iFAN: Inference-Aware Learning for Plain Mask Transformers

Anonymous authors
Paper under double-blind review

Abstract

Query-based mask transformers assemble segmentation outputs through pixel-wise competition among query predictions of the final layer, yet this inference process is not explicitly optimized during training. We identify two key mismatches: the query with the highest probability–mask score does not necessarily produce the most accurate mask, and final-layer decoding may discard superior predictions from intermediate layers. To address these issues, we propose Inference-Aware Learning (**iFAN**), a general training framework for plain mask transformers. iFAN introduces Adjusted Probability-Mask Ranking (APMR), which aligns query competition with predicted mask quality and suppresses high-confidence but inaccurate competitors. We further employ Cross-Layer Self-Distillation (CLSD) to transfer stronger intermediate predictions to the final layer. The ranking and distillation objectives are training-only, while inference retains efficient final-layer decoding. Experiments on COCO, ADE20K, and Cityscapes demonstrate consistent improvements across panoptic, instance, and semantic segmentation, as well as across different architectures, backbone scales, and input resolutions. iFAN consistently improves performance by 1.16 PQ, 1.38 AP, and 0.63 mIoU, bringing plain mask transformers to an accuracy level comparable to state-of-the-art methods while preserving the efficiency of plain mask transformers.

1 Introduction

Query-based segmentation has become a promising framework for semantic, instance, and panoptic segmentation (Wang et al., 2021; Yu et al., 2022; Jain et al., 2023; Shahabodini et al., 2026). Unlike conventional approaches that classify each pixel using a fixed set of prediction headers, modern mask transformers segment an image with a set of learnable queries. Each query independently predicts a class distribution together with a corresponding mask, and the final segmentation result is obtained by a competition among these query-wise prediction maps during output assembly (Li et al., 2023b). This query-based formulation is naturally compatible with Vision Transformers (ViTs) (Dosovitskiy et al., 2020), since image tokens and query tokens can be processed within the same attention framework. Recent encoder-only methods further show that, based on strong pre-training backbones such as DINOv2 (Oquab et al., 2024), a plain ViT can support segmentation by simply injecting queries into its later layers without bells and whistles (Kerssies et al., 2025).

In this paper, we rethink the inference process of mask transformers and identify the limitations of widely used max-prob-mask strategy (Cheng et al., 2021). In this strategy, each query predicts a class distribution and a class-agnostic foreground probability mask, which are then combined to determine the pixel-wise segments. However, the mask predicted by each query is isolated with class prediction and thus lacks global information of object, *e.g.*, the object shape and contour (Ke et al., 2022). Consequently, the predicted class may be inconsistent with the predicted mask (Hu et al., 2023). For example, in the *traffic-light* case of Fig. 1, although the max-prob-mask strategy predicts the correct category, the resulting segmentation covers only the most salient regions of the traffic light. For the object without a consistent shape, such as *kite* in Fig. 1, query-based models hardly predict the whole region. Based on the inference results of a mask transformer (Kerssies et al., 2025) on COCO val2017 (Lin et al., 2014), we find that 64.45% of the segmentation results selected by max-prob-mask strategy are different from their theoretical upper-bound counterparts by "max-IoU" strategy. With a strict threshold, 29.27% of the results exhibit high prob-mask scores while low IoUs.

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.

We claim that these suboptimal results are induced by the incomplete competition strategy of probability-mask pairs (Cheng et al., 2021). To address this issue, we introduce a global mask quality score for each query to adjust the probability-mask score map for segmentation prediction. However, mask-quality calibration alone cannot fully suppress high-scoring yet inaccurate competitors. Hungarian matching determines the optimal assignment at each iteration but does not preserve query identities across training: a query may switch repeatedly between the states of matched foreground and unmatched background (Li et al., 2023a; Liu et al., 2023; Zhang et al., 2022). Importantly, unmatched queries are supervised only by no-object classification, without explicit suppression of their mask responses (Cheng et al., 2022). They may retain plausible knowledge of foreground mask from previous iterations, causing multiple plausible masks for the same object. These plausible masks pose strong competition during inference, motivating us to introduce an explicit cross-query ranking objective (Pu et al., 2023). Specifically, for each foreground target, we identify the unmatched queries with high probability of target class yet low mask overlap as hard negatives. We then introduce a pairwise ranking objective that encourages the matched query to produce a higher probability-mask score than that of negatives. To stabilize the learning of classification and mask prediction, we only use the gradient of ranking loss *w.r.t* mask quality to update the model parameters.

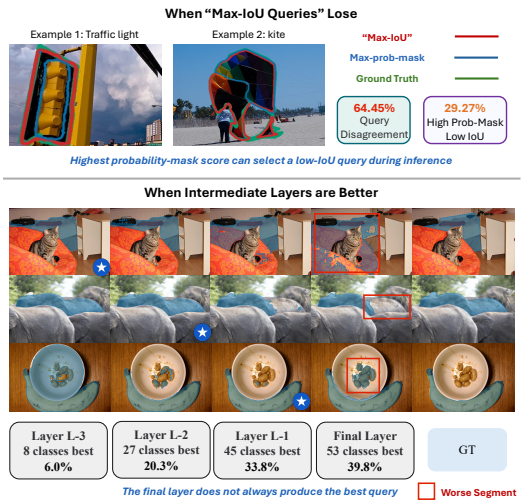


Figure 1: We study the performance of mask transformer on COCO val2017 among queries and layers with ground truth masks. Top: the *max-prob-mask* query (blue) is selected to generate the mask result of target object while achieving inferior performance than the **oracle result** of "*max-IoU*" query (red), which incurs 64.45% suboptimal result. Bottom: the distribution of "*max-IoU*" query across layers. A better query can be found in intermediate layer (star indicates max IoU), which can provide self-distillation information.

Most mask transformers use only the output of final layer to produce segmentation predictions, even if the auxiliary supervision is provided at intermediate layers. We analyze the segmentation results generated by the last several layers of a mask transformer (Kerssies et al., 2025). Only 53 of the 133 classes (39.8%) achieve their best performance at the final layer (Figure 1). Simply aggregating the predictions of multiple layers can improve performance, but it also introduces additional post-processing complexity and computational costs. Instead, we propose to extract high-quality information from preceding layers to supervise the final layer while keeping the inference strategy unchanged. Therefore, the Cross-Layer Self-Distillation (CLSD) is integrated in the training stage to distill high-quality intermediate-layer information by the final layer while keeping the inference stage unchanged (Zhang et al., 2019; Wang et al., 2024).

In summary, we upgrade the training framework of mask transformers motivated by the deficiencies of segmentation inference, which thus refers to as inference-aware learning task. Our main contributions are summarized as follows:

- We adjust the probability-mask scoring criterion with a global quality score to address the deficiency of inference. In addition, we employ object-oriented ranking task to produce more discriminative prediction scores.
- We fully exploit the useful information contained in different layers of mask transformers and propose a self-distillation strategy that improves the predictions of the final layer without additional inference cost.
- Across multiple mask-transformer architectures and datasets, iFAN consistently improves panoptic, instance, and semantic segmentation, achieving average gains of 1.16 PQ, 1.38 AP, and 0.63 mIoU, respectively, with virtually no additional inference overhead.

2 Related Work

2.1 Query-Based and Quality-Aware Segmentation

DETR introduced object queries and bipartite set prediction for detection (Carion et al., 2020). MaskFormer and Mask2Former extend this formulation to segmentation through class–mask prediction (Cheng et al., 2021; 2022), while OneFormer and Mask DINO develop it for universal segmentation and joint detection–segmentation modeling (Jain et al., 2023; Li et al., 2023b). EoMT and PMT further inserts segmentation queries directly into the late blocks of a pre-trained ViT (Kerssies et al., 2025; Cavagnero et al., 2026). Despite their architectural differences, these methods must assemble unordered class–mask predictions into a final output, making the reliability of the probability-mask score central to query competition.

Quality estimation and ranking improve prediction reliability. Mask Scoring R-CNN calibrates instance scores with predicted mask IoU (Huang et al., 2019); Panoptic SegFormer combines classification and segmentation quality to resolve mask overlaps (Li et al., 2022); and Rank-DETR learns quality-aware ranking for detection queries (Pu et al., 2023). These methods focus on instance-score calibration, inference-time overlap handling, or box-level ranking. Adjusted Probability-Mask Ranking (APMR) instead optimizes the pixel-wise competition of mask-transformer inference. It learns a soft-IoU-supervised mask-quality score, constructs an adjusted probability-mask score map, and ranks each matched query above hard unmatched competitors within its target region. Its ranking gradients update only the quality branch, preserving the original classification and mask objectives.

