A Appendix

A.1 Training Details

Pseudo mask preparation details. Empirically, in the divide stage, we set the confidence threshold $\tau = 0.3$; in the conquer stage, we choose threshold $\theta_{merge} = [0.6, 0.5, 0.4, 0.3, 0.2, 0.1]$. For each image, the divide-and-conquer pipeline generates on average 334 pseudo masks. In the self-training phase, the $\tau_{\text{self-train}} = 0.7$, and each image has 448 pseudo masks per image after merging high-confidence mask predictions generated by UnSAM. When merging the pseudo masks with the ground truths for training UnSAM+, we select $\tau_{\text{UnSAM+}} = 0.02$.

Whole-image segmentation. UnSAM picks DINO [8] pre-trained ResNet-50 [18] as the backbone and Mask2former [9] as the mask decoder. Given the abundant number of pseudo masks generated, UnSAM augments data only by cropping a 1024×1024 region from the original image. To cope with a large amount of 'ground-truth' masks per image, we find that having 2000 learnable queries produces the best result. We randomly select at most 200 'ground-truth' masks per image to speed up the training process. The default learning rate is 5×10^{-5} with batch size equals 16 and weight decay 5×10^{-2}. We train the model for 8 epochs. All model training in this paper was conducted using either 4 A100 GPUs or 8 RTX 3090 GPUs.

Promptable segmentation. UnSAM uses the self-supervised pretrained Swin-Transformer [25], specifically the Swin-Tiny model, as the backbone and leverages Semantic-SAM [23] as the base model. Given at most 6 levels of masks corresponding to one input point in SA-1B [21], we set the number of hierarchy levels to 6, which is also the number of predicted masks UnSAM generates per prompt during inference. However, one can easily train with a different number of granularity levels as needed. The default learning rate is 1×10^{-4} with a batch size of 8. The learning rate decreases by a factor of 10 at 90% and 95% of the training iterations. We train the model for 4 epochs.

A.2 Preliminary: Cut and Learn (CutLER) and MaskCut

CutLER [39] introduces a cut-and-learn pipeline to precisely segment instances without supervision. The initial phase, known as the cut stage, uses a normalized cut-based method, MaskCut [39], to generate high-quality instance masks that serve as pseudo-labels for subsequent learning phases. MaskCut begins by harnessing semantic information extracted from "key" features K_i of patch i in the last attention layer of unsupervised vision transformers. It then calculates a patch-wise cosine similarity matrix $W_{ij} = \frac{K_i K_j}{|K_i|_2 |K_j|_2}$. To extract multiple instance masks from a single image, MaskCut initially applies Normalized Cuts [31], which identify the eigenvector x corresponding to the second smallest eigenvalue. The vector x is then bi-partitioned to extract the foreground instance mask M^s. Subsequent iterations repeat this operation but adjust by masking out patches from previously segmented instances in the affinity matrix: $W_{ij}^t = \frac{\left(K_i \sum_{s=1}^{t} M_{ij}^s\right)\left(K_j \sum_{s=1}^{t} M_{ij}^s\right)}{\|K_i\|_2 \|K_j\|_2}$

Subsequently, CutLER's learning stage trains a segmentation/detection model with drop-loss, which encourages the model to explore areas not previously identified by MaskCut. An iterative self-training phase is employed for continuously refining the model's performance.

A.3 Preliminary: Segment Anything Model (SAM) and SA-1B

Inspired by achievement in the NLP field, the Segment Anything project [21] introduces the novel *promptable segmentation task*. At its core lies the Segment Anything Model (SAM) [21], which is capable of producing segmentation masks given user-provided text, points, boxes, and masks in a zero-shot manner. SAM comprises three key components: an MAE [17] pre-trained Vision Transformer [14] that extracts image embeddings, the prompt encoders that embed various types of prompts, and a lightweight Transformer [36] decoder that predicts segmentation masks by integrating image and prompt embeddings.

One significant contribution of SAM [21] is the release of the SA-1B dataset, which comprises 11 million high-resolution images and 1.1 billion segmentation masks, providing a substantial resource for training and evaluating segmentation models. In particular, annotators interactively used SAM to annotate images, and this newly annotated data was then utilized to iteratively update SAM. This cycle was repeated multiple times to progressively enhance both the model and the dataset.

