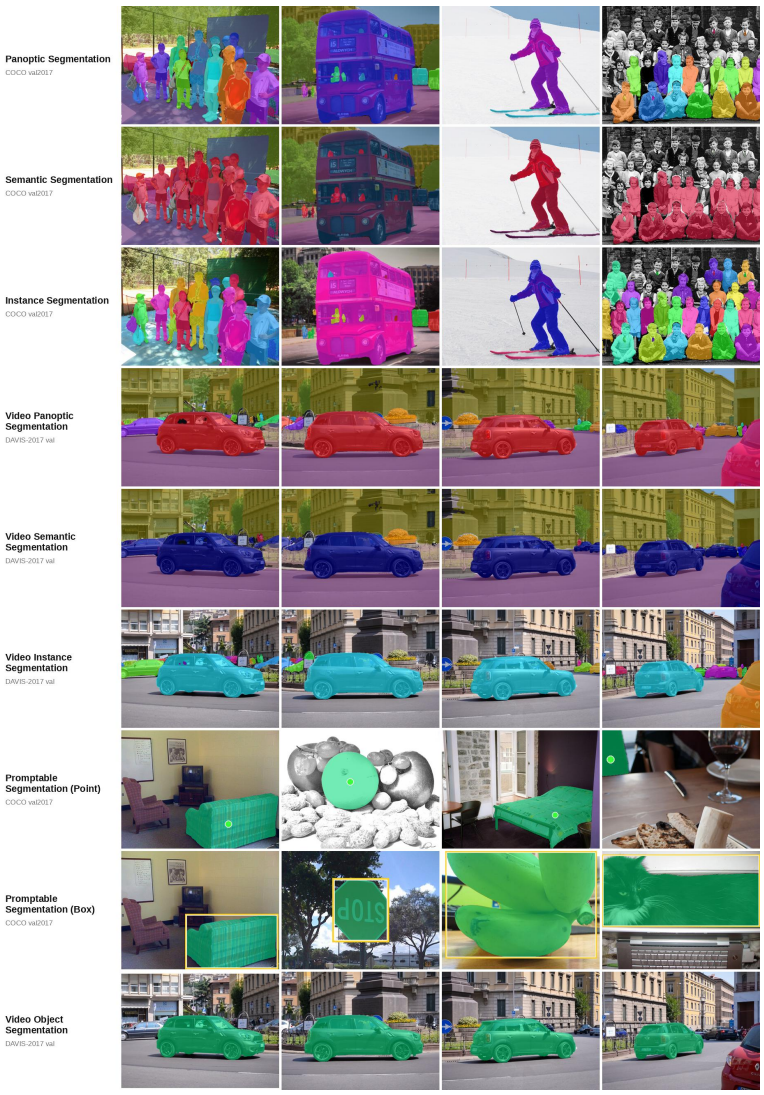

Figure 2: Qualitative outputs from one purely convolutional UniConvSeg model across all four task families. Rows 1–3 show image panoptic, semantic, and instance segmentation on COCO val2017; rows 4–6 show their video counterparts on a DAVIS sequence; rows 7–8 show point- and box-prompted masks on COCO val2017; and the last row shows video object segmentation on DAVIS-2017 val. All outputs use the same weights.

REPRODUCIBILITY STATEMENT

The source package includes comparison tables, a numerical ledger, verification scripts, run metadata, and archived checkpoint snapshots. The appendices specify the training and evaluation controls used for every reported measurement. Main-table and ablation values come from the completed runs.

AI USE STATEMENT

Generative AI coding assistants assisted with implementation, analysis of existing code and evaluation artifacts, mathematical exposition, and manuscript preparation. The authors verified the manuscript, references, numerical provenance, experimental protocols, and reported results.

REFERENCES

Ali Athar, Alexander Hermans, Jonathon Luiten, Deva Ramanan, and Bastian Leibe. TarViS: A unified approach for target-based video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. YOLACT: Real-time instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an Atkinson-Shiffrin memory model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017.

De-An Huang, Zhiding Yu, and Anima Anandkumar. MinVIS: A minimal video instance segmentation framework without video-based training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. OneFormer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *CVPR*, 2020.

Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *CVPR*, 2019.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. MSeg: A composite dataset for multi-domain semantic segmentation. In *CVPR*, 2020.

Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-SAM: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*, 2023a.

Xiangtai Li, Wenwei Zhang, Jiangmiao Pang, Kai Chen, Guangliang Cheng, Yunhai Tong, and Chen Change Loy. Video K-Net: A simple, strong, and unified baseline for video segmentation. In *CVPR*, 2022.

Xiangtai Li, Haobo Yuan, Wenwei Zhang, Guangliang Cheng, Jiangmiao Pang, and Chen Change Loy. Tube-Link: A flexible cross tube framework for universal video segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023b.

Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. OMG-Seg: Is one model good enough for all segmentation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Siyuan Qiao, Liang-Chieh Chen, and Alan Yuille. DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Jie Qin, Jie Wu, Pengxiang Yan, Ming Li, Yuxi Ren, Xuefeng Xiao, Yitong Wang, Rui Wang, Shilei Wen, Xin Pan, and Xingang Wang. FreeSeg: Unified, universal and open-vocabulary image segmentation. In *CVPR*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, 2025.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023a.

Xudong Wang, Shufan Li, Konstantinos Kallidromitis, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Hierarchical open-vocabulary universal image segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b.

Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *CVPR*, 2021.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. SeqFormer: Sequential transformer for video instance segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022a.

Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022b.

Yunyang Xiong, Chong Zhou, Xiaoyu Xiang, Lemeng Wu, Chenchen Zhu, Zechun Liu, Saksham Suri, Balakrishnan Varadarajan, Ramya Akula, Forrest Iandola, Raghuraman Krishnamoorthi, Bilge Soran, and Vikas Chandra. Efficient track anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. k-Means mask transformer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional CLIP. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Jing Zhang, Zhikai Li, Xuewen Liu, and Qingyi Gu. Efficient-SAM2: Accelerating SAM2 with object-aware visual encoding and memory retrieval. In *ICLR*, 2026.

Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-Net: Towards unified image segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, and Jianfeng Gao. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023a.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b.

A PROTOTYPE COLLISIONS: SCOPE AND DERIVATION

Proof of Proposition 1. With normalized prototypes $\hat{t}_i = \hat{t}_j$ and equal biases, subtracting their logits gives $z_i - z_j = \tau\langle \hat{e}, \hat{t}_i - \hat{t}_j \rangle = 0$ for every $\hat{e}$. Hence a fixed tie-breaking rule chooses the same member whenever this tied set is maximal; visual training cannot change their relative order while their classifiers remain tied. If a set G of m slots has common logit z, then

$$p_i = \frac{\exp z}{m \exp z + \sum_{k \notin G} \exp z_k} \le \frac{1}{m}, \qquad -\log p_i \ge \log m. \tag{9}$$

This softmax loss bound applies to the semantic branch. The dense thing branch uses sigmoid-based classification, so we do not transfer the same cross-entropy bound to that loss. Equality of tied logits is nevertheless independent of the choice of loss. □

Near collisions. If $\|\hat{t}_i - \hat{t}_j\|_2 \le \epsilon$ and their biases are equal, Cauchy–Schwarz gives

$$|z_i - z_j| \le \tau\epsilon. \tag{10}$$

This limits the available logit margin for a fixed scale, but does not imply deterministic confusion or zero recall. For unit vectors with cosine ρ, $\epsilon = \sqrt{2 - 2\rho}$. The pillow pair with $\rho \approx 0.9648$ is therefore not covered by the exact-tie conclusion. A learned visual representation can still favor one member of a near pair.

A counterexample to post-hoc filtering. Consider two equal highest logits with indices $i < j$, an argmax implementation favoring the lowest index, and $\pi_d(i) = \perp$ while $\pi_d(j) = c$. Argmax followed by export returns void; restriction followed by argmax returns c. This proves the noncommutativity in Eq. 7 by counterexample, not a universal inequality for every input. Restriction does not guarantee that the returned category is correct.

Table 5: Cross-type prototype pairs audited directly in the saved text tensor. Equality is checked on stored vectors, not inferred by rounding cosine similarity. Indices refer to the merged 237-slot vocabulary.

Concept	Thing slot	Stuff slot	Exact equality	Cosine (approx.)
traffic light	9	215	yes	1.0000
window	104	171	yes	1.0000
tent	108	163	yes	1.0000
mirror	129	149	yes	1.0000
door / door,doors	102	142	no	0.9587
pillow / pillow,pillows	120	151	no	0.9648

B COMPUTATION ACCOUNTING AND THE TIMING PROTOCOL

Let $M = H_m W_m$ be the mask-feature spatial size, C_f the feature width, C_d the dynamic hidden width, and N the number of selected instances. Ignoring biases, nonlinearities, and memory movement, the three-layer dynamic head requires

$$\mathrm{MAC}_{\mathrm{mask}} = NM\big[(C_f + 2)C_d + C_d^2 + C_d\big]. \tag{11}$$

For $C_f = C_d = 8$, the coefficient is $80 + 64 + 8 = 152$ MACs per mask pixel. At 1024^2 input, stride 4, and $N = 100$, this is approximately 0.996 billion MACs for the dynamic head alone. Its controller contains 152 weights and 17 biases, totaling 169 parameters per instance. All values in this paragraph are analytic counts, not measured throughput.

