TMI: Text-to-Image Meets Image-to-Image for Complementary Data Synthesis to Boost Long-Tailed Instance Segmentation

Hyeonseop Song*, Seokhun Choi*, and Hoseok Do†

AI Lab, CTO Division, LG Electronics, Republic of Korea
{hyeonseop.song, seokhun.choi, hoseok.do}@lge.com
Project page: https://seokhunchoi.github.io/TMI

Abstract. Large-vocabulary instance segmentation is constrained by long-tailed category distributions and fine-grained inter-class ambiguity. While data synthesis offers a promising alternative, current paradigms have complementary limitations: text-to-image (T2I) methods inherit noisy pseudo-labels and struggle on rare classes, whereas copy-paste methods compromise contextual realism. To address these issues, we propose a hybrid pipeline coupling T2I generation with context-aware image-to-image (I2I) editing. The T2I branch provides broad category and scene diversity, while a teacher-student scheme ensures label reliability by selectively retaining only prompt-specified categories. To strengthen supervision for rare classes, we introduce **VRAIN** (**V**erified **R**are-class **A**ugmentation via **IN**structed editing), a novel I2I editor. VRAIN inserts high-confidence instances at semantically appropriate locations within in-the-wild scenes, yielding semantically coherent and visually natural edits that reduce domain gaps and enable targeted augmentation. On the LVIS benchmark, our method surpasses existing baselines, improving overall AP by up to +4.0 points and rare-class AP by up to +9.5 points, while scaling effectively with backbone capacity.

Keywords: Generative Data Synthesis · Image Diffusion Models · Hybrid Data Pipeline · Long-Tailed Instance Segmentation

1 Introduction

Instance segmentation is a critical task, underpinning diverse applications from autonomous driving [4, 5] and robotics [16, 44, 50] to visual understanding [8, 26, 41] and image editing [21, 25, 38]. However, achieving robust performance relies on large, densely annotated datasets, which are costly to build. This bottleneck is particularly acute for large-vocabulary benchmarks such as LVIS [9] with their long-tailed category distributions. While methods like re-weighting [32, 33, 52], balanced sampling [3,11,52], and classifier calibration [28,36,37] help alleviate the imbalance, they cannot resolve the underlying data scarcity of rare categories [9].

*Equal contribution. †Corresponding author.

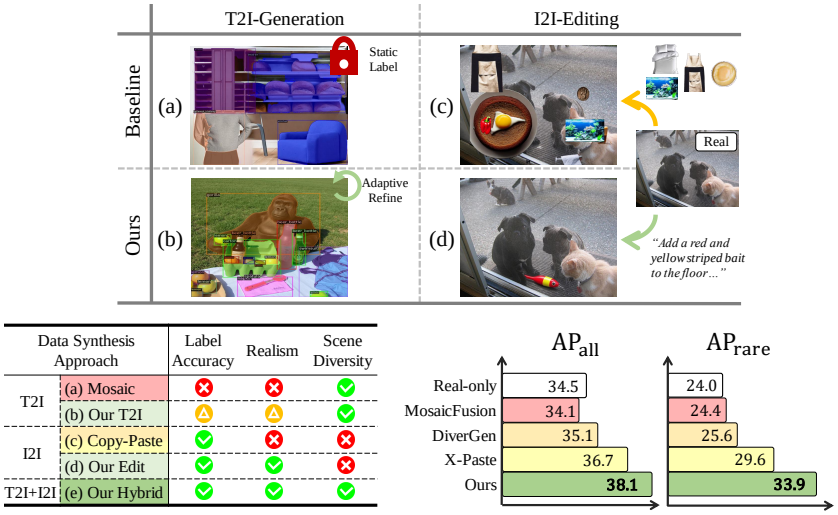


Fig. 1: Mutually Complementary T2I-I2I Data Synthesis. (a) Existing T2I methods (*e.g.*, MosaicFusion [45]) offer diverse scenes but lack label accuracy, while **(c)** Copy-Paste-based I2I methods (*e.g.*, X-Paste [51] and DiverGen [6]) offer accurate labels but reduce realism. **(e)** Our hybrid framework combines two proposed modules: **(b)** a T2I branch for scene diversity with reliable labels via teacher-student adaptive pseudo-labeling, and **(d)** context-aware I2I edits for realistic, accurate instance-level supervision. This unified approach achieves state-of-the-art results on LVIS benchmark.

This data scarcity has motivated recent work on synthetic data generation, leveraging advances in image generative modeling [12, 17, 31, 49] as a scalable alternative to manual annotation for improving long-tailed instance segmentation. Current generative data synthesis methods fall into three paradigms: text-to-image (T2I), label-to-image (L2I), and image-to-image (I2I). T2I methods [27, 42, 45], which first generate images from textual prompts and then assign semantic labels via pseudo-labeling, can introduce label noise, particularly for rare or fine-grained categories. L2I approaches [19, 47, 48] synthesize images conditioned on segmentation masks, enabling precise layout control via mask-to-image modules such as ControlNet [49]. However, scaling this approach to large category sets remains challenging, often leading to label-image mismatches and inaccurate annotations. In contrast, I2I approaches [6, 24, 51] augment real datasets by adding new objects via copy-paste or inpainting, offering accurate annotations for added instances. However, existing paste-based methods often suffer from domain gaps (*e.g.*, illumination mismatches), while inpainting struggles with plausible mask placement. Interestingly, when paste-based approaches explicitly target rare classes, performance on those classes often decreases rather than improves [6]. This counterintuitive result suggests that naive pasting often harms contextual realism, leading the segmentation model to overfit to synthetic artifacts rather than learning true representations.

To overcome these limitations, we propose a hybrid data synthesis framework that strategically integrates T2I generation with context-aware I2I editing

as illustrated in Fig. 1. Our T2I branch generates diverse images of broad categories and scenes, from which pseudo labels are obtained through an offline labeler. Since offline pseudo labels are prone to noise due to the domain gap between real and generated images, we additionally employ a teacher-student scheme [34] that adapts to the T2I domain and progressively refines pseudo-label quality. However, T2I generation alone is insufficient for rare classes, which remain difficult to label reliably. We address this limitation with our novel I2I branch, **VRAIN** (**V**erified **R**are-class **A**ugmentation via **IN**structed editing), to enhance rare-class representation. VRAIN operates as a two-stage pipeline: (i) it first performs context-aware rare-class placement via instruction-based editing to ensure natural scene integration; (ii) it then verifies the edit's semantic consistency and visual fidelity, generating a precise, high-fidelity annotation for the successfully integrated instance. This two-stage "place-and-verify" process directly mitigates the contextual inconsistency of copy-paste methods and the mask-placement ambiguity inherent in inpainting-based approaches. By yielding semantically coherent and visually natural edits, this process alleviates domain gaps and compositional artifacts, while also enabling targeted augmentation for rare categories.

