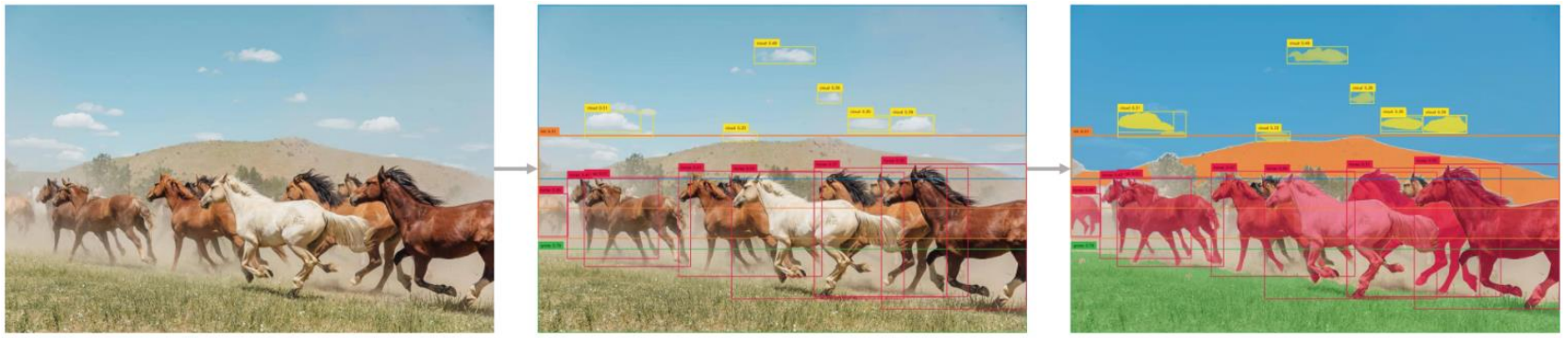
# XX-Anything:
# A demo survey on recent fundamental models and applications

2021-2023

# Introduction

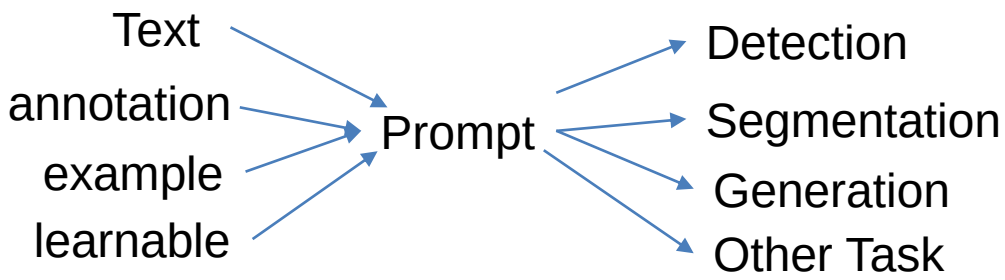
- **Segment Anything Seires**
- **Visual Grounding Seires**



**Text Prompt:**
**"Horse. Clouds. Grasses. Sky. Hill."**

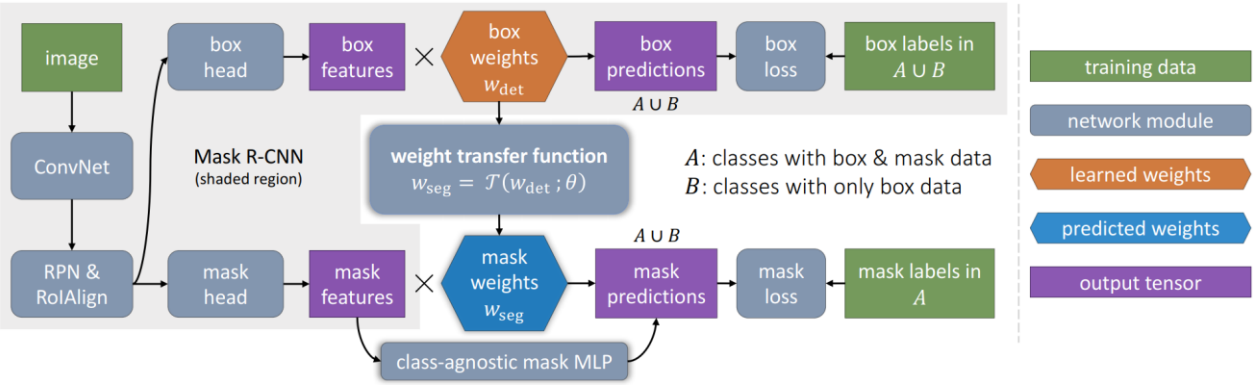
**Grounding DINO:**
**Detect Everything**