While SAM [21] significantly accelerates the labeling of segmentation masks, annotating an image still requires approximately 14 seconds per mask. Given that each image contains over 100 masks, this equates to more than 30 minutes per image, posing a substantial cost and making it challenging to scale up the training data effectively.

A.4   Evaluation Datasets

COCO (Common Objects in Context) [24] is a widely utilized object detection and segmentation dataset. It consists of 115,000 labeled training images, 5,000 labeled validation images, and more than 200,000 unlabeled images. Its object segmentation covers 80 categories and is mainly on the instance-level. We evaluate our model on COCO *Val2017* with 5000 validation images without training or fine-tuning on any images from the COCO training set. The metrics we choose are class-agnostic COCO style averaged precision and averaged recall for the whole-image inference task, and MaxIoU and OracleIoU for the promptable segmentation task.

SA-1B [21] consists of 11 million high-resolution (1500 on average) images and 1.1 billion segmentation masks, approximately 100 masks per image. All masks are collected in a class-agnostic manner with various subject themes including locations, objects, and scenes. Masks cover a wide range of granularity levels, from large scale objects to fine-grained details. In the whole-image inference task, we randomly selected 1000 SA-1B images that are not used to generate pseudo labels as the validation set.

LVIS (Large Vocabulary Instance Segmentation) [15] has 164,000 images with more than 1,200 categories and more than 2 million high-quality instance-level segmentation masks. It has a long tail distribution that naturally reveals a large number of rare categories. In the whole-image inference task, we evaluate our model using its 5000 validation images in a zero-shot manner.

EntitySeg [29] is an open-world, class-agnostic dataset that consists of 33277 images in total. There are on average 18.1 entities per image. More than 80% of its images are of high resolution with at least 1000 pixels for the width. EntitySeg also has more accurate boundary annotations. In the whole-image inference task, we evaluate our model with 1314 low resolution version images (800×1300 on average) in a zero-shot manner.

PACO (Parts and Attributes of Common Objects) [30] is a detection dataset that provides 641,000 masks for part-level entities not included in traditional datasets. It covers 75 object categories and 456 object-part categories. In the whole-image inference task, we evaluate our model with 2410 validation images in a zero-shot manner.

PartImageNet [16] is a large-scale, high quality dataset with rich part segmentation annotations on a general set of classes with non-rigid, articulated objects. It includes 158 classes and 24,000 images from ImageNet [13]. In the whole-image inference task, we evaluate our model with 2956 validation images in a zero-shot manner.

ADE20K [48] is composed of 25,574 training and 2,000 testing images spanning 365 different scenes. It mainly covers semantic-level segmentation with 150 semantic categories and 707,868 objects from 3,688 categories. In the whole-image inference task, we evaluate our model with 2000 testing images in a zero-shot manner.

A.5   More Visualizations

We provide more qualitative results of UnSAM and UnSAM+ in a zero-shot manner in Figure A1, Figure A2, and Figure A3.