Experimentally, our approach achieves state-of-the-art results on the LVIS benchmark, significantly outperforming existing paste-based augmentation and T2I baselines. Notably, it demonstrates remarkable effectiveness for both overall and underrepresented categories. Furthermore, our method scales effectively with larger backbones, demonstrating its practicality as a scalable solution for data generation in large-vocabulary instance segmentation.

In summary, our key contributions are as follows:

- We propose a unified T2I-I2I data synthesis framework that leverages T2I for scene diversity and I2I editing for enhanced realism and rare-class representation.
- We design VRAIN, a novel "place-and-verify" I2I pipeline, which leverages instruction-based editing for semantically coherent placement and introduces a verification stage to generate accurate annotations.
- We achieve state-of-the-art performance on the LVIS benchmark, especially on rare categories and demonstrate effective scaling with larger backbone models.

2 Related Work

2.1 Image Generative Models

High-quality T2I generation is dominated by diffusion models, which have rapidly evolved from DDPM [12] and latent diffusion [31] to more efficient flow-matching formulations [20,22]. Modern architectures such as Flux [17] and Qwen-Image [40] further boost fidelity and scalability by adopting Diffusion Transformer (DiT) [29] backbones with flow-matching objectives. These T2I advances also empower instruction-based I2I editing—pioneered by InstructPix2Pix [2] and extended by

unified frameworks like Flux-Kontext [18] and Qwen-Image-Edit [40]—enabling semantically consistent, language-guided generation and editing. Our dataset synthesis framework is built on these generative and editing developments.

2.2 Data Generation for Segmentation Task

Recent advances in generative modeling have led to various data generation strategies for segmentation, broadly categorized as T2I, L2I, and I2I. **T2I methods** generate images from textual prompts, often using diffusion attention maps [27, 43, 45] or fine-tuning perception decoders [42] for pseudo-labeling in few-shot settings. For example, MosaicFusion [45] divides the image canvas into mosaic [1]-like regions and simultaneously generates multiple objects. However, most T2I approaches are effective on small-scale data or for a limited number of categories [27, 43], struggling to scale to large-scale, in-the-wild scenarios.

L2I methods generate images conditioned on segmentation masks, enabling precise control over object layout. Some works [19, 47] utilize mask-to-image modules like ControlNet [46, 49], while others [48] extend this framework by designing text-to-mask modules for mask diversity. However, extending L2I to large-vocabulary datasets such as LVIS [9] remains challenging. Mask-only conditioning often fails to faithfully render the semantics of numerous fine-grained categories, leading to label-image mismatches.

I2I methods expand datasets by editing existing images. Copy-paste techniques, such as X-Paste [51] and DiverGen [6], leverage generative models to create diverse instance pools for composition onto real images. This successfully covers large categories where L2I methods struggle; however, the pasted instances often appear visually distinct from their surroundings (*e.g.*, illumination or blending mismatches), limiting training gains. Inpainting-based approaches [14, 24] synthesize objects into masked regions. While effective in constrained domains (*e.g.*, road scenes [14, 24]), they have not scaled to in-the-wild, large-category datasets like LVIS, due to the challenges of realistic mask placement and high semantic diversity of scenes.

Our hybrid framework addresses these complementary failures. We harness T2I for diversity but mitigate its label noise by employing prompt-consistent filtering and teacher-student refinement. We then use our I2I branch to address both T2I's rare-class weakness and prior I2I's realism gap, providing high-fidelity, targeted augmentation that scales effectively on LVIS.

3 Methods

We generate synthetic training data through two complementary paradigms: (i) T2I generation for diverse scene synthesis, and (ii) I2I editing for context-aware augmentation of real images. The T2I branch (Sec. 3.1) provides broad semantic coverage across all categories in the long-tailed dataset. However, assigning accurate annotations remains a challenge; conventional offline models, trained exclusively on real images, produce noisy pseudo-labels due to the domain gap

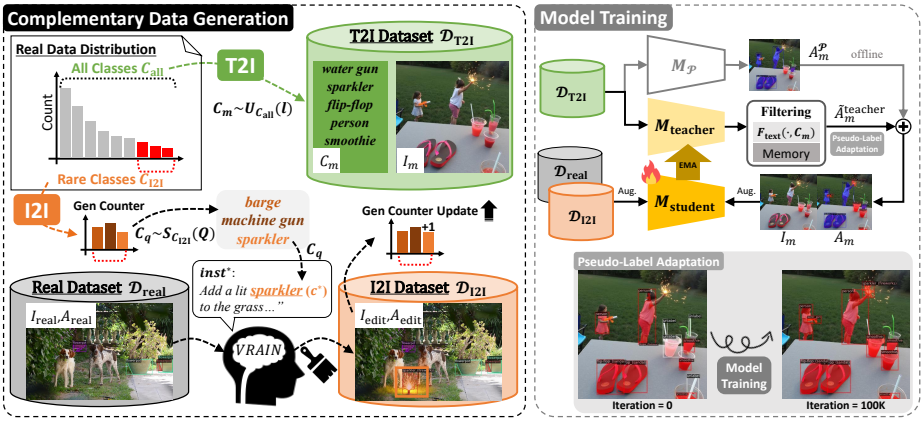


Fig. 2: Overview of Hybrid Data Generation and Training Framework. We combine two complementary paradigms: (i) T2I generation for diverse scenes over all categories C_{all} and (ii) I2I editing (VRAIN) for rare categories C_{I2I} with accurate instance-level supervision. (iii) A student M_{student} jointly learns from both $\mathcal{D}_{\text{T2I}}$ and $\mathcal{D}_{\text{I2I}}$. For $\mathcal{D}_{\text{T2I}}$ images I_m, an EMA teacher M_{teacher} generates refined pseudo-labels $\tilde{A}_m^{\text{teacher}}$, merged with offline labels $A_m^{\mathcal{P}}$ to form the final supervision A_m. As training proceeds, the teacher adapts to the T2I domain, progressively improving initially noisy pseudo-labels (pseudo-label adaptation). For example, rare classes (*e.g.*, *sparkler* and *water gun*)—missed by the offline labeler and initially unlabeled—are gradually labeled correctly as the teacher adapts with the support from rare-class-targeted I2I data.

between real and generated images. We therefore treat this T2I set as a source of weakly-labeled data, $\mathcal{D}_{\text{T2I}}$. In contrast, our I2I branch (VRAIN, Sec. 3.2) enhances rare-class representation. Because these edits are driven by specific instructions, the identity and location of the newly inserted instance are explicitly known, yielding accurate, instance-level labeled data ($\mathcal{D}_{\text{I2I}}$) that provide reliable supervision for rare categories where T2I struggles. Our final synthesized dataset is the union of these two complementary sets, $\mathcal{D}_{\text{gen}} = \{\mathcal{D}_{\text{T2I}}, \mathcal{D}_{\text{I2I}}\}$. Finally, we introduce a teacher-student scheme (Sec. 3.3) to train the student model on this hybrid dataset. In this semi-supervised loop, an EMA-updated teacher generates dynamic, online pseudo-labels for the $\mathcal{D}_{\text{T2I}}$. The student is jointly trained on these progressively refined T2I pseudo-labels to expand category and scene diversity, while also incorporating the labeled I2I data for accurate, instance-level supervision. An overview of the entire process is depicted in Fig. 2.