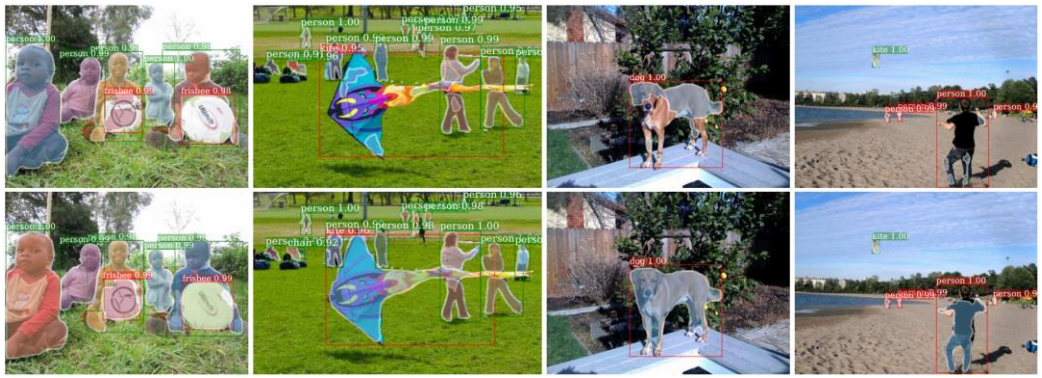
**Grounded-SAM:**
**Detect and Segment Everything**



# Segment Anything

- **Learning to segment every thing (2018)**
  - **Mask RCNN + box-to-mask transfer**
  - **Utilize box labels -> 3000 classes segmentation**



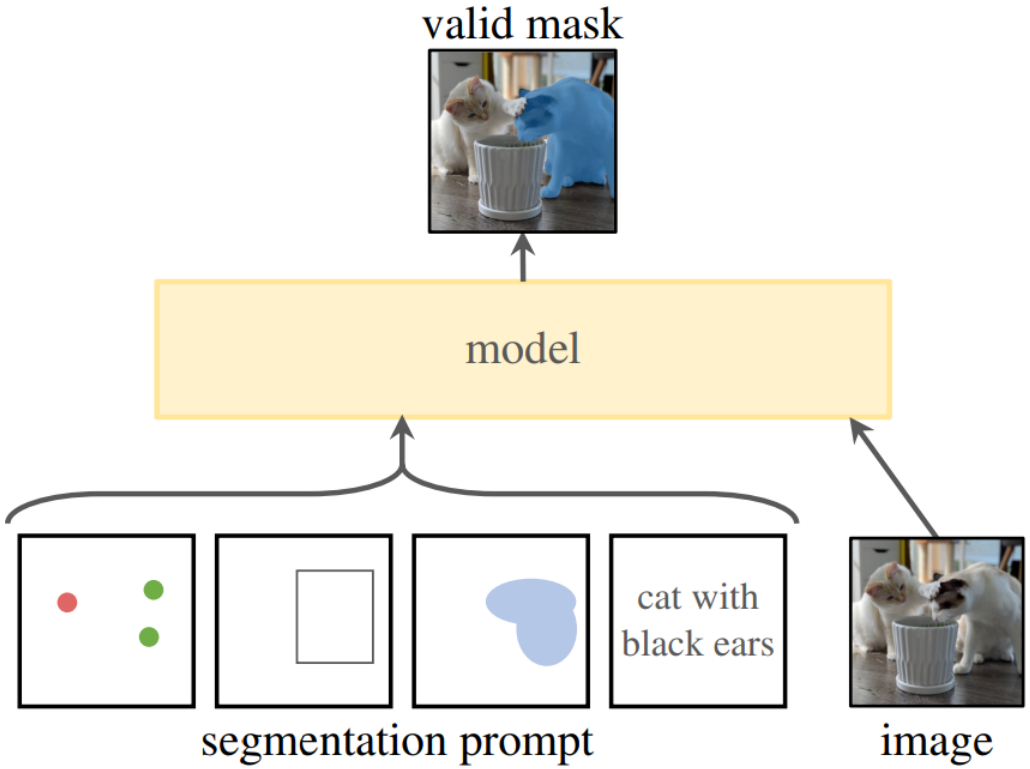



# Segment Anything

- **Learning to segment every thing**
  - Our work is complementary in the sense that bottom-up segmentation methods may be used to infer training masks for our weakly-labeled examples. We leave this extension to future work.
  - Scaling instance segmentation to thousands of categories, without full supervision, is an extremely challenging problem with ample opportunity for improved methods.

# Segment Anything

- **SAM（Segment Anything Model）**
  - **weak prompts -> mask**



# Segment Anything

- **SAM**
  - Limitations：

（1）It can miss fine structures, hallucinates small disconnected components at times, and does not produce boundaries as crisply as more computationally intensive methods that “zoom-in”

（2）SAM can process prompts in real-time, but nevertheless SAM’s overall performance is not real-time when usinga heavy image encoder

（3）Our foray into the text-to-mask task is exploratory and not entirely robust

（4）Unclear how to design simple prompts that implement semantic and panoptic segmentation