2.2 Intermediate Supervision and Self-Distillation

Intermediate supervision applies annotation-based objectives before the final output. Deeply-Supervised Nets introduced auxiliary hidden-layer losses (Lee et al., 2015), while query-based segmenters supervise intermediate class and mask predictions (Li et al., 2022). MP-Former further promotes mask consistency across decoder layers (Zhang et al., 2023). These approaches improve intermediate predictions or their evolution, but do not explicitly transfer a better intermediate probability-mask prediction to the final layer.

Self-distillation transfers knowledge between predictions within a training process, as exemplified by Be Your Own Teacher (Zhang et al., 2019). For set prediction, DETRDistill uses a pre-trained teacher to transfer matched logits, features, and query priors, whereas OD-DETR uses an exponential-moving-average teacher (Chang et al., 2023; Wu et al., 2024). Cross-Layer Self-Distillation (CLSD) requires neither. For each target, it selects the intermediate matched query with the highest soft-IoU and retains it only when it outperforms the final prediction. CLSD then distills its adjusted probability-mask score map $\hat{S}$ through a one-sided objective, transferring useful intermediate knowledge while retaining final-layer-only inference.

3 Method

In this section, we introduce the details of **Inference-Aware Learning (iFAN)** as illustrated in Fig. 2. Based on the plain mask transformer, we adjust the predicted probability-mask map by mask quality and Probability-Mask Ranking strategy, promoting the distinction of final predictions from matched queries and unmatched queries. To further acquire the knowledge of intermediate layers, the Cross-Layer Self-Distillation is injected into our framework for the promotion of matched queries. Those strategies motivated by the deficiencies in inference stage improve the performance of mask transformer, composing the framework of Inference-Aware Learning.

3.1 Preliminaries

Given an input image $I \in R^{H \times W}$, the mask transformer predicts N probability-mask pairs

$$\hat{y} = \{(\hat{\mathbf{p}}_i, \hat{M}_i)\}_{i=1}^{N} \tag{1}$$

where the pair $(\hat{\mathbf{p}}_i, \hat{M}_i)$ is the prediction result from i-th query. $\hat{\mathbf{p}}_i \in [0,1]^{K+1}$ is the class-probability distribution over K foreground classes and the no-object class $\varnothing$. $\hat{M}_i$ is the mask probability map of the foreground, which is usually class-agnostic. Similarly, the ground-truth segments

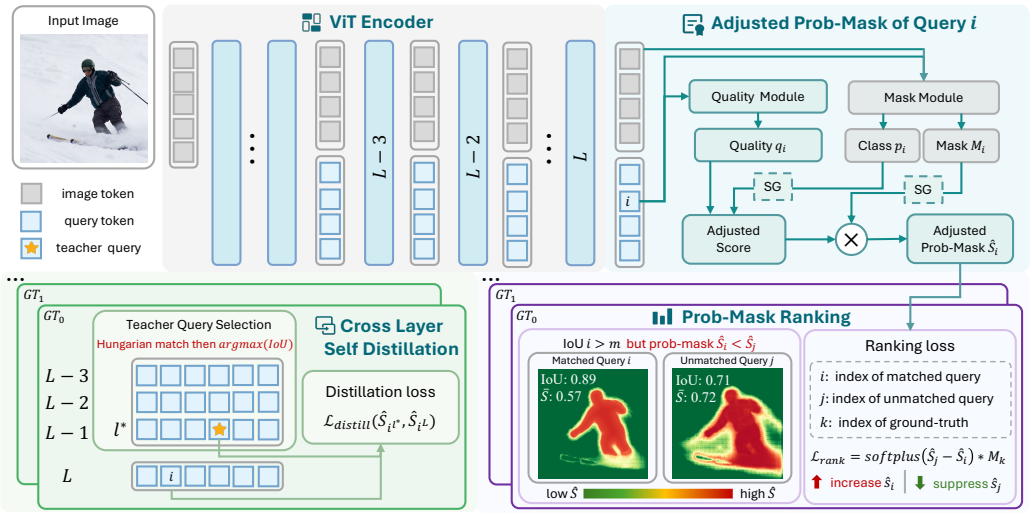


Figure 2: **Overview of iFAN.** APMR addresses the query-wise mismatch by learning mask quality, calibrating the scores used for assembly, and ranking a matched query above high-evidence unmatched query. CLSD addresses the layer-wise mismatch by selecting a stronger teacher query from an intermediate layer and transferring its assembled evidence to the final layer. At inference, the adjusted-prob-mask is only computed for final layer. The APMR ranking objective and CLSD are training-only.

are represented as

$$y = \{(c_i, M_i)|c_i \in \{1, \dots, K\}, M_i \in \{0,1\}^{H \times W}\}_{i=1}^{N^{gt}} \tag{2}$$

During training, the Hungarian matching algorithm assigns each ground-truth to a unique query by minimizing the matching cost between pairs of query predictions and ground-truths. The learning tasks on the matched pairs include object classification and foreground mask prediction. Unmatched queries are usually considered as background with only the supervision of no-object classification.

The inference stage follows a slightly different decision rule. Since queries may predict overlapping masks of objects, the segmentation assembly strategy must assign each pixel to one of the N predicted probability-mask pairs according to the maximal probability-mask score of object class and object-agnostic mask.

3.2 Adjusted Probability-Mask Ranking

The standard segmentation assembly constructs the segmentation result based on the pixel-wise competition among probability-mask responses, without considering the holistic reliability of the predicted mask. Therefore, we introduce a lightweight shared quality head to estimate the global mask quality $\hat{q}_i$ of the i-th query, resulting in the prediction tuple $(\hat{q}_i, \hat{\mathbf{p}}_i, \hat{M}_i)$. The quality head consists of a single linear layer followed by a sigmoid function, and its parameters are shared across all queries and prediction layers, introducing only negligible parameter overhead. The ground-truth of $\hat{q}_i$ for matched queries is the *soft_IoU* of predicted mask $\hat{M}_i$ and its corresponding ground-truth segmentation mask, and is set as 0 for unmatched queries. In the training stage, the BCE loss is integrated into the segmentation task for the learning of mask quality, comprising the augmented segmentation loss, and the unmatched queries with high class probability are selected as hard negatives.

To complement global information in segmentation competition, we integrate predicted quality into probability-mask and obtain the adjusted probability-mask pair $((\hat{q}_i * \hat{\mathbf{p}}_i)^\gamma, \hat{M}_i)$, where γ is set to 2 to enhance discrimination in the high-score range. Together with the hard mining strategy, the adjusted score $(\hat{q}_i * \hat{\mathbf{p}}_i)^\gamma$ can suppress the plausible result of unmatched queries induced by modest predicted masks.

With the adjusted probability-mask pair, we introduce an object-oriented ranking objective to directly model the segmentation competition between matched and unmatched queries. This objective requires that the adjusted probability-mask score of matched query should be notably higher than that of unmatched queries over the foreground region of target. Given target object (c_k, M_k) and the adjusted probability-mask pair of matched query $((\hat{q}_i * \hat{\mathbf{p}}_i)^\gamma, \hat{M}_i)$, the probability-mask score map of i-th query for k-th target object is formulated as

$$\hat{S}_i = (\hat{q}_i * \hat{\mathbf{p}}_i(c_k))^\gamma * \hat{M}_i \tag{3}$$

where $\hat{\mathbf{p}}_i(c_k)$ is the c_k-th value of $\hat{\mathbf{p}}_i$. Similarly, denoting by the probability-mask score map of unmatched query for k-th target object $\hat{S}_j = ((\hat{q}_j * \hat{\mathbf{p}}_j(c_k))^\gamma, \hat{M}_j)$, the ranking loss between the score maps of matched query $\hat{S}_i$ and unmatched query $\hat{S}_j$ with respect to k-th target object is formulated as:

$$\mathcal{L}'_{\text{rank}} = \frac{1}{|M_k|} \sum (\textit{softplus}\left(\hat{S}_j - \hat{S}_i\right) * M_k), \tag{4}$$

Then, the overall ranking loss $\mathcal{L}_{\text{rank}}$ is the average of $\mathcal{L}'_{\text{rank}}$ over all target objects and their corresponding hard negatives. In our implementation, we select the top N_H unmatched queries according to the order of $\hat{S}_j$. Softplus provides smooth gradients to continuously enlarge the score gap between matched queries and hard negatives. The gradients with respect to $\hat{\mathbf{p}}$ and $\hat{M}$ are detached to prevent over-suppression of the class probabilities and mask predictions.