3.1 Text-to-Image Generation

Our T2I method consists of two stages: (i) generating a set of text prompts and corresponding images, and (ii) offline pseudo-labeling with text-consistent filtering. First, to generate text prompt set $\mathcal{T}$ of size N_{text}, we randomly sample a subset C_m of size l from the total category set C_{all}. We then create a descriptive prompt t_m composed of these selected classes using GPT-4o [13]:

$$\mathcal{T} = \{t_m \mid t_m = \text{GPT-4o}(C_m)\}. \tag{1}$$

Method	AP^{box}	AP^{mask}	AP^{box}_r	AP^{mask}_r
(a) Real-only	34.5	30.8	24.0	21.6
(b) High-threshold Filtering	35.8	32.0	31.2	28.0
(c) Text-consistent Filtering	**36.7**	**33.0**	**32.1**	**28.9**

Table 1: Comparison of T2I Pseudo-label Filtering Strategies. We compare models trained with T2I data, where the raw predictions $\hat{A}^{\mathcal{P}}_m$ are filtered using either (b) a high-threshold strategy or (c) our text-consistent strategy. Our method (c) retains semantically valid, low-confidence instances that high-threshold filtering discards, resulting in improved overall and rare-class AP.

Subsequently, we generate an image I_m from the created text prompt using the image generation model Φ_{T2I} [17] and construct N_{T2I} unlabeled T2I dataset as image-category pairs: $\mathcal{D}_{\text{T2I}} = \{(I_m, C_m) \mid I_m = \Phi_{\text{T2I}}(t_m)\}$

Second, we pre-compute a set of offline labels for this dataset. We process each image I_m with a pre-trained public instance segmentation model [55] $M_{\mathcal{P}}$ to obtain an initial set of raw predictions, $\hat{A}^{\mathcal{P}}_m = M_{\mathcal{P}}(I_m)$. To leverage the prior knowledge that I_m was generated from the prompt t_m containing classes C_m, we apply a text-consistent filtering process F_{text}. For the k-th predicted instance annotation $a^{\mathcal{P}}_k$, this filter retains only the instances whose predicted classes belong to C_m:

$$A^{\mathcal{P}}_m = F_{\text{text}}(\hat{A}^{\mathcal{P}}_m, C_m) = \left\{ a^{\mathcal{P}}_k \mid a^{\mathcal{P}}_k \in \hat{A}^{\mathcal{P}}_m,\ \text{class}(a^{\mathcal{P}}_k) \in C_m \right\}. \tag{2}$$

The effectiveness of this text-consistent filtering is validated in Table 1. Our filtering method yields superior results compared to high-threshold instance mining. This demonstrates that our filtering strategy, which leverages prompt knowledge, is more effective than relying solely on high-confidence scores. It successfully retains valid instances, even those with low-confidence, that are consistent with the text prompt. These filtered offline labels, $A^{\mathcal{P}}_m$, are then used as one of the supervision sources in our training pipeline, detailed in Sec. 3.3.

3.2 Image-to-Image Editing

While our T2I branch provides broad diversity, it struggles to produce accurate instance-level annotations, especially for rare categories prone to label noise. A high-quality seed set of verified rare-class annotations is therefore essential, as it bootstraps the online pseudo-labeling methods (*e.g.*, EMA teacher [34]). I2I approaches, which can add new objects with accurate annotations, offer a promising solution. However, prior I2I methods suffer from critical flaws: copy-paste techniques produce contextually inconsistent artifacts, while inpainting-based methods struggle to identify suitable mask regions in complex scenes. Motivated by these limitations, we propose VRAIN (**V**erified **R**are-class **A**ugmentation via **IN**structed editing), illustrated in Fig. 3. VRAIN is a principled two-stage pipeline designed to overcome these realism and placement challenges, performing (1) context-aware rare-class placement via instruction-based editing, and (2) verification of the edit's fidelity to generate trustworthy annotation.

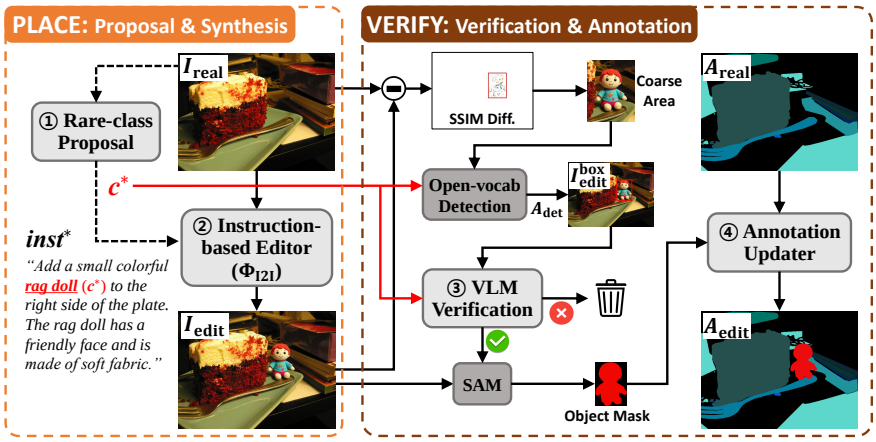


Fig. 3: VRAIN "Place-and-Verify" Pipeline. Our two-stage framework ensures high-fidelity I2I editing. **(i) Place:** A VLM proposes a semantically coherent instruction ($inst^*$) for inserting a rare class (c^*) that fits naturally within I_{real}. An instruction-based editor Φ_{I2I} then synthesizes I_{edit}. **(ii) Verify:** I_{edit} is then validated. The new instance is localized via SSIM difference and open-vocab detection, semantically confirmed by a VLM filter, and masked using SAM. Finally, an annotation updater resolves occlusions with A_{real}, integrating the new instance to produce the trustworthy final annotation A_{edit}.

