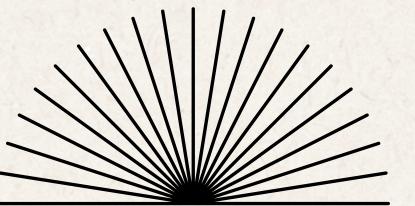


EDGETAM: ON-DEVICE TRACK ANYTHING MODEL

Zhou et al. - 2025



Pourquoi EdgeTAM ?

SAM 2 (Segment Anything Model 2)
→ 1 FPS sur iPhone 15 Pro Max

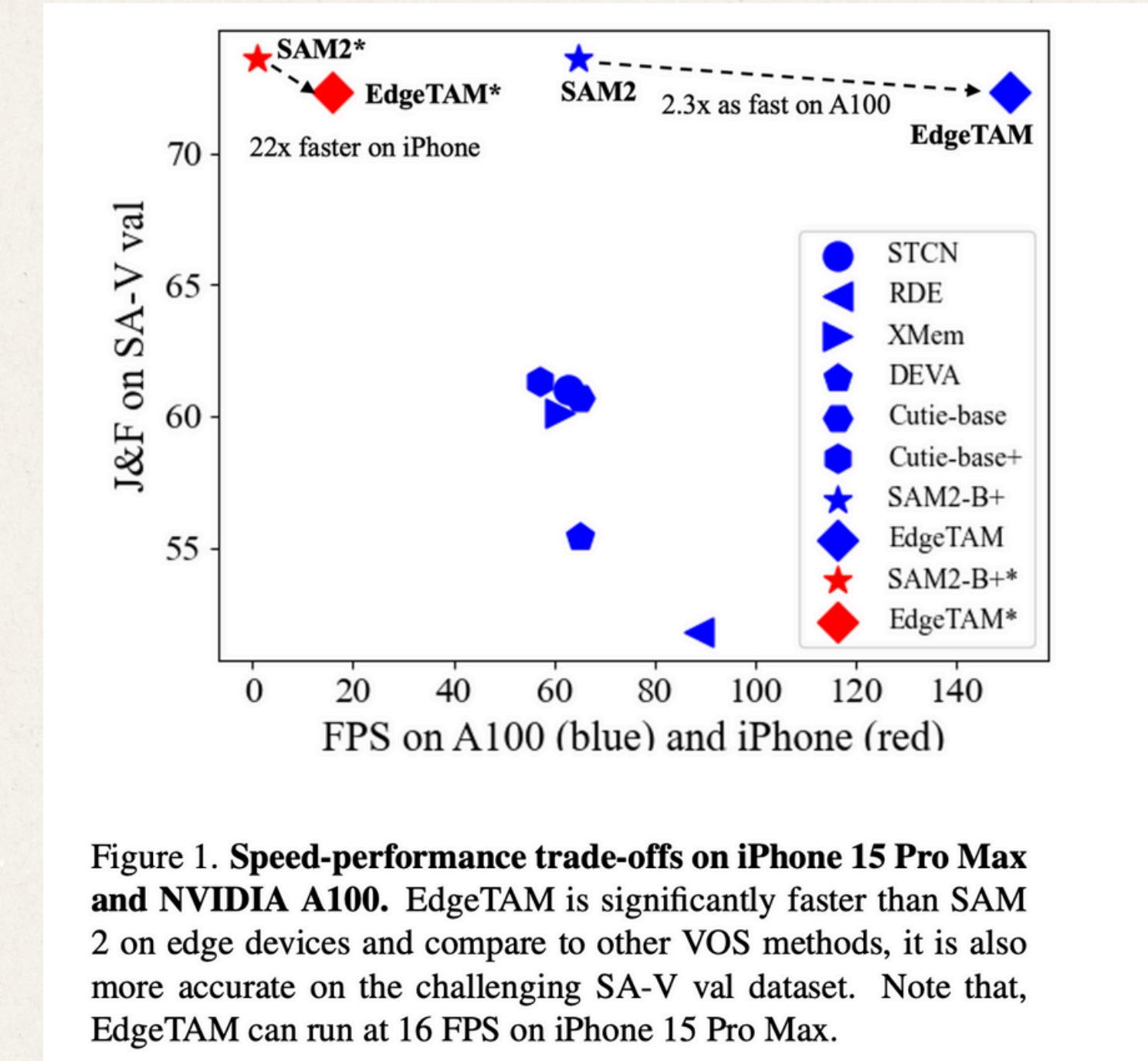


Figure 1. Speed-performance trade-offs on iPhone 15 Pro Max and NVIDIA A100. EdgeTAM is significantly faster than SAM 2 on edge devices and compare to other VOS methods, it is also more accurate on the challenging SA-V val dataset. Note that, EdgeTAM can run at 16 FPS on iPhone 15 Pro Max.

Objectif : Atteindre 16+ FPS tout en gardant la qualité de SAM 2

L'hypothèse initiale

SAM 2 = **Encodeur d'images** + Modules mémoire + Décodeur



Remplacer l'encodeur original (Hiera, très lourd) par des alternatives plus compactes
Essayé ViT-Tiny, RepViT, etc.

RepViT-M1 = encodeur d'images léger (petit nombre de paramètres) conçu pour les appareils mobiles



iPhone toujours aussi lent

La vraie cause - Les Blocs d'Attention Mémoire

SAM 2 = Encodeur d'images + **Modules mémoire** + Décodeur



7 frames gardées en mémoire

Nouvelle frame → comparaison avec les 7 autres pour savoir si l'objet à segmenter est le même

Multiplication matricielles complexes → problème computationnel

Frame mémoire de taille 64×64 pixels
→ $O(T \times H \times W^2)$
→ 4,096 pixels

La solution : 2D Spatial Perceiver

EdgeTAM **comprime la mémoire** avant de la comparer avec la frame actuelle

Approche initiale = réduire la résolution de la mémoire en utilisant du **pooling spatial**
→ Beaucoup plus rapide, mais la qualité s'effondre

Autre approche = **Perceiver** = compresser les données complexes de manière intelligente