3.3 Cross-Layer Self-Distillation

During training, predictions from multiple layers receive auxiliary supervision, whereas inference only retains the prediction of final layer. Although cascaded supervision allows different layers to capture complementary information, valuable intermediate representations may not be fully preserved in later layers (Liu et al., 2023; Zhang et al., 2023), as illustrated in Fig. 1. Therefore, we propose the strategy of Cross-Layer Self-Distillation (CLSD) to extract high-quality intermediate knowledge as teacher supervision for the final layer.

For a target object (c_k, M_k), the Hungarian matching algorithm is applied at each ℓ layer, where $\ell \in [L-3, L]$ and L is the index of final layer. Let i^ℓ denotes the index of matched query at ℓ-th layer, which may be different in various layers. Among all matched queries in intermediate layers, we select the layer with the highest soft IoU as the teacher candidate:

$$\ell^\star = \underset{L-3 \le \ell < L}{\arg\max}\ \textit{soft_IoU}(\, M_k, \hat{M}_{i^\ell}). \tag{5}$$

Considering that i^{ℓ^*}-th query (at ℓ^*-layer) and i^L-th query (at L-layer) are all matched queries of target object (c_k, M_k) at different layers, we select the real teacher supervision if the *soft_IoU*$(\, M_k, \hat{M}_{i^{\ell^*}})$ of i^{ℓ^*}-th query (at ℓ^*-layer) is better than that of i^L-th query (at L-layer). In this way, we design the loss function of Cross-Layer Self-Distillation for target object (c_k, M_k):

$$\mathcal{L}'_{\text{distill}} = \frac{1}{|M_k|} \sum \left[(sg(\hat{S}_{i^{\ell^*}}) - \hat{S}_{i^L}) * M_k \right]_+ \tag{6}$$

where $[\cdot]_+ = \max(0, \cdot)$ and $sg(\cdot)$ denotes the stop-gradient operator. The hinge function is activated when the "soft_IoU-qualified" teacher provides stronger adjusted probability-mask score map than that of final layer. Analogously, the overall distillation loss $\mathcal{L}_{\text{distill}}$ is average of $\mathcal{L}'_{\text{distill}}$ over all target objects. Note that CLSD applied on training stage brings no additional cost in inference.

3.4 Training and Inference

The overall objective combines the augmented segmentation loss with adjusted probability-mask ranking loss and self-distillation loss

$$\mathcal{L} = \mathcal{L}_{\text{aug_seg}} + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}} + \lambda_{\text{distill}}\mathcal{L}_{\text{distill}}. \tag{7}$$

where $\mathcal{L}_{\text{aug_seg}}$ is the augmented segmentation loss, including object classification, pixel-wise mask prediction, Dice and proposed mask-quality regression losses. The coefficients λ_{rank} and λ_{distill}

Table 1: **Panoptic segmentation on COCO *val2017* and ADE20K *val*.** Blue rows apply iFAN to the corresponding PMT or EoMT baseline; ↑ gives the absolute PQ improvement. † denotes shortest-side resizing, ‡ re-implementation, and $_c$ COCO panoptic pre-training.

Method	Backbone	Pre-training	Params	COCO *val2017* (Lin et al., 2014)				ADE20K *val* (Zhou et al., 2017)			
				Input	GFLOPs	FPS	PQ	Input	GFLOPs	FPS	PQ
Mask2Former†	Swin-L	IN21K	216M	800^2	868	24	57.8	640^2	–	33	48.1
kMaX-DeepLab	ConvNeXt-L	IN21K	232M	1281^2	–	–	58.0	1281^2	1302	–	50.9
OneFormer†	DiNAT-L	IN21K	223M	800^2	736	20	58.0	1280^2	1369	10	53.5_c
MaskDINO†	Swin-L	IN21K	223M	800^2	1326	14	58.3	–	–	–	–
OneFormer+ViT-P†	DiNAT-L	IN21K	309M	1024^2	1013	8	58.0	1280^2	1955	3	54.0_c
Mask2Former‡	ViT-AL‡	DINOv2	349M	640^2	830	29	57.1	640^2	830	29	51.8_c
Mask2Former‡	ViT-AL‡	DINOv2	354M	1280^2	4817	10	59.7	1280^2	4817	10	53.0_c
Mask2Former‡	ViT-Adapter-g‡	DINOv2	1209M	640^2	2510	20	57.7	640^2	2510	20	52.6_c
Mask2Former‡	ViT-Adapter-g‡	DINOv2	1216M	1280^2	13790	6	59.9	1280^2	13790	6	54.2_c
PMT	ViT-L	DINOv3	357M	640^2	767	141	56.1	640^2	767	141	$49.4^‡$
PMT-iFAN	ViT-L	DINOv3	357M	640^2	**767**	**141**	**56.7**$^{\uparrow 0.6}$	640^2	**767**	**141**	**50.7**$^{\uparrow 1.3}$
PMT	ViT-L	DINOv3	357M	1280^2	4925	29	58.1	1280^2	4925	29	$50.5^‡$
PMT-iFAN	ViT-L	DINOv3	357M	1280^2	**4925**	**29**	**58.6**$^{\uparrow 0.5}$	1280^2	**4925**	**29**	**53.0**$^{\uparrow 2.5}$
EoMT	ViT-L	DINOv2	316M	640^2	669	128	56.0	640^2	669	128	50.6_c
EoMT-iFAN	ViT-L	DINOv2	316M	640^2	**669**	**128**	**57.0**$^{\uparrow 1.0}$	640^2	**669**	**128**	**52.2**$_c^{\uparrow 1.6}$
EoMT	ViT-L	DINOv2	322M	1280^2	4146	30	58.3	1280^2	4146	30	51.7_c
EoMT-iFAN	ViT-L	DINOv2	322M	1280^2	**4146**	**30**	**58.8**$^{\uparrow 0.5}$	1280^2	**4146**	**30**	**53.9**$_c^{\uparrow 2.2}$
EoMT	ViT-G	DINOv2	1164M	640^2	2261	55	57.0	640^2	2261	55	51.3_c
EoMT-iFAN	ViT-G	DINOv2	1164M	640^2	**2261**	**55**	**57.8**$^{\uparrow 0.8}$	640^2	**2261**	**55**	**52.7**$_c^{\uparrow 1.4}$
EoMT	ViT-G	DINOv2	1171M	1280^2	12712	12	59.2	1280^2	12712	12	52.8_c
EoMT-iFAN	ViT-G	DINOv2	1171M	1280^2	**12712**	**12**	**59.6**$^{\uparrow 0.4}$	1280^2	**12712**	**12**	**53.9**$_c^{\uparrow 1.1}$

balance the contributions of probability-mask score ranking and cross-layer self-distillation objectives, which are set to 0.1 and 0.4, respectively. The epoch-dependent coefficient λ_{distill} gradually activates selective depth distillation after an initial warm-up stage and anneals near convergence stage, mitigating the influence of unreliable early teachers and avoiding excessive constraints on the final representation. During inference, we extract the adjusted probability-mask score map $\hat{S}_i = (\hat{q}_i * max(\hat{\mathbf{p}}_\mathbf{i}))^\gamma * \hat{M}_i$ for i-th query at the final layer for segmentation competition.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. We evaluate iFAN on panoptic, instance, and semantic segmentation. COCO val2017 (Lin et al., 2014) is used for panoptic and instance segmentation, ADE20K val (Zhou et al., 2017) for panoptic and semantic segmentation, and Cityscapes val (Cordts et al., 2016) for semantic segmentation. Following the evaluation protocols, we report Panoptic Quality (PQ) (Kirillov et al., 2019), mask Average Precision (AP) (Lin et al., 2014), and mean Intersection over Union (mIoU) (Everingham et al., 2010), respectively.