Figure A1: More visualizations on SA-1B [21]. From top to bottom are raw images, segmentation by SAM, segmentation by UnSAM, and segmentation by UnSAM+.
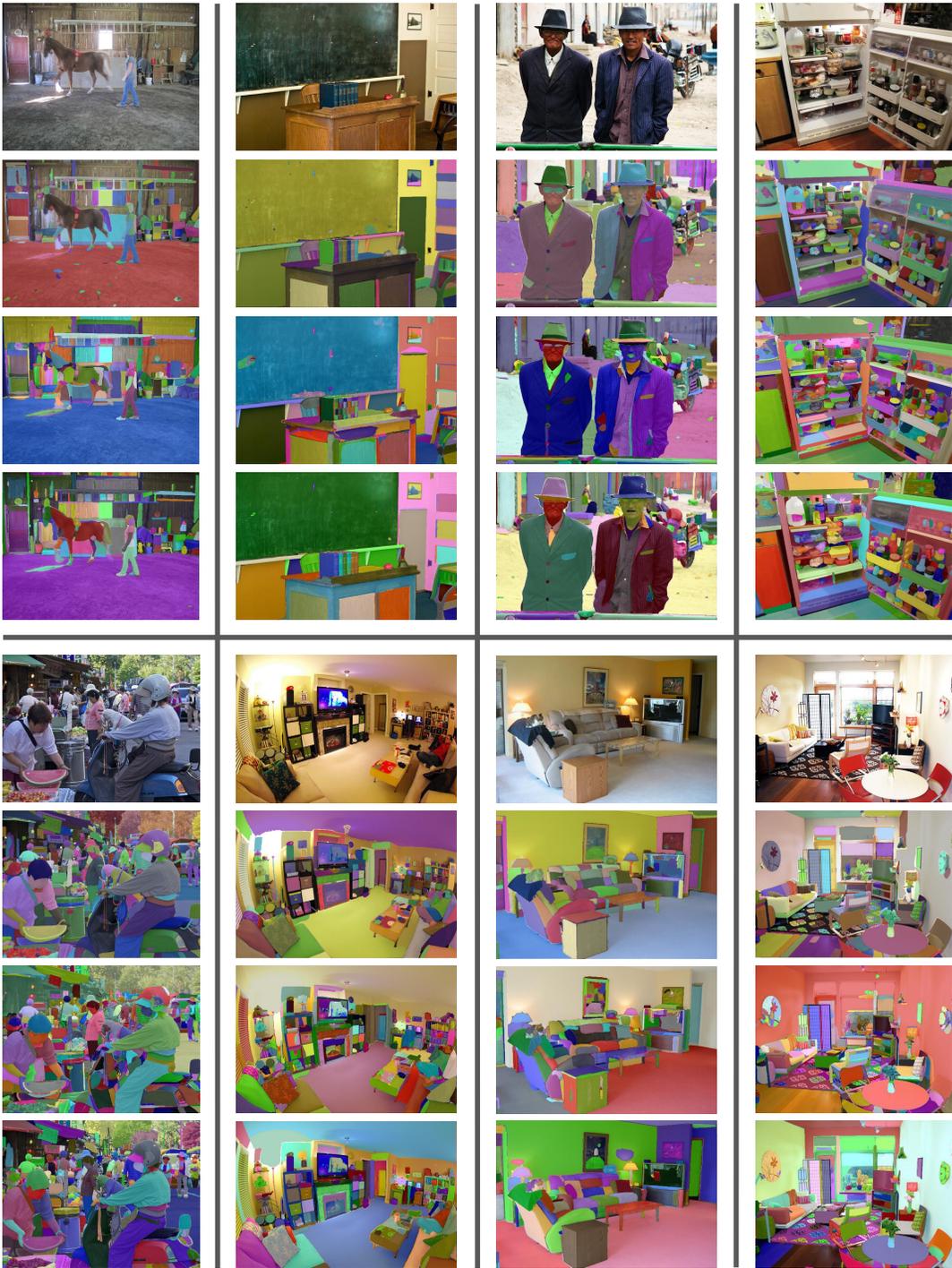
Figure A2: More visualizations on COCO [24]. From top to bottom are raw images, segmentation by SAM, segmentation by UnSAM, and segmentation by UnSAM+.
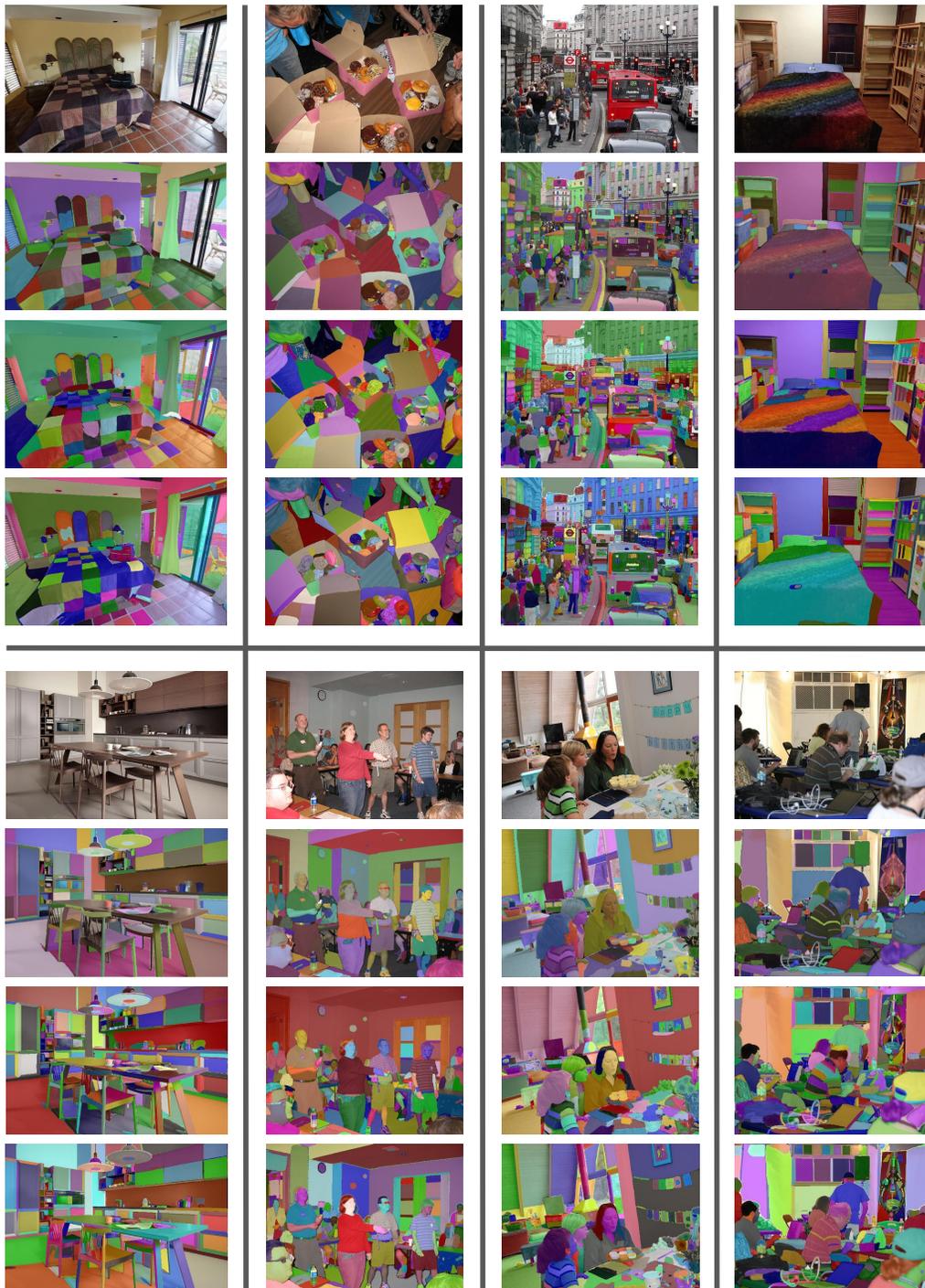
Figure A3: More visualizations on PACO [30]. From top to bottom are raw images, segmentation by SAM, segmentation by UnSAM, and segmentation by UnSAM+.