Rare-class Proposal and Synthesis. A central challenge in rare-class generation is selecting a semantically fitting category and determining its placement, ensuring the inserted instances are contextually coherent with the scene. This process is key to reducing the unrealistic compositions of previous copy-paste methods. To this end, we leverage a Vision Language Model (VLM) [54] to ensure semantic consistency, by selecting the most suitable category from a candidate set for a given image, and then generating its placement instruction. This candidate set (C_q) is formed by sampling Q categories from all target categories C_{I2I} for a given real image I_{real} from the training dataset $\mathcal{D}_{\text{real}} = \{(I_{\text{real}}, A_{\text{real}})\}$. The sampling is performed using a softmax-based sampler ($C_q \sim S_{C_{\text{I2I}}}(Q)$) that is weighted by the generation count of each class (initialized to zero), penalizing frequently generated classes. The VLM is then prompted with I_{real} and the candidate set C_q to recommend a suitable category c^* along with the corresponding textual instruction $inst^*$:

$$(c^*,\ inst^*) = \text{VLM}(I_{\text{real}},\ C_q), \tag{3}$$

where $inst^*$ is the natural language instruction fed to the editor, while c^* is the explicit category label passed to the annotation stage (see the supplementary material for details). Finally, $inst^*$ is used by an instruction-based editing model Φ_{I2I} [18] to produce the edited image: $I_{\text{edit}} = \Phi_{\text{I2I}}(I_{\text{real}}, inst^*)$.

Verification and Annotation. Although the instruction-based editor Φ_{I2I} produces visually plausible results, these edits may suffer from contextual mismatches or rendering artifacts. To ensure annotation quality, VRAIN performs a

verification and annotation stage that evaluates both semantic consistency and visual fidelity of each edited image. First, to locate the edited region, we compute a structural similarity (SSIM [39]) difference map between I_{real} and I_{edit}, thresholding at τ_{edit} to identify modified pixels. Second, to detect the new object, we apply an open-vocabulary detector [7] within this coarse area, searching for the intended category c^* and producing a set of candidate detections A_{det}.

However, A_{det} may contain false-positives or artifacts. To semantically verify these candidates, we introduce a VLM-based filter, $F_{\text{VLM}}(A_{\text{det}}, c^*)$. For each detection $a_k \in A_{\text{det}}$, we prepare an image $I^{\text{box}}_{\text{edit}}(a_k)$ by overlaying its bounding box in red on the edited image. The VLM is then prompted with this visualization along with a template-based question t^{c^*} (*e.g.*, "Is the red bounding box in the image a c^*?"). The VLM outputs a binary 'Yes' or 'No' response. A detection a_k is confirmed only if the VLM responds 'Yes':

$$F_{\text{VLM}}(A_{\text{det}}, c^*) = \Big\{ a_k \mid a_k \in A_{\text{det}}, \text{VLM}(I^{\text{box}}_{\text{edit}}, t^{c^*}; a_k) = \text{Yes} \Big\}. \tag{4}$$

For all verified objects in F_{VLM}, we obtain instance masks utilizing Segment Anything Model (SAM) [30]. These new masks are then added to the original annotation A_{real}. However, this integration can create occlusions, where new SAM masks overlap with an existing ground-truth mask from A_{real}. Thus, we update in A_{real} by subtracting the new SAM masks from any overlapping original masks to produce the final annotation A_{edit}. If this process yields at least one verified instance mask, the generation count for category c^* is incremented by one. This entire procedure is repeated until the total count for C_{I2I} reaches the target N_{I2I}, resulting in the final dataset $\mathcal{D}_{\text{I2I}} = \{(I_{\text{edit}}, A_{\text{edit}})\}$ of size N_{I2I}.

3.3 Model Training

The straightforward way to train an instance segmentation model [53] using the two complementary synthesized datasets, $\mathcal{D}_{\text{gen}} = \{\mathcal{D}_{\text{T2I}}, \mathcal{D}_{\text{I2I}}\}$, is to combine them with the real dataset $\mathcal{D}_{\text{real}}$ for joint training. However, unlike $\mathcal{D}_{\text{real}}$ or $\mathcal{D}_{\text{I2I}}$, all images in $\mathcal{D}_{\text{T2I}}$ are fully generated, which can lead to domain gaps and unreliable pseudo-labels using a static labeler trained only on real data. To address this issue, we adopt an EMA-based teacher-student scheme [34], where the student model M_{student} learns from supervision signals generated by the teacher model M_{teacher}. Its parameters are updated as $\Theta_{\text{teacher}} \leftarrow \gamma\Theta_{\text{teacher}} + (1-\gamma)\Theta_{\text{student}}$, where Θ_{teacher} and Θ_{student} denote the parameters of the teacher and student models, respectively, and γ is the EMA decay rate. The teacher gradually adapts to the T2I domain during training, thereby producing more accurate pseudo-labels for T2I data. Moreover, because the teacher is also trained on the high-quality, rare-class-focused $\mathcal{D}_{\text{I2I}}$, it becomes more adept at assigning more precise labels even for rare categories in $\mathcal{D}_{\text{T2I}}$.

Specifically, at each training step, a batch is sampled from $\{\mathcal{D}_{\text{real}}, \mathcal{D}_{\text{I2I}}, \mathcal{D}_{\text{T2I}}\}$. For images in $\mathcal{D}_{\text{real}}$ and $\mathcal{D}_{\text{I2I}}$, the student is trained directly using their ground-truth annotations (A_{real} and A_{edit}). On the other hand, for images in $\mathcal{D}_{\text{T2I}}$, the

student is trained on hybrid pseudo-labels created by processing and merging the offline ($A_m^{\mathcal{P}}$) and new online (A_m^{teacher}) labels generated by the teacher.

This training-time process for online labels is as follows. First, the online EMA-based teacher M_{teacher} produces raw predictions $\hat{A}_m^{\text{teacher}}$, which are then split into two groups according to their predicted class $\hat{c}$ and the sampled category set C_m. Prompt-consistent instances ($\hat{c} \in C_m$) are filtered using a class-wise adaptive threshold. This threshold is dynamic; we maintain a per-category running memory of recent confidences, and the threshold for a given class is set to τ_{label} times the mean confidence of that memory, allowing the filter to adapt as the teacher improves. Prompt-inconsistent instances ($\hat{c} \notin C_m$) are assigned an 'unlabeled' class if they are detected with high confidence above $\tau_{unlabel}$ times the mean confidence. These 'unlabeled' instances are excluded only from the classification loss but are still included in box and mask losses [53], providing implicit localization supervision from objects that were confidently detected but not explicitly specified in the prompt. These two sets of processed labels (the adaptively-thresholded and the unlabeled instances) together form the complete online pseudo-label set, $\tilde{A}_m^{\text{teacher}}$. Finally, we merge this dynamic online set $\tilde{A}_m^{\text{teacher}}$ with the static offline pseudo-labels $A_m^{\mathcal{P}}$ via IoU-based matching to produce the student's final supervision, A_m. This strategy combines their complementary strengths: the offline set $A_m^{\mathcal{P}}$ provides a stable, precision-oriented baseline of high-confidence instances with conservative coverage of common categories, while the online set $\tilde{A}_m^{\text{teacher}}$ introduces progressively refined, domain-adapted labels, especially for rare classes thanks to the auxiliary guidance from I2I data. The student M_{student} thus learns from both a robust baseline and dynamic, in-domain corrections.