Implementation We implement iFAN on two plain mask-transformer baselines: EoMT (Kerssies et al., 2025) uses DINOv2-pretrained ViT-S/B/L/G backbones, while PMT (Cavagnero et al., 2026) uses DINOv3-pretrained ViT-L. Models are trained on two NVIDIA B200 GPUs with a batch size of 4, using mixed-precision training and AdamW (Loshchilov & Hutter, 2017). We use an initial learning rate of 10^{-4}, layer-wise learning-rate decay of 0.8, and polynomial learning-rate decay with a power of 0.9. Training lasts 12 epochs on COCO, 31 epochs on ADE20K, and 56 epochs on Cityscapes. We set the number of hard negatives N_H to 10. CLSD starts at epoch 4 and is linearly annealed thereafter. Unless stated otherwise, all ablations use EoMT-iFAN with ViT-L and 640^2 inputs on COCO.

4.2 Main Results

Table 2: **Semantic segmentation on Cityscapes and ADE20K *val*.** Blue rows apply iFAN to the corresponding PMT or EoMT baseline, and ↑ gives the absolute mIoU improvement. † denotes ADE20K shortest-side resizing, ‡ re-implementation, and $_c$ COCO panoptic pre-training. PMT and EoMT use DINOv3- and DINOv2-pretrained backbones, respectively.

Method	Backbone	Pre-training	Params	Cityscapes *val* (Cordts et al., 2016)				ADE20K *val* (Zhou et al., 2017)			
				Input	GFLOPs	FPS	mIoU	Input	GFLOPs	FPS	mIoU
Mask2Former†	Swin-L	IN21K	216M	1024×2048	–	14	83.3	640^2	–	33	56.1
MaskDINO†	Swin-L	IN21K	223M	–	–	–	–	640^2	–	–	56.6
OneFormer†	ConvNeXt-XL	IN21K	373M	1024×2048	775	7	83.6	640^2	607	21	57.4
OneFormer†	DiNAT-L	IN21K	223M	1024×2048	450	14	83.1	896^2	678	19	58.1
kMaX-DeepLab	ConvNeXt-L	IN21K	232M	1025×2049	1673	–	83.5	–	–	–	–
Mask2Former	ViT-L	DINOv2	–	896×1792	–	–	84.8	896^2	–	–	59.4
OneFormer+ViT-P‡	DiNAT-L‡	IN21K	309M	–	–	–	–	1280^2	1955	5	59.9_c
Mask2Former‡	ViT-AL‡	DINOv2	351M	1024^2	5200	7	84.5	512^2	910	21	58.9
PMT	ViT-L	DINOv3	357M	1024^2	5113	27	83.1‡	512^2	823	128	58.5
PMT-iFAN	ViT-L	DINOv3	357M	1024^2	**5113**	**27**	**83.8**$^{\uparrow 0.7}$	512^2	**823**	**128**	**59.4**$^{\uparrow 0.9}$
EoMT	ViT-L	DINOv2	319M	1024^2	4350	25	84.2	512^2	721	92	58.4
EoMT-iFAN	ViT-L	DINOv2	319M	1024^2	**4350**	**25**	**84.5**$^{\uparrow 0.3}$	512^2	**721**	**92**	**59.0**$^{\uparrow 0.6}$

iFAN on different benchmarks. As shown in Table 1, iFAN consistently improves PMT and EoMT with virtually no additional parameters, FLOPs, or inference overhead, narrowing their accuracy gap to non-PMT methods. On average, iFAN improves PMT by 0.55/1.90 PQ and EoMT by 0.68/1.58 PQ on COCO/ADE20K. At 1280^2, EoMT-iFAN with ViT-G reaches 59.6 PQ on COCO and 53.9 PQ on ADE20K, only 0.3 PQ below the best non-PMT results, while running at 12 FPS versus 6 FPS, i.e., $2\times$ faster. Thus, iFAN brings PMT-style models close to state-of-the-art accuracy while preserving their efficiency.

Table 3: **Instance segmentation on COCO *val2017*.** Blue rows apply iFAN to the corresponding PMT or EoMT baseline, and ↑ gives the absolute AP improvement. † denotes shortest-side resizing and ‡ re-implementation. PMT and EoMT use DINOv3- and DINOv2-pretrained backbones, respectively.

Method	Backbone	Params	Input	FPS	AP
OneFormer†	DiNAT-L	223M	800^2	20	49.2
Mask2Former†	Swin-L	216M	800^2	24	50.1
MaskDINO†	DiNAT-L	223M	800^2	14	52.3
OneFormer+ViT-P†	DiNAT-L	309M	1024^2	4	49.5
Mask2Former‡	ViT-AL‡	349M	640^2	29	47.6
Mask2Former‡	ViT-AL‡	354M	1280^2	10	51.4
PMT	ViT-L	357M	640^2	141	45.4
PMT-iFAN	ViT-L	357M	640^2	**141**	**46.6**$^{\uparrow 1.2}$
PMT	ViT-L	357M	1280^2	29	49.3
PMT-iFAN	ViT-L	357M	1280^2	**29**	**50.6**$^{\uparrow 1.3}$
EoMT	ViT-L	316M	640^2	128	44.8
EoMT-iFAN	ViT-L	316M	640^2	**128**	**46.0**$^{\uparrow 1.2}$
EoMT	ViT-L	322M	1280^2	30	48.8
EoMT-iFAN	ViT-L	322M	1280^2	**30**	**50.6**$^{\uparrow 1.8}$

As shown in Table 3, iFAN improves PMT by 1.25 AP and EoMT by 1.50 AP on average. EoMT-iFAN reaches 50.6 AP at 30 FPS, approaching the strongest non-PMT result of 52.3 AP while being $2.1\times$ faster (30 vs. 14 FPS). This again shows that iFAN closes the accuracy gap without losing the throughput advantage of plain mask transformers.

Table 2 shows the same trend. On Cityscapes, EoMT-iFAN reaches 84.5 mIoU, matching the best non-PMT method with reported throughput while running $3.6\times$ faster (25 vs. 7 FPS). On ADE20K, PMT-iFAN achieves 59.4 mIoU, only 0.5 below the best non-PMT result, yet runs at 128 FPS versus 5 FPS ($25.6\times$). Overall, iFAN consistently pushes efficient PMT-style models toward non-PMT state-of-the-art accuracy across segmentation tasks.

4.3 Hyperparameter Sensitivity

Figure 3 evaluates the sensitivity of iFAN with EoMT-L and PMT-L to the loss weights of the APMR and CLSD objectives. Both models remain robust across a broad range of hyperparameter values. We therefore adopt representative values within the stable regions, namely $\lambda_{\text{rank}} = 0.10$ and $\lambda_{\text{distill}} = 0.40$, and keep them fixed across all datasets, backbones, input resolutions, and segmentation tasks.

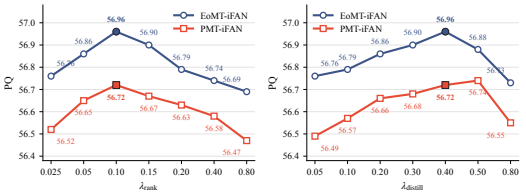


Figure 3: **Hyperparameter sensitivity.** Left: sensitivity to the APMR weight λ_{rank}. Right: sensitivity to the CLSD weight λ_{distill}. Each loss weight is varied independently with ViT-L and 640^2 inputs, while all other settings are fixed. Filled markers indicate the selected values.

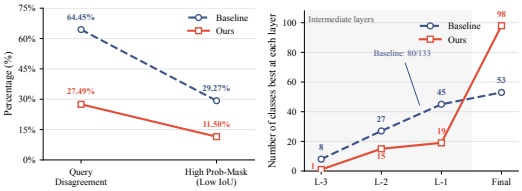


Figure 4: **Inference diagnostics on COCO *val2017*.** Left: query-selection disagreement and high-score, low-IoU predictions. Right: the numbers of classes whose best predictions occur at each layer. Both panels compare EoMT with iFAN; lower disagreement and more final-layer optima indicate better inference alignment.

4.4 Ablation Studies

Overall Ablation. Table 4(a) isolates the global mask-quality branch (Quality), APMR, and CLSD; checkmarks enable the corresponding components, and crosses disable them. The EoMT baseline obtains 56.00 PQ. Adding mask-quality learning yields 56.43 PQ; further introducing ranking raises performance to 56.68 PQ, whereas combining quality learning with CLSD obtains 56.54 PQ. The complete iFAN model achieves 57.00 PQ. Thus, ranking suppresses competitive unmatched queries within the final prediction set, while CLSD transfers better adjusted probability-mask responses from intermediate layers to the final layer.