  - Outlooks：

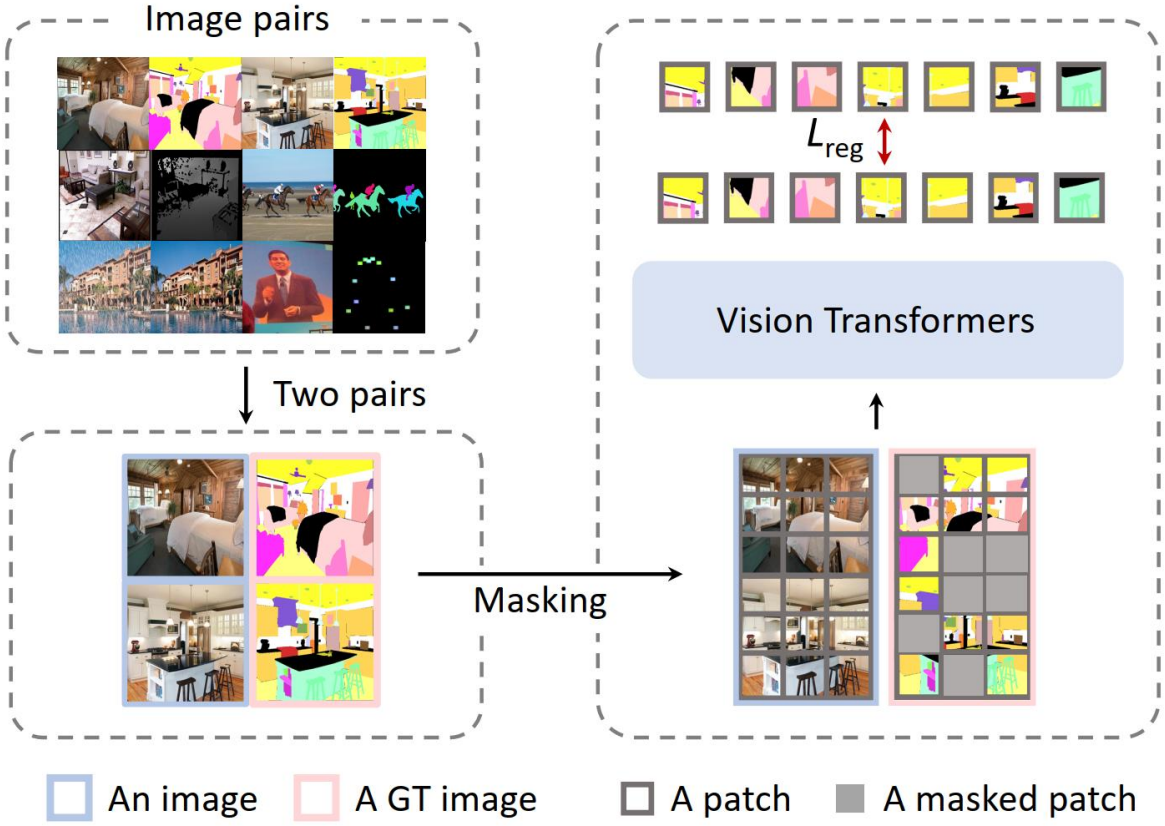
（1）We expect dedicated interactive segmentation methods to outperform SAM when many points are provided

（2）There are domain-specific tools, such as ‘ilastik’, that we expect to outperform SAM in their respective domains

（3） The release of over 1B masks, and our promptable segmentation model will help pave the path ahead

# Segment Anything

- **Painter（2023）**
  - **Examples as prompts -> dense output**



# Segment Anything

- **Painter**
  - Limitations：

（1）There is still much room for boosting our approach especially on the difficult task of panoptic segmentation

（2）Our approach is designed based on visual signals as contexts, this general interface does not seem natural for modeling language signals.