4 Experiments

4.1 Experimental Settings

Dataset. We conducted experiments on LVIS [9], a large-scale instance segmentation dataset featuring a long-tailed distribution of 1,203 categories. It contains 100K training and 20K validation images with 2M annotations. Based on the number of training images per category, categories are divided into three frequency groups: rare (1–10 images; 337 categories), common (11–100 images; 461 categories), and frequent (>100 images; 405 categories).

Baselines. We compare our method against three baselines: MosaicFusion [45] (a T2I method), X-Paste [51], and DiverGen [6] (I2I copy-paste methods). We use CenterNet2 [53], a widely used model for large-vocabulary instance segmentation, for all our experiments, with two backbone configurations: ResNet-50 [10] (640 resolution, batch size 32) and Swin-L [23] (896 resolution, batch size 16). Following the convention of prior work [6,51], all models are trained for 180K iterations, initializing from their respective "Real-only" pretrained checkpoints. To ensure a fair comparison under identical experimental settings, we reproduced

Backbone	Method	T2I Dataset	I2I Dataset	AP^{box}	AP^{mask}	AP_r^{box}	AP_r^{mask}
ResNet-50	Real-only [53]	-	-	34.5	30.8	24.0	21.6
ResNet-50	MosaicFusion [45]	✓		34.1	30.4	24.4	22.5
ResNet-50	DiverGen [6]		✓	35.1	31.2	25.6	23.8
ResNet-50	X-Paste [51]		✓	36.7	33.0	29.6	27.8
ResNet-50	**Ours**	✓	✓	**38.1**	**34.0**	**33.9**	**31.7**
Swin-L	Real-only [53]	-	-	47.5	42.3	41.4	36.8
Swin-L	MosaicFusion [45]	✓		47.7	42.8	41.3	37.5
Swin-L	DiverGen [6]		✓	49.6	44.2	44.5	39.8
Swin-L	X-Paste [51]		✓	50.1	44.4	48.2	43.3
Swin-L	**Ours**	✓	✓	**50.7**	**45.2**	**49.1**	**44.0**

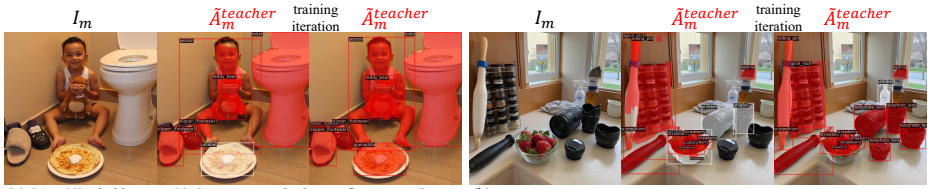
Table 2: Comparison with Baselines on LVIS validation Dataset. Our hybrid T2I-I2I approach consistently outperforms all baselines: the Real-only model, the T2I-only method (MosaicFusion), and copy-paste I2I methods (DiverGen, X-Paste). This superior performance holds across both ResNet-50 and Swin-L backbones, delivering particularly strong gains on rare-class metrics and demonstrating the effectiveness of combining both data synthesis paradigms.

DiverGen using their official code, generating 1K synthetic instance pool per category (1.2M in total) with two generative models, and also reproduced MosaicFusion following its original implementation. For the evaluation metrics, we evaluate performance using LVIS box average precision AP^{box} and mask average precision AP^{mask}. We also report rare-category results AP_r^{box} and AP_r^{mask}.

Implementation Details. We generated 200K T2I images ($N_{\text{T2I}} = 200\text{K}$) using Flux.1-dev [17] from a set of 40K prompts ($N_{\text{text}} = 40\text{K}$), where each prompt was created by randomly sampling $l \in [5, 10]$ categories from the total LVIS category set. We also generated 80K I2I images ($N_{\text{I2I}} = 80\text{K}$) from the LVIS training set using Flux.1 Kontext-dev [18], targeting rare LVIS categories with $Q = 5$ and an edit-region threshold $\tau_{\text{edit}} = 5$. We set the EMA decay rate to $\gamma = 0.999$, the same as the baselines, but in our framework it is used to adaptively refine pseudo-labels rather than just for evaluation. We set $\tau_{label} = 0.7$ and $\tau_{unlabel} = 1.2$ for dynamic threshold in online pseudo-labeling with running memory size 100 per class. We used InternVL3-14B [54] as the VLM for VRAIN's recommendation and verification steps. Further details on VLM implementation including prompt creation are provided in the supplementary material.

4.2 Comparison

We evaluated our approach against various data synthesis methods, as shown in Table 2. MosaicFusion, a rare-class-targeted T2I method, offers only marginal gains over the real-only model, indicating that despite its scene diversity, the limited label accuracy of T2I generation constrains performance on long-tailed datasets. Copy-paste I2I methods like DiverGen and X-Paste provide more accurate instance-level labels and achieve larger improvements, yet still lag behind our hybrid T2I-I2I framework. Our method uniquely combines the strengths of



(a) *"A **toddler** holding a **teddy bear** sits on a bathroom floor near **slippers**, a **toilet** in the background, and a **quesadilla** on a plate beside them, warm indoor lighting, candid photography style."*

(b) *"A kitchen counter with a **spice rack** beside a **rolling pin** and **strawberries**, a **screwdriver** and **telephoto lens** near a **beanie**, **traffic light** visible through the window, warm natural lighting, professional photography."*

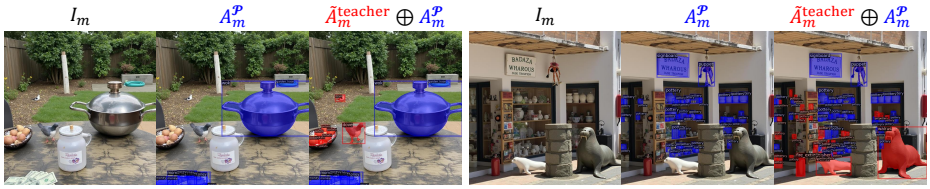
Fig. 4: Examples of Pseudo-label Adaptation in T2I images. In each T2I prompt, boldfaced and underlined words indicate the target categories for generation. As the teacher model updates, it adapts to the T2I domain and enhances pseudo-label ($\tilde{A}_m^{\text{teacher}}$) quality—(a) improving the *toddler* mask and labeling previously missed objects (*e.g.*, *quesadilla*), and (b) refining masks (*e.g.*, *spice rack*, *rolling pin*) while newly labeling a *telephoto lens*.