Ablation on Cross-Layer Self-Distillation. Table 4(b) studies best-teacher selection strategy (BT) and $\hat{S}$-map distillation. A checkmark under BT selects the intermediate teacher with the highest soft-IoU (Eq. 5); a cross uses the layer immediately preceding the final layer. When distillation is enabled, a checkmark under $\hat{S}$ Distill. uses the adjusted probability-mask score map $\hat{S}$ as the distillation target, whereas a cross uses soft-IoU. Without self-distillation, the model obtains 56.00 PQ. Naively distilling soft-IoU from the preceding layer reduces the result to 55.96 PQ, indicating that the teacher from the preceding layer does not necessarily perform better than the output from the final layer, and a low-quality teacher may even negatively affect an otherwise correct result. Enabling BT increases the result to 56.39 PQ, while using $\hat{S}$-map distillation with the preceding-layer teacher gives 56.02 PQ. Combining best-teacher selection with $\hat{S}$-map distillation reaches 56.54 PQ, 0.58 points above naive distillation and 0.15 points above best-teacher selection alone. Teacher selection therefore removes the main source of noisy cross-layer supervision, and $\hat{S}$-map distillation provides an additional benefit by transferring the same score map used for inference-time competition.

Figure 4 (right) provides a depth-wise diagnostic. For EoMT, the final layer is optimal for 53 of the 133 classes, whereas an intermediate layer is optimal for the remaining 80. With CLSD, the number of final-layer-optimal classes increases to 98 and the intermediate-layer count decreases to 35. This shift indicates that CLSD more effectively preserves useful intermediate predictions in the final layer.

Ablation on Adjusted Probability-Mask Ranking. Table 5(a) incrementally evaluates the ranking design using EoMT with ViT-L at 640^2. The variants cumulatively enable the APMR loss, its computation on the adjusted probability-mask score map $\hat{S}_i$, and stop-gradient (SG) on the class-probability and mask branches. Adding the ranking objective to the 56.00-PQ baseline reaches 56.26 PQ. Computing the ranking loss on the adjusted probability-mask score map $\hat{S}$ further improves PQ to 56.45. Finally, stopping the ranking gradients to the class-probability and mask branches, so that this objective updates the global quality score, yields 56.68 PQ. This final 0.23 point gain supports the design choice in the APMR section: ranking should calibrate the quality used in segmentation competition without disrupting the original classification and mask objectives.

The left panel of Figure 4 measures disagreement between the *max-prob-mask* and *max-IoU* queries, as well as the frequency of high probability-mask scores paired with low IoU. The results are consis-

Table 4: **Overall iFAN and CLSD ablations** on COCO *val2017* using ViT-L at 640^2. (a) Contributions of global mask-quality learning (Quality), APMR, and CLSD. (b) Effects of best-teacher selection (BT) and adjusted probability-mask distillation ($\hat{S}$ Distill.).

(a) Overall module ablation

Quality	APMR	CLSD	PQ
×	×	×	56.00
✓	×	×	56.43
✓	✓	×	56.68
✓	×	✓	56.54
✓	✓	✓	**57.00**

(b) CLSD component analysis

$\mathcal{L}_{\text{distill}}$	BT	$\hat{S}$ Distill.	PQ
×	×	×	56.00
✓	×	×	55.96
✓	✓	×	56.39
✓	×	✓	56.02
✓	✓	✓	**56.54**

Table 5: **APMR and model-size ablations** on COCO using 640^2 inputs. (a) Contributions of the ranking loss, adjusted score map, and stop-gradient (SG) with ViT-L. (b) PQ across model sizes with DINOv2 pre-training. Detailed settings are described in the text.

(a) APMR component analysis

$\mathcal{L}_{\text{rank}}$	$\hat{S}$ Ranking	SG	PQ
×	×	×	56.00
✓	×	×	56.26
✓	✓	×	56.45
✓	✓	✓	**56.68**

(b) Effect of model size

Size	Params	EoMT	iFAN
ViT-S	24M	44.7	**46.3**
ViT-B	93M	50.6	**51.8**
ViT-L	316M	56.0	**57.0**
ViT-G	1164M	57.0	**57.8**

tent with the ablation. iFAN reduces the disagreement rate between the *max-prob-mask* and *max-IoU* queries from 64.45% to 27.49%, and reduces the proportion of high probability-mask scores with low IoU from 29.27% to 11.50%. APMR therefore makes the query selected for inference more consistent with the query that best overlaps the target.

Reliability of mask-quality prediction. We evaluate the reliability of global mask-quality prediction with EoMT-iFAN on all 5,000 images in COCO *val2017*. Specifically, we compute the pooled Spearman rank correlation between the predicted quality score $\hat{q}$ and the soft IoU with the ground-truth mask over all class-consistent query–target pairs. The quality head achieves a correlation of $\rho = 0.726$ over 63,935 pairs. After applying the calibrated class–quality score threshold of 0.60, the correlation remains high at $\rho = 0.710$ over the remaining 49,005 pairs. These results indicate that the lightweight quality head provides a reliable global ranking of mask quality and therefore serves as an informative quality signal for adjusted probability-mask scoring. See the appendix A.4 for details.

Effect of model size. Table 5(b) evaluates EoMT-iFAN from ViT-S to ViT-G using 640^2 inputs and DINOv2 pre-training; all results report PQ. iFAN improves all backbones by 1.6, 1.2, 1.0, and 0.8 PQ, respectively. The gain decreases as the backbone scales, suggesting that smaller models leave more probability-mask competition errors for iFAN to correct; the improvement remains consistent even for ViT-G.

5 Conclusion

We presented iFAN, an inference-aware learning framework for plain mask transformers that improves both query competition and cross-layer knowledge transfer. APMR augments the probability-mask score with global mask quality and explicitly ranks matched queries above hard unmatched competitors, making inference-time query selection more consistent with mask quality. CLSD identifies a better intermediate prediction for each target and distills its adjusted probability-mask score map to the final layer, without requiring multi-layer inference. Experiments on COCO, ADE20K, and Cityscapes show consistent improvements across panoptic, instance, and semantic segmentation, two plain mask-transformer baselines, multiple backbone sizes, and different input resolutions. Importantly, iFAN introduces negligible additional parameters, FLOPs, or inference overhead, preserving the efficiency advantage of plain mask transformers while bringing their accuracy close to strong non-PMT methods. Diagnostic results further show that iFAN reduces the disagreement between the max-prob-mask and max-IoU queries from 64.45% to 27.49%, while increasing the number of classes for which the final layer achieves the best performance from 53 to 98. These results demonstrate that iFAN effectively alleviates cross-query competition and preserves valuable intermediate information in final predictions, while maintaining efficient inference.

AI USE STATEMENT

Generative AI tools were used to polish the manuscript and assist with simple utility code and tests. The authors reread the AI-assisted text and inspected the AI-assisted code, incorporating outputs into the manuscript or Git repository only after checking them and finding no issues. Research design, interpretation of results, and mathematical reasoning and proofs were conducted by the authors without generative-AI assistance. The authors take responsibility for the final text, code, claims, and results.

REFERENCES

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Niccolò Cavagnero, Narges Norouzi, Gijs Dubbelman, and Daan de Geus. Pmt: Plain mask transformer for image and video segmentation with frozen vision encoders. *arXiv preprint arXiv:2603.25398*, 2026.

Jiahao Chang, Shuo Wang, Hai-Ming Xu, Zehui Chen, Chenhongyi Yang, and Feng Zhao. Detrdistill: A universal knowledge distillation framework for detr-families. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6898–6908, 2023.

Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 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*, pp. 1290–1299, 2022.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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.

Jie Hu, Chen Chen, Liujuan Cao, Shengchuan Zhang, Annan Shu, Guannan Jiang, and Rongrong Ji. Pseudo-label alignment for semi-supervised instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16337–16347, 2023.

Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6409–6418, 2019.

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*, pp. 2989–2998, 2023.

Lei Ke, Martin Danelljan, Xia Li, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Mask transfiner for high-quality instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4412–4421, 2022.