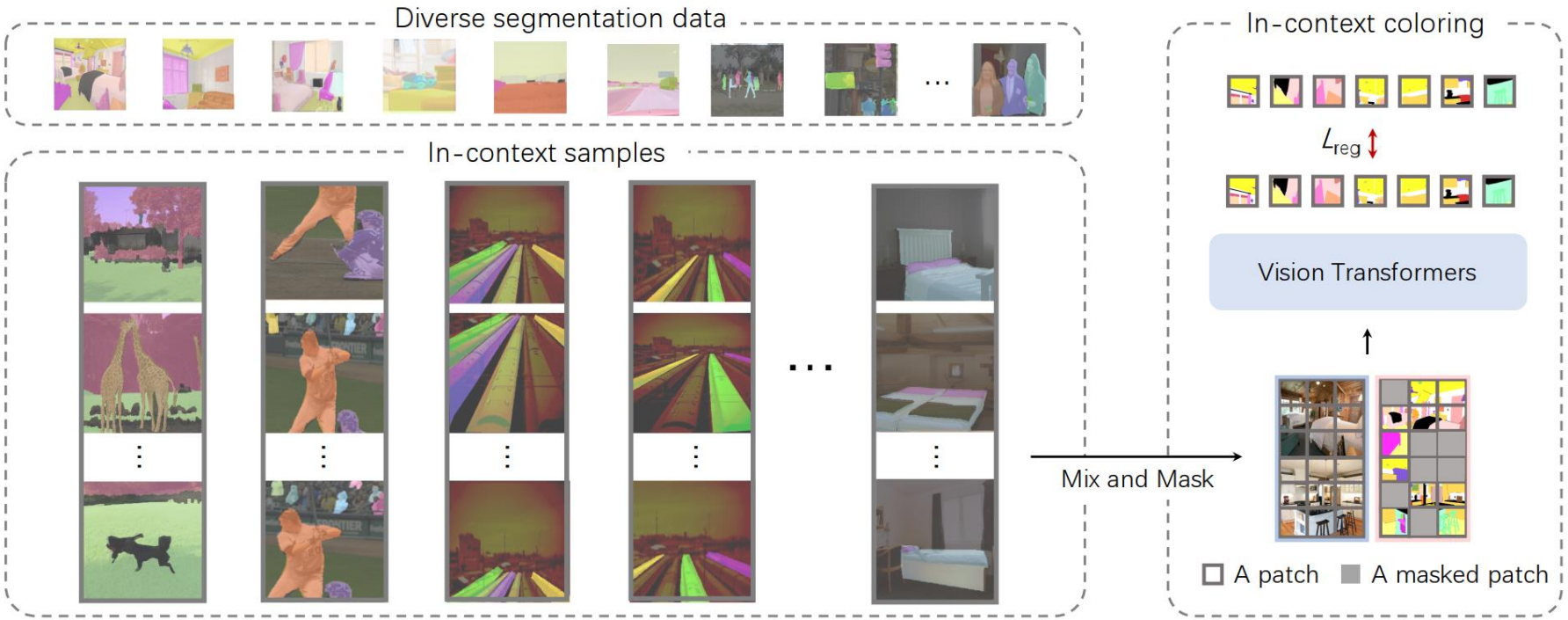
  - Outlooks：

（1）How to model discrete language signals as continuous ones seems to be an impressive direction, and some work has started to emerge recently

（2）We believe that the best GPT-3 moment in the vision field is yet to come.

# Segment Anything

- **SegGPT（2023）**
  - **Examples as prompts -> mask**



# Segment Anything

- **SegGPT**

| method | venue | YouTube-VOS 2018 [52] | | | | | DAVIS 2017 [37] | | | MOSE [12] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $G$ | $J_s$ | $F_s$ | $J_u$ | $F_u$ | $J\&F$ | $J$ | $F$ | $J\&F$ | $J$ | $F$ |
| *with video data* | | | | | | | | | | | | |
| AGAME [21] | CVPR'19 | 66.0 | 66.9 | - | 61.2 | - | 70.0 | 67.2 | 72.7 | - | - | - |
| AGSS [29] | ICCV'19 | 71.3 | 71.3 | 65.5 | 75.2 | 73.1 | 67.4 | 64.9 | 69.9 | - | - | - |
| STM [36] | ICCV'19 | 79.4 | 79.7 | 84.2 | 72.8 | 80.9 | 81.8 | 79.2 | 84.3 | - | - | - |
| AFB-URR [27] | NeurIPS'20 | 79.6 | 78.8 | 83.1 | 74.1 | 82.6 | 74.6 | 73.0 | 76.1 | - | - | - |
| RDE [25] | CVPR'22 | 83.3 | 81.9 | 86.3 | 78.0 | 86.9 | 86.1 | 82.1 | 90.0 | 48.8 | 44.6 | 52.9 |
| SWEM [31] | CVPR'22 | 82.8 | 82.4 | 86.9 | 77.1 | 85.0 | 84.3 | 81.2 | 87.4 | 50.9 | 46.8 | 54.9 |
| XMem [9] | ECCV'22 | 86.1 | 85.1 | 89.8 | 80.3 | 89.2 | 87.7 | 84.0 | 91.4 | 57.6 | 53.3 | 62.0 |
| *without video data* | | | | | | | | | | | | |
| Painter | CVPR'23 | 24.1 | 27.6 | 35.8 | 14.3 | 18.7 | 34.6 | 28.5 | 40.8 | 14.5 | 10.4 | 18.5 |
| SegGPT | this work | 74.7 | 75.1 | 80.2 | 67.4 | 75.9 | 75.6 | 72.5 | 78.6 | 45.1 | 42.2 | 48.0 |