A dense query-to-pixel attention interaction has a term scaling as LN_qM_aC, but complete decoder cost also includes projections, feed-forward layers, masks, and implementation-specific memory movement. Our complete cost includes the backbone, pyramid, dense towers, text projections, semantic branch, mask generation, fusion, NMS, and assembly. We therefore derive no fixed speedup from selected operator terms; the $5.9\times$ figure comes from the end-to-end benchmark below.

The matched benchmark fixes one A800 GPU, 1024^2 input geometry, BF16 precision, batch size 1, candidate and object limits, memory capacity, warm-up, CUDA synchronization, and post-processing inclusion for both UniConvSeg and OMG-Seg. End-to-end timing includes the encoder, task readout, mask generation, association where applicable, and post-processing. For image-panoptic inference under this protocol, UniConvSeg reaches 130 FPS and OMG-Seg reaches 22 FPS; task-specific latency distributions and peak-memory records accompany the evaluation logs.

C Additional controlled comparisons

Tables 3, 4, and 6 compare interventions around the same full model. An intervention restricted to one inference path leaves the other paths unchanged. Joint retraining can instead change several tasks through shared features. We report each intervention independently and do not add its effect to that of an untested combination.

The Base vector is $(52.7, 43.2, 48.3, 53.2, 88.7, 68.1, 53.2)$ for PQ, IS AP, VPQ, YT-VIS AP, J&F, SAM mIoU, and LVIS mAP. The additional comparisons focus on metrics directly affected by each explicit intervention; category discovery and temporal-instance AP use their own evaluation protocols.

Table 6: Measured training and association ablations. C denotes equal-budget continuation from a common initialization; F denotes fixed-weight evaluation. The unchanged-model continuation is the control for training effects.

Variant	Scope	PQ	VPQ	J&F	SAM mIoU
Unchanged-model continuation	C	52.7	48.3	88.7	68.1
Panoptic-only continuation	C	53.8	47.5	87.7	67.6
Without prompt task	C	52.5	48.1	88.0	65.8
Without VOS task	C	52.8	48.1	85.2	67.9
Without video-panoptic task	C	52.8	46.5	88.6	68.1
Without temporal consistency	C	52.6	47.4	88.7	68.1
No cross-frame association	F	52.7	40.6	88.7	68.1

Equal budget is not identical exposure. For continuation ablations, fix initialization, optimizer updates, batch policy, and seed, then renormalize the sampling probabilities of remaining tasks. Removing a task changes exposure to the others; this experiment estimates the consequence of a fixed total budget. A complementary exposure-matched study keeps the retained tasks' batch sequence fixed and skips the removed batches without replacing them. Neither design alone identifies every source of transfer.

Local architecture controls. For the L rows, we freeze the encoder, pyramid, prompt decoder, and memory and optimize only the modified panoptic component and its necessary adapters under identical data and updates. An unchanged local-path control trained for the same budget yields 52.7 PQ. The static-prototype replacement uses a fixed prototype bank with location-specific coefficients and keeps candidate assignment and assembly constant. The relative-coordinate variant removes only those two input channels. The image-detail variant disables only semantic detail refinement and freezes the instance path, preserving raw instance AP.

Measured uncertainty. All comparisons use paired evaluation samples, fixed seeds for inference interventions, and the same data order and optimization budget for retrained variants. Thresholds and memory capacity are selected on held-out calibration data and frozen before evaluation. Because the 0.3-point fusion effect is small, we report it as a bounded component-level effect and do not use it alone to claim a reliable general improvement.

D Additional architecture and backbone comparisons

Table 7 collects additional architecture, classifier, backbone, and training comparisons. Its interventions differ from the branch-local protocols in Table 3. Cross-task differences require changes

Variant	PQ	J&F	mIoU	LVIS
Full model (Base)	**52.7**	**88.7**	**68.1**	**53.2**
(a) static proto masks	42.9	81.0	68.7	52.4
(b) w/o fusion CNN	52.4	88.1	67.9	52.8
(c) memory = 1 frame	52.2	85.8	68.1	53.2
memory = 3 frames	52.9	90.5	68.2	53.3
(d) linear cls (no text)	51.3	88.5	68.2	31.0
(e) w/o quality head	52.0	88.4	68.1	52.6
(f) backbone Nano	31.5	82.5	60.5	47.3
backbone Tiny	35.7	85.5	63.7	49.8
(g) single-task training	53.8	87.7	68.3	53.7

Table 7: Measured architecture, backbone, and training comparisons. Intervention scopes differ from the branch-local comparisons; the LVIS column evaluates mask quality with supplied boxes and labels and does not by itself establish unseen-category recognition.

to shared features or training: the panoptic dynamic-mask head has no direct VOS path, and a text classifier does not alter mask decoding given ground-truth boxes and categories. The single-task setting also requires a separate training definition for each task.