Tommie Kerssies, Niccolo Cavagnero, Alexander Hermans, Narges Norouzi, Giuseppe Averta, Bastian Leibe, Gijs Dubbelman, and Daan De Geus. Your vit is secretly an image segmentation model. In *Proceedings of the computer vision and pattern recognition conference*, pp. 25303–25313, 2025.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9404–9413, 2019.

Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pp. 562–570. Pmlr, 2015.

Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2239–2251, 2023a.

Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3041–3050, 2023b.

Zhiqi Li, Wenhai Wang, Enze Xie, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, Ping Luo, and Tong Lu. Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1280–1289, 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.

Shilong Liu, Tianhe Ren, Jiayu Chen, Zhaoyang Zeng, Hao Zhang, Feng Li, Hongyang Li, Jun Huang, Hang Su, Jun Zhu, et al. Detection transformer with stable matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6491–6500, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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. *Transactions on Machine Learning Research Journal*, 2024.

Yifan Pu, Weicong Liang, Yiduo Hao, Yuhui Yuan, Yukang Yang, Chao Zhang, Han Hu, and Gao Huang. Rank-detr for high quality object detection. *Advances in Neural Information Processing Systems*, 36:16100–16113, 2023.

Sajjad Shahabodini, Mobina Mansoori, Farnoush Bayatmakou, Jamshid Abouei, Konstantinos Plataniotis, and Arash Mohammadi. The missing point in vision transformers for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6314–6324, June 2026.

Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5463–5474, 2021.

Yu Wang, Xin Li, Shengzhao Weng, Gang Zhang, Haixiao Yue, Haocheng Feng, Junyu Han, and Errui Ding. Kd-detr: Knowledge distillation for detection transformer with consistent distillation points sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16016–16025, 2024.

Shengjian Wu, Li Sun, and Qingli Li. Od-detr: online distillation for stabilizing training of detection transformer. *arXiv preprint arXiv:2406.05791*, 2024.

Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. k-means mask transformer. In *European conference on computer vision*, pp. 288–307. Springer, 2022.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *The eleventh international conference on learning representations*, 2022.

Hao Zhang, Feng Li, Huaizhe Xu, Shijia Huang, Shilong Liu, Lionel M Ni, and Lei Zhang. Mp-former: Mask-piloted transformer for image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18074–18083, 2023.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3713–3722, 2019.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.

A SUPPLEMENTARY MATERIAL

Table 6: **Detailed panoptic segmentation results for PMT and EoMT on COCO *val2017* and ADE20K *val*.** th and st denote things and stuff, respectively. ΔPQ compares each iFAN model with its matched baseline. Blue rows denote iFAN. Parameters, GFLOPs, and FPS in iFAN rows follow the matched baseline under shared-configuration accounting; "–" indicates that a metric is unavailable.

Method	Dataset	Backbone	Input	Params	GFLOPs	FPS	PQ	RQ	SQ	PQ_{th}	RQ_{th}	SQ_{th}	PQ_{st}	RQ_{st}	SQ_{st}	ΔPQ
PMT	COCO	ViT-L	640^2	357M	767	141	56.1	67.0	82.9	61.5	73.0	83.6	48.1	58.0	81.7	–
PMT-iFAN	COCO	ViT-L	640^2	357M	**767**	**141**	**56.7**	67.2	83.5	62.3	73.5	84.2	48.2	57.6	82.4	**+0.6**
PMT	COCO	ViT-L	1280^2	357M	4925	29	58.1	68.7	83.7	64.6	75.7	84.8	48.3	58.1	81.9	–
PMT-iFAN	COCO	ViT-L	1280^2	357M	**4925**	**29**	**58.6**	69.0	84.1	65.0	76.0	85.1	48.9	58.4	82.7	**+0.5**
EoMT	COCO	ViT-L	640^2	316M	669	128	56.0	67.2	82.5	61.2	73.1	83.1	48.2	58.2	81.5	–
EoMT-iFAN	COCO	ViT-L	640^2	316M	**669**	**128**	**57.0**	68.0	83.0	62.1	73.7	83.6	49.3	59.3	82.1	**+1.0**
EoMT	COCO	ViT-L	1280^2	322M	4146	30	58.3	69.0	83.5	64.7	75.9	84.7	48.6	58.6	81.8	–
EoMT-iFAN	COCO	ViT-L	1280^2	322M	**4146**	**30**	**58.8**	69.5	83.8	65.2	76.5	84.7	49.2	58.9	82.4	**+0.5**
EoMT	COCO	ViT-G	640^2	1164M	2261	55	57.0	68.2	82.7	62.6	74.5	83.5	48.6	58.8	81.4	–
EoMT-iFAN	COCO	ViT-G	640^2	1164M	**2261**	**55**	**57.8**	68.8	83.2	63.5	75.4	83.8	49.1	58.8	82.4	**+0.8**
EoMT	COCO	ViT-G	1280^2	1171M	12712	12	59.2	69.8	83.8	66.0	77.2	85.1	48.8	58.8	81.8	–
EoMT-iFAN	COCO	ViT-G	1280^2	1171M	**12712**	**12**	**59.6**	70.4	83.9	66.3	77.6	85.0	49.6	59.5	82.3	**+0.4**
PMT	ADE20K	ViT-L	640^2	357M	767	141	49.4	58.9	82.7	48.3	57.9	82.3	52.0	60.8	83.5	–
PMT-iFAN	ADE20K	ViT-L	640^2	357M	**767**	**141**	**50.7**	–	–	50.4	–	–	51.2	–	–	**+1.3**
PMT	ADE20K	ViT-L	1280^2	357M	4925	29	50.5	59.6	83.2	50.5	59.6	83.7	51.4	59.6	82.3	–
PMT-iFAN	ADE20K	ViT-L	1280^2	357M	**4925**	**29**	**53.0**	–	–	53.7	–	–	51.6	–	–	**+2.5**
EoMT	ADE20K	ViT-L	640^2	316M	669	128	50.6	59.9	82.4	49.7	59.2	82.2	52.3	61.5	82.8	–
EoMT-iFAN	ADE20K	ViT-L	640^2	316M	**669**	**128**	**52.2**	62.0	81.3	51.8	61.9	81.9	53.1	62.1	80.1	**+1.6**
EoMT	ADE20K	ViT-L	1280^2	322M	4146	30	51.7	60.4	82.5	52.0	61.2	82.5	51.0	58.9	82.6	–
EoMT-iFAN	ADE20K	ViT-L	1280^2	322M	**4146**	**30**	**53.9**	63.2	84.6	53.7	63.3	84.1	54.1	62.9	85.5	**+2.2**
EoMT	ADE20K	ViT-G	640^2	1164M	2261	55	51.3	60.7	83.3	50.5	59.9	82.6	53.0	62.3	84.6	–
EoMT-iFAN	ADE20K	ViT-G	640^2	1164M	**2261**	**55**	**52.7**	62.3	83.2	52.3	62.2	83.1	53.6	62.7	83.4	**+1.4**
EoMT	ADE20K	ViT-G	1280^2	1171M	12712	12	52.8	61.3	84.0	52.7	61.6	83.9	52.6	60.8	84.1	–
EoMT-iFAN	ADE20K	ViT-G	1280^2	1171M	**12712**	**12**	**53.9**	63.0	84.3	54.0	63.2	83.8	53.7	62.6	85.1	**+1.1**

A.1 MORE ABLATION

Figure 5 evaluates the sensitivity to the number of hard negatives N_H and the weighting exponent γ. Performance remains consistently strong for $N_H \in [7, 15]$. Similarly, varying γ from 0.5 to 3 results in only a 0.18 PQ difference. The setting $\gamma = 2$ effectively suppresses low-valued responses while avoiding the excessive attenuation caused by a larger exponent. We therefore use $N_H = 10$ and $\gamma = 2$ as the default settings.