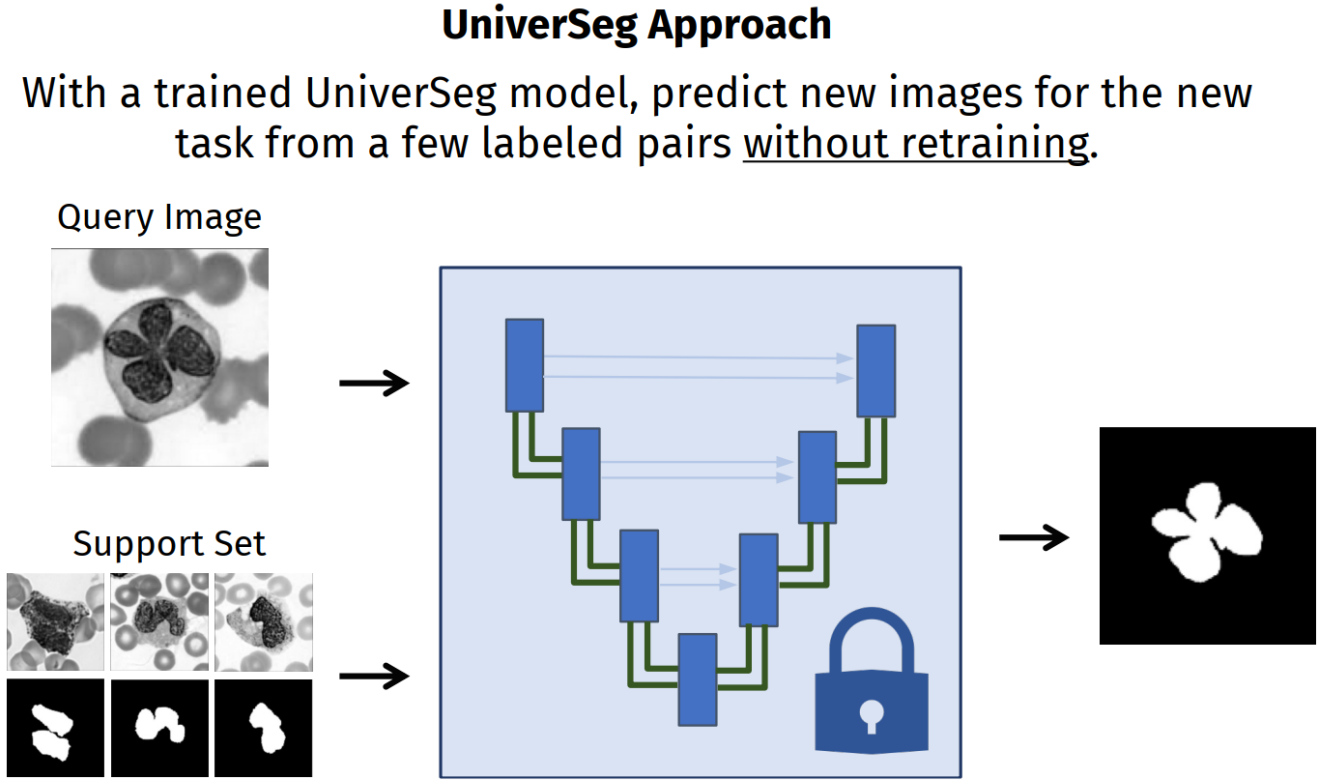
- Outlooks:

（1）We could optimize a prompt image for a specific scene, e.g., your apartment, or a specific character, e.g., Bert's face. This opens up opportunities for a broad range of applications

（2）Different from the in-context tuning, we only randomly select several samples in the training set as examples，...，These experiments inspire us to explore in the future what makes good examples and how many examples we need to approach the results of in-context tuning.

# Segment Anything

- **UniverSeg （2023）**
  - **Examples as prompts -> mask**



# Segment Anything

- **UniverSeg**
  - Limitations：

Using 2D data and single labels.

  - Outlooks：

（1）We are excited by future extensions to segment 3D volumes using 2.5D or 3D models and multi-label maps, and further closing the gap with the upper bounds.

（2）easily adapt to new segmentation tasks determined by scientists and clinical researchers, without model retraining.

# Segment Anything

- **Summary**
  - Promptable，Prompt具有灵活的输入形式，是实现通用性的普遍做法。不同的prompt方式：（1）SAM的预定义形式（点/框/区域）（2）SegGPT、UniverSeg中的example形式。

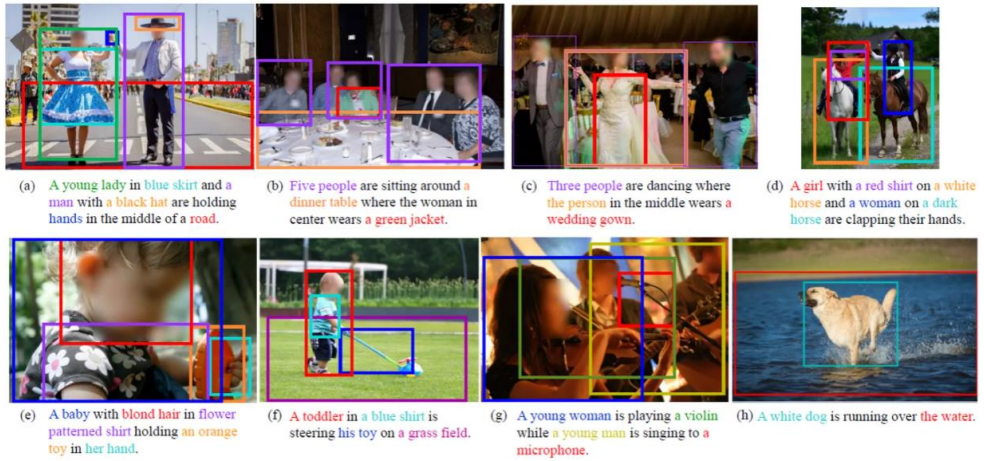
  - 性能上，当前通用分割模型在很多下游任务上仍然不及任务专用模型，作者也认为仍然存在很大改进空间：（1）视频目标分割，SegGPT vs XMem（74.7 vs 86.1）；（2）医学影像分割任务，UniverSeg vs nnUNet（71.8 vs 84.4）。短时间内，细分领域的研究应当不至于被通用分割模型完全覆盖，仍然有足够的发展空间。