Method	AP^{box}	AP^{mask}	AP_r^{box}	AP_r^{mask}
(a) Real-only	47.5	42.3	41.4	36.8
(b) MosaicFusion [45]	47.7	42.8	41.3	37.5
(c) DiverGen (rare-target) [6]	47.3	42.2	34.5	31.2
(d) $\mathcal{D}_{\text{I2I}}$ only (VRAIN)	**48.1**	**43.3**	**42.9**	**39.5**

Table 3: Comparison with Rare-class-targeted Data Synthesis Approaches. Among methods generating data for rare categories, (c) DiverGen shows decreased rare-class performance due to reduced realism, whereas (d) our I2I approach effectively and practically improves rare-class performance with high visual fidelity, accurate labels.

both paradigms. It leverages T2I for broad diversity like MosaicFusion, but uses our teacher-student adaptive pseudo-labeling to ensure high quality label. By further incorporating VRAIN (I2I) to inject rare-class data with precise instance-level labels, it enhances performance on underrepresented categories and bootstraps online pseudo-labeling, boosting overall accuracy. Fig. 4 demonstrates the EMA teacher's progressive improvement of pseudo-labels on T2I images during training—including the detection and mask refinement of previously missed rare-class objects such as *quesadilla* and *telephoto lens*—effectively bootstrapping the online pseudo-labeling process. This synergy achieves the largest gains across all categories and especially for rare classes, improving AP^{box} from 34.5 → 38.1 (ResNet-50) and 47.5 → 50.7 (Swin-L) overall, and AP_r^{box} from 24.0 → 33.9 (ResNet-50) and 41.4 → 49.1 (Swin-L) for rare categories.

To further analyze rare-class-focused strategies, we compare approaches that synthesize data exclusively for rare categories, as shown in Table 3. Although designed for rare-class enhancement, MosaicFusion again provides a marginal improvement over the real-only model, indicating that the limited label reliability of T2I generation constrains its effectiveness. Notably, the rare-class-focused variant of DiverGen catastrophically underperforms the real-only baseline on rare categories (AP_r^{box} 41.4 → 34.5). This contradictory result highlights that naive copy-pasting harms performance due to its poor contextual coherence and realism, encouraging overfitting to pasted instances rather than learning true representations. In contrast, our rare-class-focused I2I approach (VRAIN) successfully boosts rare-class performance, benefiting from high visual fidelity and



(a) *"A backyard scene with a **garden hose** coiled near a **wok** on a patio table, a **dove** perched nearby, **potatoes** in a basket, and **money** scattered on the ground, soft natural daylight photography."*

(b) *"A coastal souvenir shop with a **walrus** statue near a **signboard**, **pottery** on shelves, a **fire extinguisher** on the wall, a **puppet** hanging from the **cornice**, warm afternoon light, casual photography."*

Fig. 5: Examples of Final Pseudo-labels for T2I images. In each T2I prompt, boldfaced and underlined words indicate the target categories for generation. Teacher-generated pseudo-labels (red) are merged ($\oplus$) with offline pseudo-labels (blue) to form the final supervision. As the teacher adapts to the T2I domain, it successfully labels rare classes—(a) *dove* and (b) *walrus*.

Method	$\mathbf{AP}^{\text{box}}$	$\mathbf{AP}^{\text{mask}}$	$\mathbf{AP}_{\text{r}}^{\text{box}}$	$\mathbf{AP}_{\text{r}}^{\text{mask}}$
(a) Real-only	47.5	42.3	41.4	36.8
(b) M_{teacher} only	49.4	43.9	**47.9**	43.0
(c) $M_{\mathcal{P}}$ only	49.9	44.9	45.3	41.7
(d) M_{teacher} and $M_{\mathcal{P}}$	**50.3**	**45.1**	47.5	**43.8**

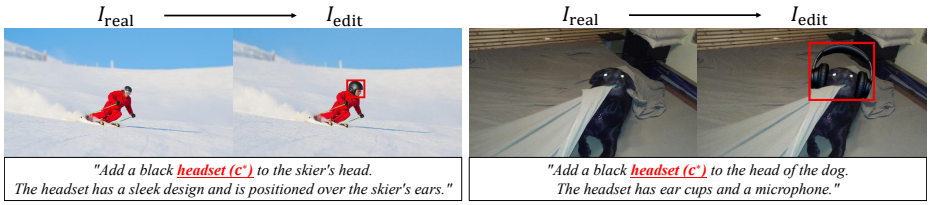
Table 4: Ablation Study on T2I Pseudo-label Supervision. We evaluate the impact of different pseudo-label sources: offline pseudo labeler ($M_{\mathcal{P}}$), EMA teacher (M_{teacher}), and their combination. The results show that $M_{\mathcal{P}}$ provides stable supervision for common categories, M_{teacher} adaptively improves rare-class coverage, and combining both leverages complementary strengths.

precise instance-level labeling. These results demonstrate the effectiveness and practical applicability of our method for improving underrepresented categories.

4.3 Ablation Study

Effect of EMA-Teacher Labeling. To evaluate the effectiveness of our EMA-teacher framework, we conducted an ablation study on T2I data labeling, as shown in Table 4. Specifically, we trained separate instance segmentation models using either the offline pseudo labels ($A_m^{\mathcal{P}}$) or the EMA teacher-generated labels ($\tilde{A}_m^{\text{teacher}}$) as supervision. The model trained with $A_m^{\mathcal{P}}$ achieves relatively higher AP on common categories, reflecting the stable supervision provided by the offline pseudo labels. In contrast, the model trained with $\tilde{A}_m^{\text{teacher}}$ leverages pseudo labels that are progressively updated by the EMA teacher throughout training, adaptively correcting errors and improving AP on underrepresented rare classes. Combining both $M_{\mathcal{P}}$ and M_{teacher} leverages their complementary strengths, maintaining high accuracy for common categories while improving rare-class coverage, as visually illustrated in Fig. 5.