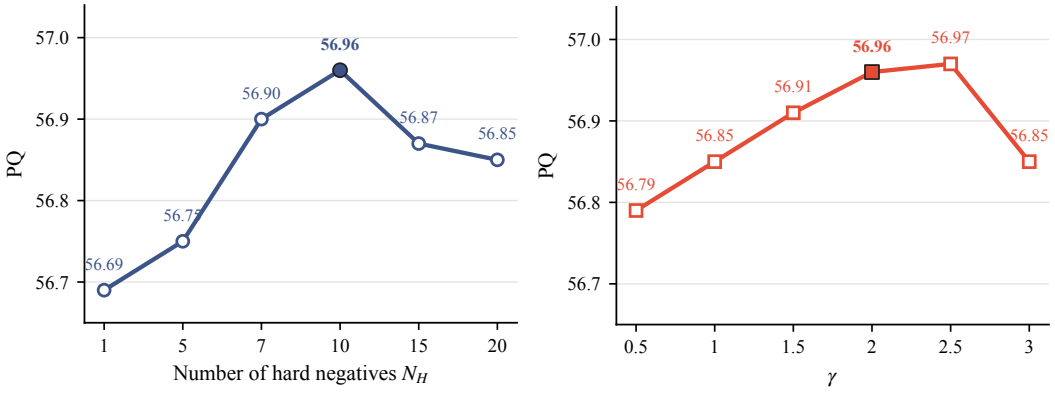


Figure 5: **Hyperparameter sensitivity on COCO panoptic segmentation.** Left: performance remains stable across a broad range of N_H. Right: γ balances the suppression of low-valued responses against the preservation of useful responses. Filled markers denote the selected settings, $N_H = 10$ and $\gamma = 2$.

A.2 DETAILED PMT AND EOMT RESULTS

Tables 6–8 provide more detailed results of EoMT, PMT, and iFAN for panoptic, semantic, and instance segmentation.

Table 7: **Detailed semantic segmentation results for PMT and EoMT on Cityscapes and ADE20K *val*.** Block $-k$ gives the validation mIoU predicted from the k-th block counted backward from the final block. ΔmIoU compares each iFAN model with its matched baseline. Blue rows denote iFAN, and "–" indicates that a metric was not logged for that configuration.

Method	Dataset	Backbone	Input	Params	GFLOPs	FPS	mIoU	Block -1	Block -2	Block -3	Block -4	ΔmIoU
EoMT	Cityscapes	ViT-L	1024^2	319M	4350	25	84.2	–	–	–	–	–
EoMT-iFAN	Cityscapes	ViT-L	1024^2	319M	**4350**	**25**	**84.5**	84.5	84.4	82.6	74.3	**+0.3**
PMT	ADE20K	ViT-L	512^2	357M	823	128	58.5	–	–	–	–	–
PMT-iFAN	ADE20K	ViT-L	512^2	357M	**823**	**128**	**59.4**	–	–	–	–	**+0.9**
EoMT	ADE20K	ViT-L	512^2	319M	721	92	58.4	–	–	–	–	–
EoMT-iFAN	ADE20K	ViT-L	512^2	319M	**721**	**92**	**59.0**	58.5	58.2	57.5	14.1	**+0.6**

Table 8: **Detailed local instance segmentation results for PMT and EoMT on COCO *val2017*.** ΔAP compares each iFAN model with its matched baseline; EoMT deltas are computed from the unrounded records before rounding. Blue rows denote iFAN. Parameters and FPS in iFAN rows follow the matched baseline under shared-configuration accounting.

Method	Backbone	Input	Params	FPS	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Epoch	ΔAP
PMT	ViT-L	640^2	357M	141	45.4	69.5	48.2	21.0	50.9	72.3	11	–
PMT-iFAN	ViT-L	640^2	357M	**141**	**46.6**	69.8	50.2	21.7	51.2	72.5	10	**+1.2**
EoMT	ViT-L	640^2	316M	128	44.8	69.3	47.4	19.7	50.6	72.2	–	–
EoMT-iFAN	ViT-L	640^2	316M	**128**	**46.0**	69.7	48.9	20.4	50.6	71.9	–	**+1.2**
PMT	ViT-L	1280^2	357M	29	49.3	73.5	53.5	28.7	53.9	72.9	11	–
PMT-iFAN	ViT-L	1280^2	357M	**29**	**50.6**	74.0	55.1	28.8	54.6	72.8	10	**+1.3**
EoMT	ViT-L	1280^2	322M	30	48.8	73.5	52.5	27.4	54.1	72.1	–	–
EoMT-iFAN	ViT-L	1280^2	322M	**30**	**50.6**	74.2	54.8	29.6	54.7	72.9	–	**+1.8**

A.3 Score–Quality Alignment in Query Competition

This section details the query-wise diagnostic summarized in Fig.1 (top) of the Introduction. It tests whether the query-level score that determines probability-mask competition is consistent with the actual mask quality of that query, thereby isolating the inference mismatch targeted by Adjusted Probability-Mask Ranking (APMR).

Protocol. We evaluate EoMT and EoMT-iFAN on all 5,000 images of COCO *val2017*. For a given image, let $\{(c_j, M_j)\}_{j=1}^{N_{\text{gt}}}$ denote its ground-truth segments, where c_j and M_j are the class label and binary mask of segment j. Query i outputs a class-probability distribution $\hat{\mathbf{p}}_i$ over the foreground classes and the no-object class $\varnothing$, together with a soft mask $\hat{M}_i \in [0,1]^{H\times W}$. Let $\hat{c}_i = \arg\max_c \hat{\mathbf{p}}_i(c)$ denote its predicted class. We first compute its soft-IoU with every ground-truth segment:

$$s_{ij} = \textit{soft_IoU}(\hat{M}_i, M_j) \tag{8}$$

The sums are taken over all image pixels; unlike thresholded IoU, Eq. equation 8 retains the predicted mask probabilities.

To exclude queries with essentially no spatial support on any annotated segment, we define the evaluated query set as

$$\mathcal{Q}_{\text{eval}} = \left\{ i \,\middle|\, \hat{c}_i \neq \varnothing,\ \max_{1\le j\le N_{\text{gt}}} s_{ij} > 0.01 \right\}. \tag{9}$$

This initial overlap test is class-agnostic and serves only as a candidate filter. It does not select the ground-truth segment used to define mask quality.

For each $i \in \mathcal{Q}_{\text{eval}}$, let $\mathcal{J}_i = \{j \mid c_j = \hat{c}_i\}$ be the set of ground-truth segments whose class matches the predicted class of query i. We define the query's reference **mask quality** as

$$u_i = \begin{cases} \max_{j\in\mathcal{J}_i} s_{ij}, & \mathcal{J}_i \neq \varnothing, \\ 0, & \mathcal{J}_i = \varnothing. \end{cases} \tag{10}$$

Hence, u_i is the best soft-IoU that query i attains with any ground-truth segment of its predicted class. Overlap with a ground-truth segment of a different class cannot increase u_i. Unlike one-to-

Table 9: **Score–quality alignment on COCO *val2017*.** All images are evaluated. "Change" denotes iFAN minus EoMT; percentage changes are in percentage points.

Metric	EoMT	iFAN	Change
Pairwise ordering accuracy (%)	76.07	**79.42**	+3.35
Top-query accuracy (%)	35.55	**72.51**	+36.96
Top-query conflict (%)	64.45	**27.49**	-36.96
High-score low-quality rate (%)	29.27	**11.5**	-17.8
Mask-quality gap	0.1070	**0.0591**	-0.0479

one Hungarian assignment, this diagnostic comparison is many-to-one: multiple queries may use the same ground-truth segment as their best class-consistent reference. It therefore measures the class-consistent mask quality of every retained competing query without tying the analysis to its training-time assignment.

For EoMT, the query-level **competition score** is the predicted-class confidence

$$r_i^{\mathrm{EoMT}} = \max_{c \neq \varnothing} \hat{\mathbf{p}}_i(c). \tag{11}$$

For iFAN, the corresponding score incorporates the predicted global mask quality,

$$r_i^{\mathrm{iFAN}} = \hat{q}_i \max_{c \neq \varnothing} \hat{\mathbf{p}}_i(c). \tag{12}$$

The adjusted probability-mask score map used by iFAN is therefore $\hat{S}_i = (r_i^{\mathrm{iFAN}})^2 \hat{M}_i$. Because squaring is monotonic for these non-negative scores, it does not change the pairwise ordering or the identity of the highest-scoring query.

Metrics. We report five complementary measurements:

Pairwise ordering accuracy (higher is better): For every pair of retained queries from the same image with $|u_i - u_k| > 0.05$, this metric checks whether their competition scores have the same ordering as their reference mask qualities. It reports the percentage of such pairs for which the higher-quality query also has the higher score.

Top-query accuracy (higher is better): For each eligible image, we identify the retained query with the largest competition score and the retained query with the largest reference mask quality. This metric reports the percentage of images for which these two selections identify the same query.