  - 训练方式：普遍使用多样化任务数据，训练数据多样性对于实现新任务泛化应该是很重要的。框架设计：追求“大道至简”，简洁和统一的范式。

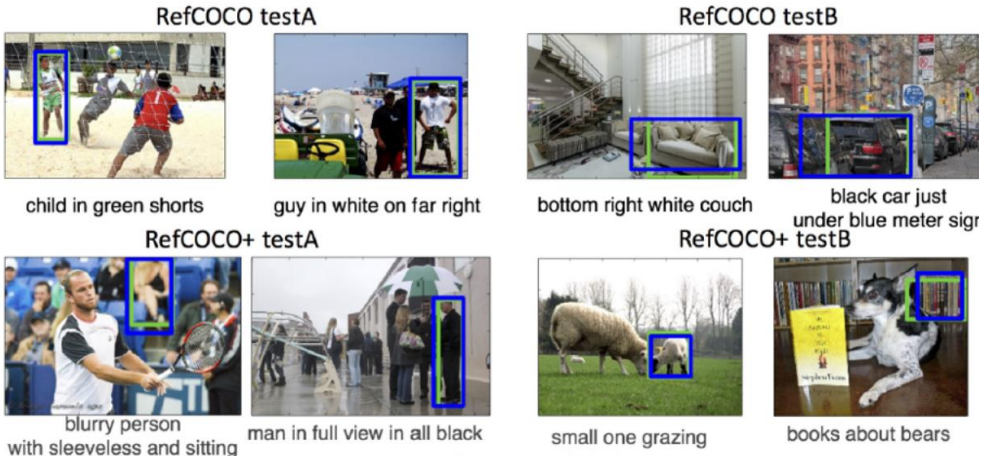
# Visual Grounding

- **Visual Grounding**
  - **Text -- > Detection**

**Phrase Localization**

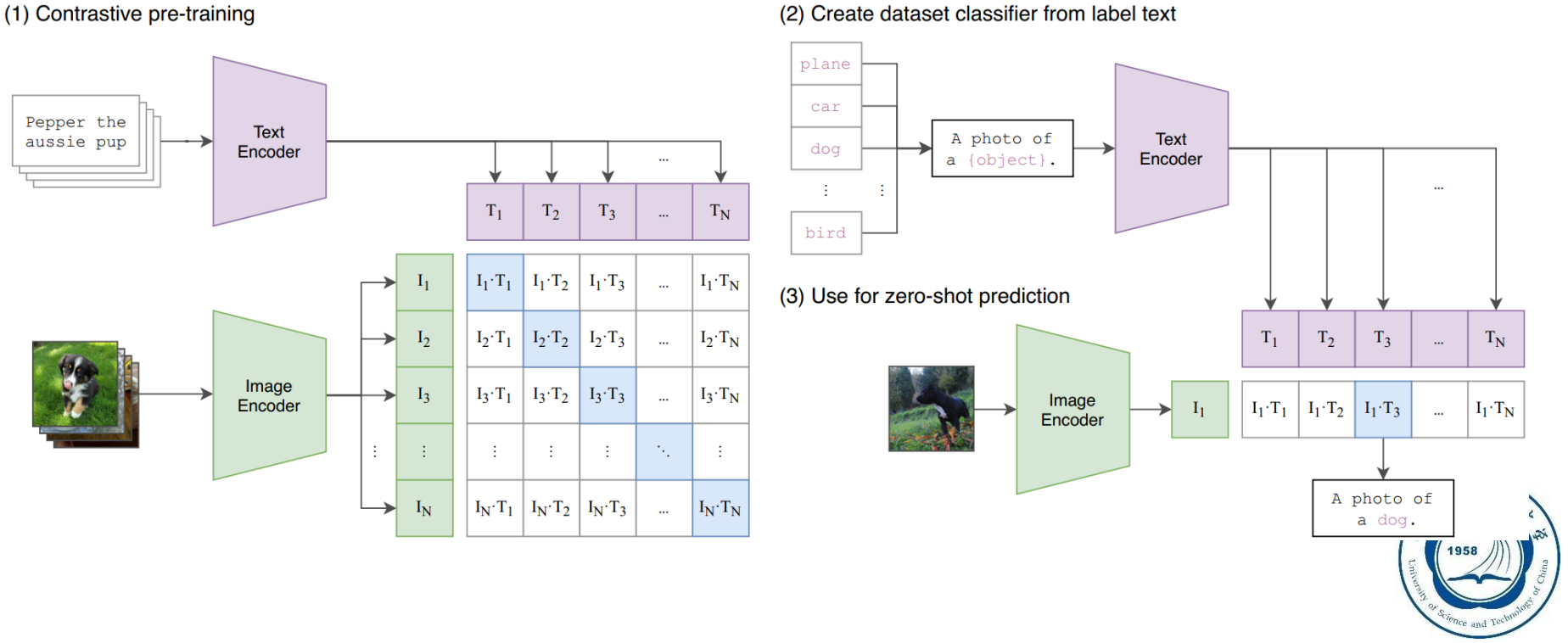



**Referring Expression Comprehension**



# Visual Grounding

- **CLIP（2021）**
  - **Text < -- > Image**
  - **symmetric loss**



# Introduction

- **BLIP（2022）**