Reliability of VRAIN's Verification and Annotation. To quantitatively assess the reliability of our VLM-based verification module in VRAIN, we construct a controlled perturbation benchmark from the LVIS validation set (244k instances). We generate negative samples by randomly perturbing instance class



(a) Filtered-out Case (b) Filtered-in Case

Fig. 6: Qualitative Examples of the VRAIN Verification Stage. (a) A 'helmet' is added instead of the requested 'headset' (c^*), creating a semantic mismatch, and is thus filtered out; (b) The target 'headset' is correctly added, passing the verification.

Method	Precision	Recall	FP Rate	FN Rate
(a) CLIP (Crop)	0.886	0.623	0.187	0.377
(b) VLM (Crop)	0.975	0.521	**0.031**	0.478
(c) VLM (Ours)	**0.984**	**0.846**	0.033	**0.154**

Table 5: VLM Verification Reliability on LVIS Validation. Compared to prior CLIP-based verification, our spatially grounded VLM verifier achieves a superior precision–recall trade-off with lower FP and FN rates.

Method	AP^{box}	AP^{mask}	AP_r^{box}	AP_r^{mask}
(a) Real-only	47.5	42.3	41.4	36.8
(b) $\mathcal{D}_{\text{I2I}}$ only (w/o verification)	47.4	42.5	39.5	36.0
(c) $\mathcal{D}_{\text{I2I}}$ only (VRAIN)	**48.1**	**43.3**	**42.9**	**39.5**

Table 6: Ablation Study on VLM Verification in VRAIN. We compare I2I data with and without VLM-based filtering. Removing the verification step (b) reduces both overall and rare-class AP, indicating that the VLM filter is crucial for retaining semantically correct and relevant instances.

labels with 30% probability while preserving the original images and bounding boxes. We measure false positive (FP) rate when the verifier incorrectly accepts perturbed labels and false negative (FN) rate when it rejects correct ones. We compare three verification strategies: (a) CLIP similarity on crops ("a photo of [class]"), (b) crop-based VLM verification ("Is the image a [class]?"), and (c) our spatially grounded red-box VLM ("Is the red bounding box a [class]?"). As shown in Table 5, the VLM verifier significantly reduces the FP rate compared to CLIP (0.187 → 0.033). Moreover, explicitly providing spatial grounding significantly reduces the FN rate (0.478 → 0.154) compared to the crop-based VLM, achieving a balanced trade-off between precision and recall. We further evaluate the impact of the verification step by generating I2I data without applying VLM filtering. As shown in Table 6-(b), generating I2I data without VLM verification degrades both overall and rare-class performance, highlighting the importance of semantic consistency filtering for reliable synthetic supervision. Qualitative examples of filtered and retained samples are shown in Fig. 6.

Beyond semantic verification, we validate the mask annotation quality of VRAIN using the ORIDa benchmark [15], which contains 108k real before/after image pairs with physically placed objects and ground-truth annotations. By treating the before and after images as I_{real} and I_{edit} respectively, we apply

Method	AP^{box}	AP^{mask}	AP_r^{box}	AP_r^{mask}
(a) Real-only	47.5	42.3	41.4	36.8
(b) w/o $\mathcal{D}_{\text{I2I}}$	50.3	45.1	47.5	43.8
(c) Ours	**50.7**	**45.2**	**49.1**	**44.0**

Table 7: Ablation Study on I2I Dataset. Excluding $\mathcal{D}_{\text{I2I}}$ lowers overall and rare-class performance, emphasizing the value of high-fidelity, rare-class-targeted I2I data that collaboratively complements T2I supervision.

our annotation pipeline and compare the resulting masks against the ground-truth. On this benchmark, our method achieves 0.925 mAP on mask annotations, demonstrating high annotation fidelity comparable to human-level quality [9,35].

Contribution of the I2I Data. Finally, we assessed the overall contribution of the generated high-fidelity I2I data ($\mathcal{D}_{\text{I2I}}$) to the hybrid training framework. As shown in Table 7, excluding $\mathcal{D}_{\text{I2I}}$ leads to a noticeable performance drop, particularly for rare categories (AP_r^{box} 49.1 $\rightarrow$ 47.5, AP_r^{mask} 44.0 $\rightarrow$ 43.8). This confirms that $\mathcal{D}_{\text{I2I}}$ successfully provides targeted rare-class coverage that complements the T2I data. By integrating these reliable, instance-level annotations, the model learns more robust representations for underrepresented classes, further boosting overall accuracy.

5 Conclusion

We present a unified data synthesis framework that bridges T2I generation and context-aware I2I editing for large-vocabulary instance segmentation. Our approach effectively couples the diversity and scalability of T2I synthesis with the realism and precise rare-class supervision of I2I editing. Through prompt-consistent filtering and teacher-student refinement, the T2I branch achieves reliable label quality despite domain gaps, while the proposed VRAIN framework provides high-fidelity, contextually coherent rare-class augmentation that reinforces the teacher-student learning process. Experiments on LVIS show that this hybrid strategy surpasses existing T2I and paste-based baselines, particularly on rare categories, and scales favorably with larger backbones. We believe this integration of complementary generative paradigms offers a promising approach to boosting model performance across diverse and underrepresented categories in long-tailed instance segmentation.

Limitation. Our hybrid approach relies on off-the-shelf generative models, which do not perfectly adhere to input instructions. For example, the T2I branch uses prompt-consistent filtering to retain relevant instances, but it may still fail to realize certain categories specified in the prompt. Similarly, the I2I can ignore positional instructions and place objects incorrectly. Although these approaches help mitigate label noise and contextual inconsistency, achieving perfect fidelity and precise instance-level control across all synthetic data remains a persistent challenge.

References

1. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
2. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
3. Chang, N., Yu, Z., Wang, Y.X., Anandkumar, A., Fidler, S., Alvarez, J.M.: Image-level or object-level? a tale of two resampling strategies for long-tailed detection. In: International conference on machine learning. pp. 1463–1472. PMLR (2021)
4. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
5. De Brabandere, B., Neven, D., Van Gool, L.: Semantic instance segmentation for autonomous driving. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 478–480 (2017). https://doi.org/10.1109/CVPRW.2017.66
6. Fan, C., Zhu, M., Chen, H., Liu, Y., Wu, W., Zhang, H., Shen, C.: Divergen: Improving instance segmentation by learning wider data distribution with more diverse generative data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3986–3995 (2024)
7. Fu, S., Yang, Q., Mo, Q., Yan, J., Wei, X., Meng, J., Xie, X., Zheng, W.S.: Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14987–14997 (2025)
8. Guo, H., Zhu, H., Peng, S., Wang, Y., Shen, Y., Hu, R., Zhou, X.: Sam-guided graph cut for 3d instance segmentation. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLVIII. p. 234–251. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-73195-2_14, https://doi.org/10.1007/978-3-031-73195-2_14
9. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5356–5364 (2019)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. He, Y.Y., Zhang, P., Wei, X.S., Zhang, X., Sun, J.: Relieving long-tailed instance segmentation via pairwise class balance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7000–7009 (2022)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
13. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
14. de Jorge, P., Volpi, R., Dokania, P.K., Torr, P.H., Rogez, G.: Placing objects in context via inpainting for out-of-distribution segmentation. In: European Conference on Computer Vision. pp. 456–473. Springer (2024)