→ transforme les pixels en résumés généraux mais perte d'information spatiale

La solution : 2D Spatial Perceiver

Solution = diviser les requêtes du Perceiver en deux groupes + passage de 4 blocs mémoire à 2

Groupe 1 - Requêtes Globales :

Quelques requêtes qui regardent toute la carte mémoire et font un résumé général

Groupe 2 - Requêtes Spatiales 2D :

Des requêtes qui sont assignées à des petites zones locales non-chevauchantes (des "patches"). Chaque requête ne regarde qu'un seul patch et le compressse

→ **structure spatiale préservée**

De $O(T \times H^2 \times W^2)$ à $O(T \times H \times W \times (Ng + Ni))$
(256 chacun dans EdgeTAM)
→ environ 8 fois plus rapide sans perte de qualité notable

La solution : 2D Spatial Perceiver

Approche	J&F (SA-V val)	J&F (SA-V test)	FPS iPhone	Evaluation
Baseline (encodeur léger seul)		63.5	62.1	2.5 ✖ Lent ✖ Moins bon
Average Pooling 4×4		61.8	59.8	15.7 ✓ Rapide ✖ Mauvaise qualité
Perceiver Standard		64.4	62.5	15.7 ✓ Rapide ✖ Structure perdue
2D Spatial Perceiver		64.4	62.5	15.7 ✓ Rapide ✓ Qualité OK

La distillation de connaissances

Même après cette optimisation, EdgeTAM est un peu moins précis que SAM 2. Les auteurs utilisent donc une technique appelée distillation de connaissances pour récupérer cette précision perdue.

Phase 1 - Segmentation d'images :

EdgeTAM apprend à segmenter les images (sans utiliser la mémoire). Pendant l'entraînement, on compare les représentations internes que produit EdgeTAM avec celles que produit SAM 2. Si elles sont différentes, on pénalise EdgeTAM jusqu'à ce qu'elles soient similaires.

Phase 2 - Segmentation vidéo :

Ensuite, EdgeTAM apprend à traiter les vidéos avec sa mémoire.