  - **Image -- > Caption**
  - **Captioner + Filter**

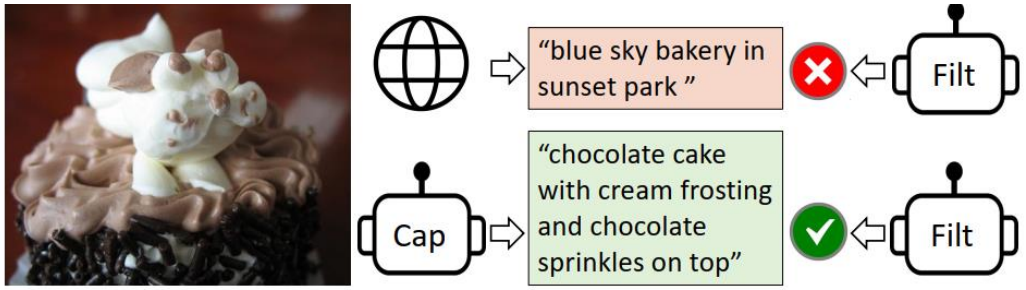
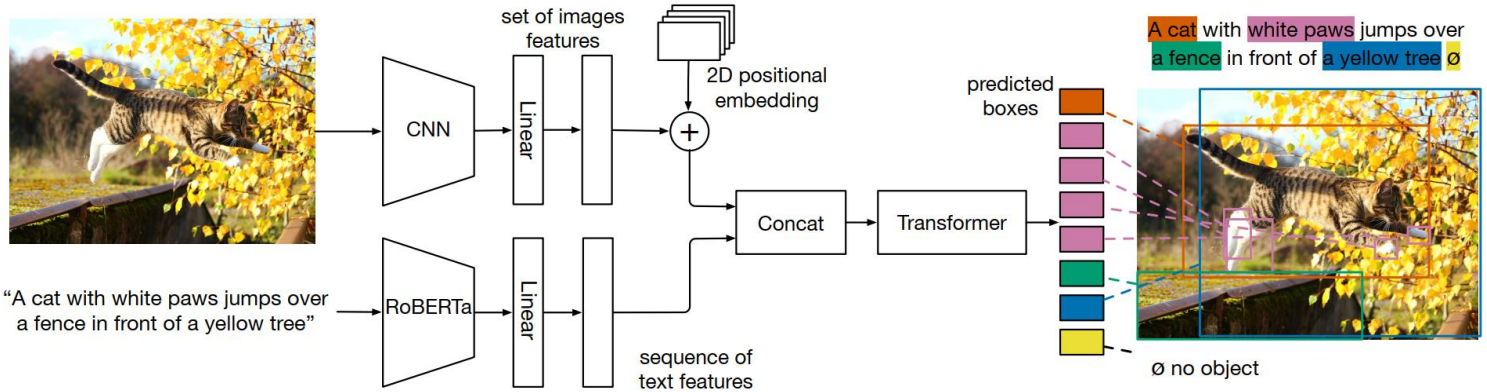



*Figure 1.* We use a Captioner (Cap) to generate synthetic captions for web images, and a Filter (Filt) to remove noisy captions.


*Figure 6.* Examples of the web text $T_w$ and the synthetic text $T_s$. Green texts are accepted by the filter, whereas red texts are rejected.