15. Kim, J., Han, S., Jeong, J., Choi, J., Kim, D., Kim, S.J.: Orida: Object-centric real-world image composition dataset. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3051–3060 (2025)
16. Kimhi, M., Vainshtein, D., Baskin, C., Di Castro, D.: Robot instance segmentation with few annotations for grasping. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 7939–7949. IEEE (2025)
17. Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024)
18. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
19. Li, Y., Keuper, M., Zhang, D., Khoreva, A.: Adversarial supervision makes layout-to-image diffusion models thrive. In: ICLR (2024)
20. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
21. Liu, C., Li, X., Ding, H.: Referring image editing: Object-level image editing via referring expressions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13128–13138 (June 2024)
22. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
23. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
24. Loiseau, T., Vu, T.H., Chen, M., Pérez, P., Cord, M.: Reliability in semantic segmentation: Can we use synthetic data? In: European Conference on Computer Vision. pp. 442–459. Springer (2024)
25. Luo, W., Yang, S., Zhang, X., Zhang, W.: Siedob: Semantic image editing by disentangling object and background. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1868–1878 (June 2023)
26. Molina, J.M., Llerena, J.P., Usero, L., Patricio, M.A.: Advances in instance segmentation: Technologies, metrics and applications in computer vision. Neurocomputing **625**, 129584 (2025). https://doi.org/https://doi.org/10.1016/j.neucom.2025.129584, https://www.sciencedirect.com/science/article/pii/S0925231225002565
27. Nguyen, Q., Vu, T., Tran, A., Nguyen, K.: Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. Advances in Neural Information Processing Systems **36**, 76872–76892 (2023)
28. Pan, T.Y., Zhang, C., Li, Y., Hu, H., Xuan, D., Changpinyo, S., Gong, B., Chao, W.L.: On model calibration for long-tailed object detection and instance segmentation. Advances in Neural Information Processing Systems **34**, 2529–2542 (2021)
29. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
30. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)

32. Tan, J., Lu, X., Zhang, G., Yin, C., Li, Q.: Equalization loss v2: A new gradient balance approach for long-tailed object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1685–1694 (2021)
33. Tan, J., Wang, C., Li, B., Li, Q., Ouyang, W., Yin, C., Yan, J.: Equalization loss for long-tailed object recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11662–11671 (2020)
34. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
35. Tschirschwitz, D.E., Rodehorst, V.: Label convergence: Defining an upper performance bound in object recognition through contradictory annotations. In: Proceedings of the Winter Conference on Applications of Computer Vision. pp. 6848–6857 (2025)
36. Wang, T., Li, Y., Kang, B., Li, J., Liew, J.H., Tang, S., Hoi, S., Feng, J.: Classification calibration for long-tail instance segmentation. arXiv preprint arXiv:1910.13081 (2019)
37. Wang, T., Li, Y., Kang, B., Li, J., Liew, J., Tang, S., Hoi, S., Feng, J.: The devil is in classification: A simple framework for long-tail instance segmentation. arXiv preprint arXiv:2007.11978 (2020)
38. Wang, X., Darrell, T., Rambhatla, S.S., Girdhar, R., Misra, I.: Instancediffusion: Instance-level control for image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6232–6242 (2024)
39. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
40. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
41. Wu, S., Fei, H., Chua, T.s.: Universal scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14158–14168 (June 2025)
42. Wu, W., Zhao, Y., Chen, H., Gu, Y., Zhao, R., He, Y., Zhou, H., Shou, M.Z., Shen, C.: Datasetdm: Synthesizing data with perception annotations using diffusion models. Advances in Neural Information Processing Systems **36**, 54683–54695 (2023)
43. Wu, W., Zhao, Y., Shou, M.Z., Zhou, H., Shen, C.: Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1206–1217 (2023)
44. Xie, C., Mousavian, A., Xiang, Y., Fox, D.: Rice: Refining instance masks in cluttered environments with graph neural networks. In: Conference on Robot Learning. pp. 1655–1665. PMLR (2022)
45. Xie, J., Li, W., Li, X., Liu, Z., Ong, Y.S., Loy, C.C.: Mosaicfusion: Diffusion models as data augmenters for large vocabulary instance segmentation. International Journal of Computer Vision **133**(4), 1456–1475 (2025)
46. Xue, H., Huang, Z., Sun, Q., Song, L., Zhang, W.: Freestyle layout-to-image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14256–14266 (2023)
47. Yang, L., Xu, X., Kang, B., Shi, Y., Zhao, H.: Freemask: Synthetic images with dense annotations make stronger segmentation models. Advances in Neural Information Processing Systems **36**, 18659–18675 (2023)

48. Ye, H., Kuen, J., Liu, Q., Lin, Z., Price, B., Xu, D.: Seggen: Supercharging segmentation models with text2mask and mask2img synthesis. In: European Conference on Computer Vision. pp. 352–370. Springer (2024)
49. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
50. Zhang, Y., Yin, M., Bi, W., Yan, H., Bian, S., Zhang, C.H., Hua, C.: Zisvfm: Zero-shot object instance segmentation in indoor robotic environments with vision foundation models. IEEE Transactions on Robotics (2025)
51. Zhao, H., Sheng, D., Bao, J., Chen, D., Chen, D., Wen, F., Yuan, L., Liu, C., Zhou, W., Chu, Q., et al.: X-paste: Revisiting scalable copy-paste for instance segmentation using clip and stablediffusion. In: International Conference on Machine Learning. pp. 42098–42109. PMLR (2023)
52. Zhao, Y., Chen, S., Chen, Q., Hu, Z.: Combining loss reweighting and sample resampling for long-tailed instance segmentation. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023). https://doi.org/10.1109/ICASSP49357.2023.10094303
53. Zhou, X., Koltun, V., Krähenbühl, P.: Probabilistic two-stage detection. arXiv preprint arXiv:2103.07461 (2021)
54. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)
55. Zong, Z., Song, G., Liu, Y.: Detrs with collaborative hybrid assignments training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6748–6758 (2023)