Figure A4: Failure cases of UnSAM. From left to right are raw images, segmentation by SAM, and segmentation by UnSAM.

A.6 Limitations

In images with very dense fine-grained details, UnSAM tends to miss repetitive instances with similar texture. As showed in Figure A4, in the first row, although UnSAM accurately segments the leaves in the center of the picture, it misses some leaves located at the top of the image. Additionally, UnSAM occasionally over-segment images. In the second row, the right sleeve cuff of the dancer has meaningless segmentation masks. This issue mainly arises because the unsupervised clustering method mistakenly considers some information, such as folds and shadows on clothing, as criteria for distinguishing different entities. In contrast, human annotators can use prior knowledge to inform the model that such information should not be valid criteria. In this regard, unsupervised methods still need to close the gap with supervised methods.

A.7 Ethical Considerations

We train UnSAM and UnSAM+ on ground truths of and pseudo masks generated on SA-1B [21]. SA-1B contains licensed images that are filtered for objectionable content. It is geographically diverse, but some regions and economic groups are underrepresented. Downstream use of UnSAM and UnSAM+ may create their own potential biases concerns for specific use cases.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist",**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in our abstract and introduction precisely summarize this paper's contributions, assumptions, and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We discuss limitations of our method in appendix A.6.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All information needed to reproduce the main experimental results is included in Appendix A.1.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will open-source the codes and models.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All basic settings of the pseudo mask preparation process and model training are included in Appendix A.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Following established protocols from prior studies, we report our experimental results. We observed robustness in our results against the variability of random seeds used for initializing model weights. Additionally, our pseudo-mask generation process does not require retraining a parameterized model, thus ensuring deterministic results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide full information on the compute resources we use in Appendix A.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The research conducted in this paper fully conform with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Broader impacts of our research are discussed in Appendix A.7.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: [Yes]

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper doesn't include crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper doesn't include crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.