- On continue à comparer les features de la segmentation d'images
- On compare aussi les features après la fusion mémoire → force EdgeTAM à traiter la mémoire de manière similaire à SAM 2

+1.3 à +3.3 points de précision sans aucun coût de calcul supplémentaire

La distillation de connaissances

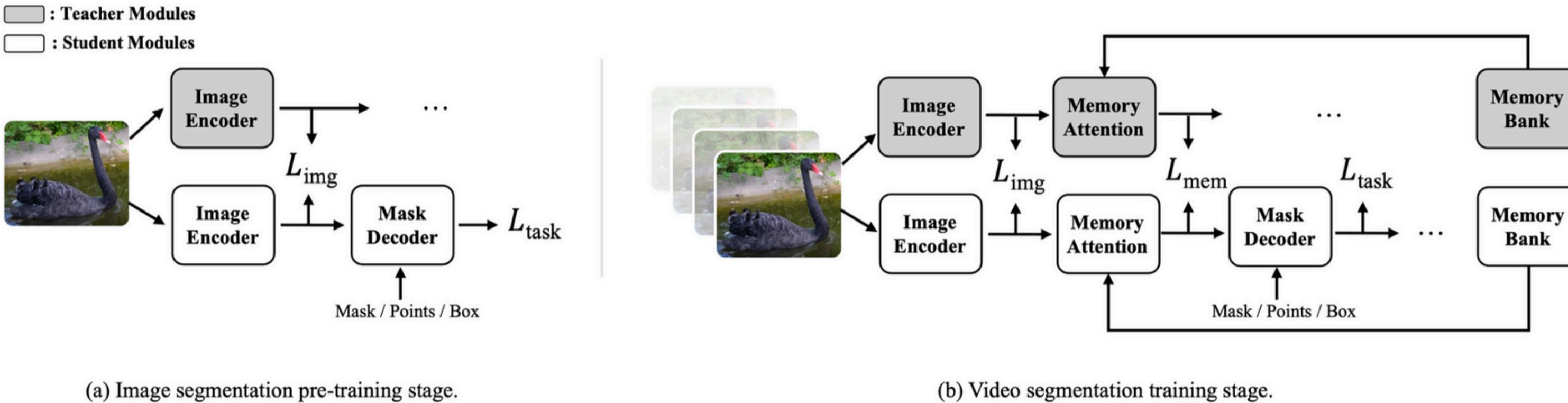


Figure 4. **The distillation pipeline in EdgeTAM.** In the image pre-training stage, we align the features from teacher’s and student’s image encoder. And in the video training stage, we additionally align the features output from memory attention between teacher and student. For both stages, task-specific losses are used.

Performance de Vitesse

Métrique	SAM 2	EdgeTAM	Improvement
iPhone 15 Pro Max	0.7 FPS	15.7 FPS	22× plus rapide
Temps par frame	~1.4 sec	~0.064 sec	22× plus rapide

Performance de Précision

SA-V val	SAM 2.1-B+	EdgeTAM	Différence
MOSE val	90.2 J&F	87.7 J&F	-2.5 (très proche)
YouTube VOS	88.6 G	86.2 G	-2.4 (acceptable)
SA-V val	76.8 J&F	72.3 J&F	-4.5 (acceptable)
MOSE val	76.6 J&F	70.0 J&F	-6.6 (acceptable)

- J&F = métrique de qualité standard pour la Segmentation Vidéo (Video Object Segmentation, VOS).
 - J (Region Similarity/IoU) : Mesure la précision de la forme (similaire au mIoU).
 - F (Boundary Accuracy) : Mesure la précision des bords/contours.
 - Le score J&F est souvent la moyenne harmonique de J et F.
- G : Représente le score de G-mean ou d'une autre métrique spécifique utilisée pour les données de YouTube VOS.

L'importance des prompts

Configuration	SAM 2	EdgeTAM	Différence
1 click	64.30%	54.40%	-9.90%
3 clicks	73.20%	72.70%	-0.50%
5 clicks	75.40%	75.50%	0.10%
Bounding box	72.90%	71.30%	-1.60%

Avec des prompts plus précis, EdgeTAM égale presque parfaitement SAM 2.

Modèle final