Top-query conflict (lower is better): Let $i_r = \arg\max_{i \in \mathcal{Q}_{\mathrm{eval}}} r_i$ be the highest-scoring retained query. A conflict is counted when its quality is more than 0.05 below the best available quality, i.e., $\max_{i \in \mathcal{Q}_{\mathrm{eval}}} u_i - u_{i_r} > 0.05$. The metric is the percentage of eligible images containing such a conflict.

High-score low-quality rate (lower is better): Among retained queries with $r_i \geq 0.8$, this metric reports the percentage whose reference mask quality satisfies $u_i < 0.5$. It measures how often a highly confident competitor nevertheless has poor class-consistent mask quality.

Mask-quality gap (lower is better): For each eligible image, we compute $\max_{i \in \mathcal{Q}_{\mathrm{eval}}} u_i - u_{i_r}$, namely the quality lost by selecting the highest-scoring query instead of the best-quality retained query. We report its average over eligible images; a value of zero indicates no quality loss.

Results and interpretation. Pairwise ordering accuracy increases from 76.07% to 79.42%. The effect is larger at the top of the ranking: top-query accuracy increases from 35.55% to 72.51%, top-query conflict decreases from 64.45% to 27.49%, and the average mask-quality gap decreases from 0.1070 to 0.0591. The high-score low-quality rate also decreases from 29.27% to approximately 11.5%. APMR therefore improves the high-score region that determines which query dominates probability-mask competition. The statistics contain 72,619 valid queries and 1,474 eligible images for EoMT, and 72,209 queries and 1,473 images for iFAN; the small count difference results from applying the candidate criterion independently to each model.

A.4 Quality Head Output and Mask Quality

Figure 6 visualizes the association between quality head outputs and ground-truth soft-IoU on all 5,000 COCO *val2017* images using EoMT-iFAN. We retain class-consistent query–target pairs with non-background predictions, overlap between the query mask binarized at 0.5 and a target, and soft-IoU > 0.01. The Spearman correlations over all retained pairs are $\rho = 0.726$ before score filtering and $\rho = 0.710$ after filtering. These correlations describe the global association across query–target pairs.

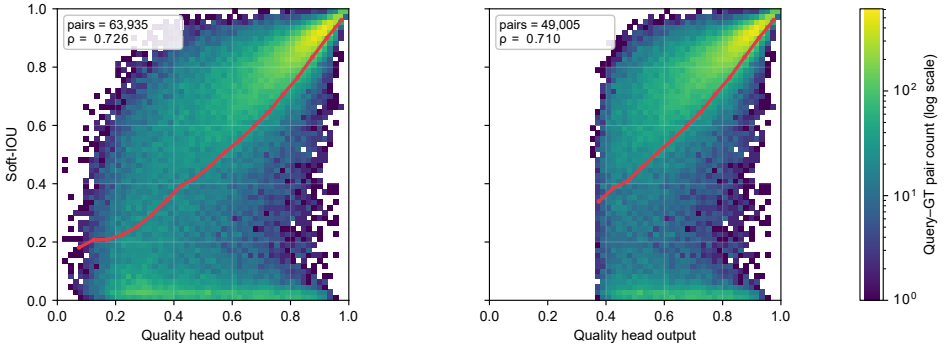


Figure 6: **Quality head output versus mask quality on COCO *val2017*.** Left: 63,935 same-class overlapping query–target pairs. Right: 49,005 pairs retained after applying a calibrated class–quality score threshold of 0.60. Color indicates pair counts on a logarithmic scale; the red curve shows mean soft-IoU in each quality-score bin. ρ denotes the Spearman correlation across all pairs in each panel.

A.5 Qualitative Segmentation Results

Figures 7, 8, and 9 show representative iFAN predictions for panoptic, instance, and semantic segmentation. Each row presents the input image, ground truth, and standalone prediction. These examples complement, rather than replace, the benchmark comparisons.



Figure 7: **Qualitative results for panoptic segmentation.** Each row shows an input image, ground truth, and the standalone prediction. The examples contain complex street layouts with large stuff regions and multiple thing instances, including vehicles, pedestrians, traffic signs, and roadside objects.



Figure 8: **Qualitative results for instance segmentation.** The three columns show the input image, the ground truth, and the standalone instance masks. The selected scenes emphasize dense arrangements, substantial occlusion, scale variation, and adjacent instances with similar appearance.

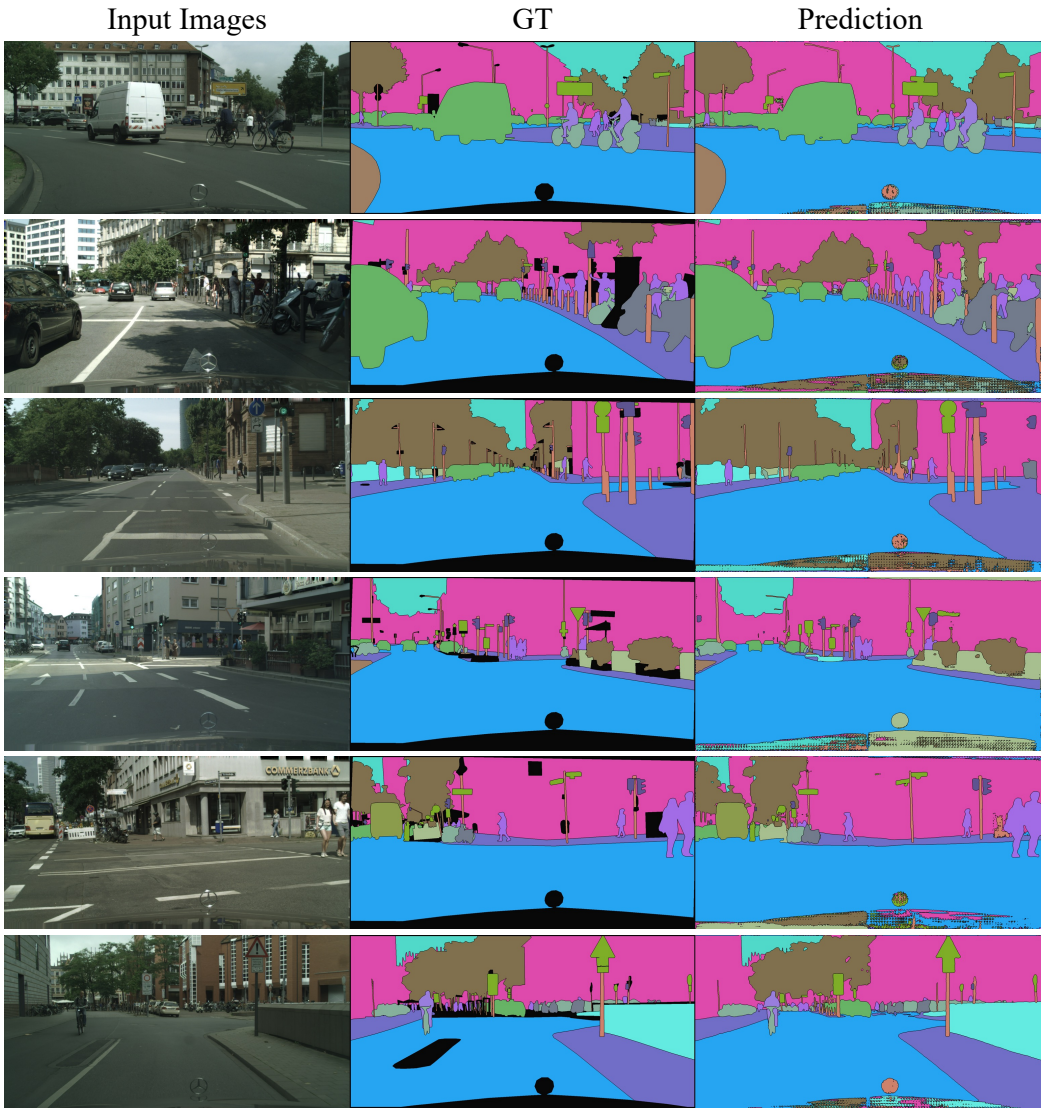


Figure 9: **Qualitative results for semantic segmentation.** Each row presents the input image, the ground truth, and the standalone prediction. The examples cover diverse indoor and outdoor scenes and illustrate coherent region assignment around objects with different shapes, scales, and surrounding context.