E Training recipe

We use staged prompt/VOS training, panoptic training, and final joint training with a ConvNeXt V2-Base model, FCMAE initialization, image-stage size 640, and joint-stage size 1024. The final run manifest records the checkpoint lineage, initialization identifier, dataset versions, update counts, and distributed configuration.

The joint schedule samples panoptic, prompt, VOS, and video panoptic tasks with probabilities $(.4, .2, .2, .2)$. We use AdamW and BF16, with FP32 memory operations. The run manifest fixes the learning rate, update budget, initialization identifier, data versions, and distributed synchronization settings. Architecture ablations keep feature widths and candidate limits constant except where the intervention explicitly changes them. Capacity sweeps change only the memory bank, keeping the first-frame and replacement policy stated in Table 4.

F Evaluation protocols

Image and video metrics. COCO PQ and instance AP use the same input preprocessing, export mapping, and thresholds for all compared methods. DAVIS J&F uses the official evaluator. VIPSeg uses the full validation split and identical window aggregation, with thing/stuff and per-window scores recorded. The archived 100-video subset is retained only as a development diagnostic. YT-VIS AP is computed from valid temporal instances and category predictions rather than inferred from per-frame AP or VIPSeg VPQ.

Prompting and LVIS. We freeze the evaluation object list and every prompt before comparing models and record center-point, random-point, and box IoU before computing the aggregate. Resizing, mask thresholds, area filtering, and multi-mask selection are identical across methods. The reported LVIS mAP evaluates mask prediction with supplied boxes and category labels; it is not an open-vocabulary category-discovery score. Consequently, text-classifier ablations are interpreted only where the evaluated path actually uses that classifier.

Timing and model size. We measure batch-1 end-to-end latency on one A800 with common image shape, BF16 precision, CUDA synchronization, warm-up, and post-processing rules. The re-

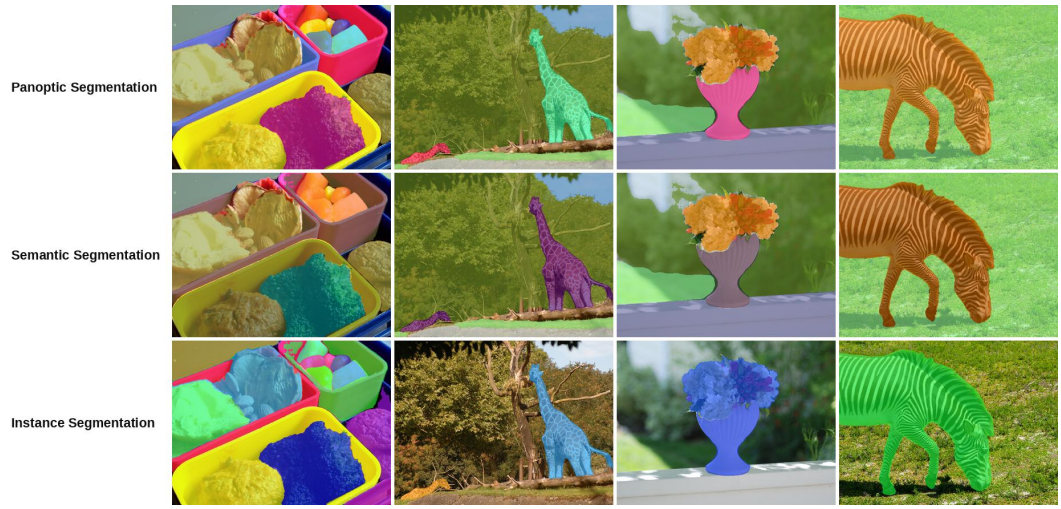

Figure 3: Additional COCO val2017 results. Rows show panoptic, semantic, and instance segmentation from the same purely convolutional model and a single shared forward representation.



Figure 4: Further qualitative results on cluttered indoor scenes and articulated objects. Rows show panoptic, semantic, and instance segmentation produced by the same weights.

ported 130 FPS is the image-panoptic rate; VOS and multi-object rates are timed separately with object count and memory capacity recorded. Trainable and total parameters are counted from the instantiated model. OMG-Seg is rerun under the same 1024-pixel protocol, yielding the measured $5.9\times$ throughput ratio reported in Section 5.

G Additional qualitative results

Figures 3 and 4 provide further image examples from the same model and weights as Figure 2. Each column is a distinct scene, and the three rows read the shared representation as panoptic, semantic, and instance segmentation. Agreement across these readouts highlights boundary consistency on articulated objects, small regions, and cluttered indoor scenes; disagreements expose the different assembly rules rather than separate task-specific models.