# Introduction

- **MDETR（2021）**
  - **Text -> Detection** （1.3M Image-text Pairs）



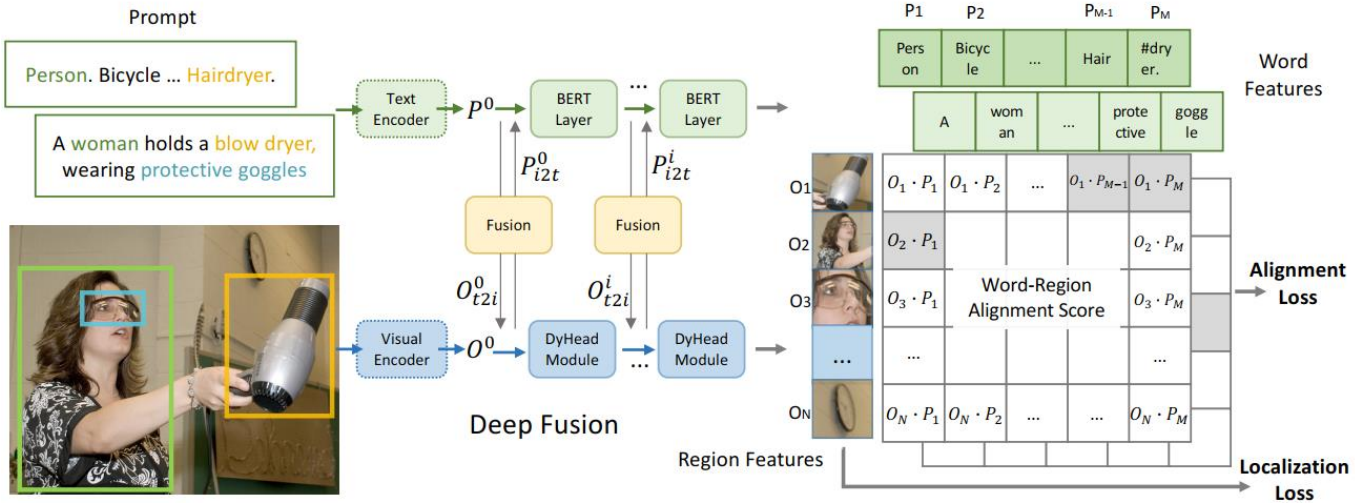


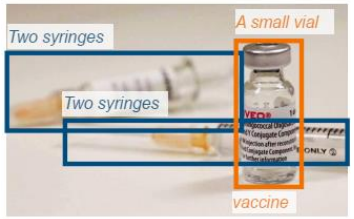
$$l_o = \sum_{i=0}^{N-1} \frac{1}{|T_i^+|} \sum_{j \in T_i^+} -\log\left(\frac{\exp(o_i^\top t_j/\tau)}{\sum_{k=0}^{L-1}\exp(o_i^\top t_k/\tau)}\right)$$

$$l_t = \sum_{i=0}^{L-1} \frac{1}{|O_i^+|} \sum_{j \in O_i^+} -\log\left(\frac{\exp(t_i^\top o_j/\tau)}{\sum_{k=0}^{N-1}\exp(t_i^\top o_k/\tau)}\right)$$

# Introduction

- **GLIP（2022）**
  - **Text -> Detection** （ 27M Image-text Pairs ）



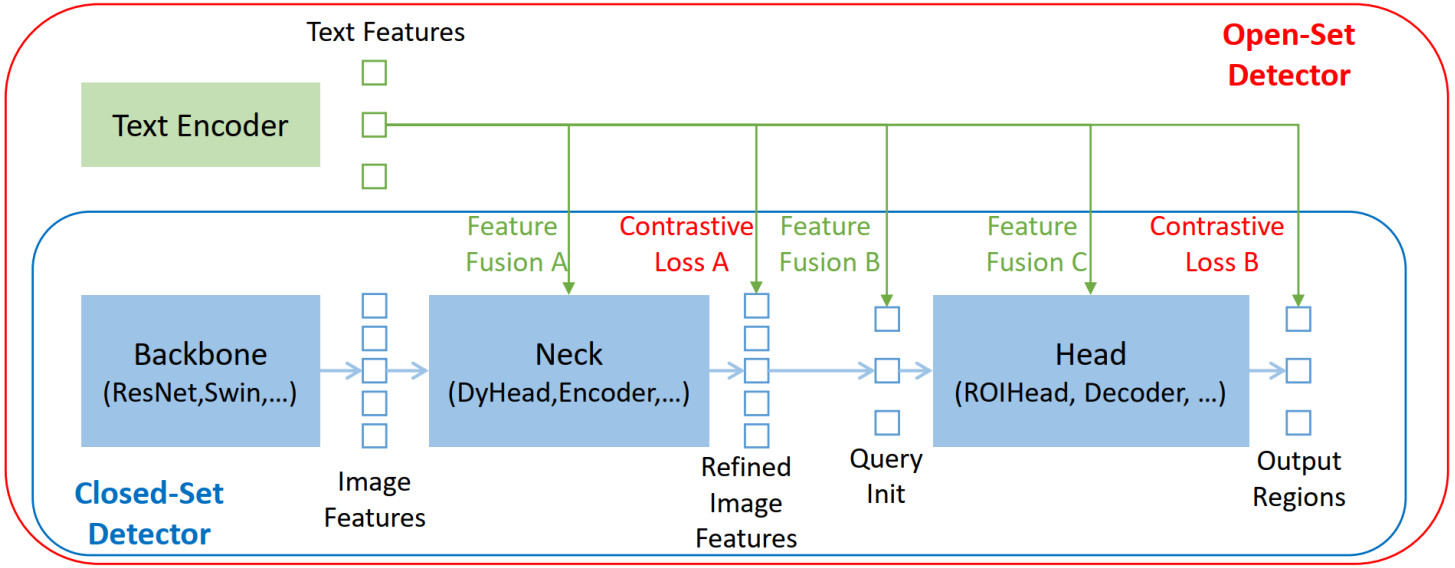




Two syringes and a small vial of vaccine.


playa esmeralda in holguin, cuba. the view from the top of the beach. beautiful caribbean sea turquoise

# Introduction

- **Grounding DINO（2023）**
  - **Text -> Detection**




| Model | Backbone | Pre-Training Data | Zero-Shot 2017val | Fine-Tuning 2017val/test-dev |
|---|---|---|---|---|
| DyHead-T† [5] | Swin-T | O365 | 43.6 | 53.3 / - |
| GLIP-T (B) [26] | Swin-T | O365 | 44.9 | 53.8 / - |
| GLIP-L [26] | Swin-L | FourODs,GoldG,Cap24M | 49.8 | 60.8 / 61.0 |
| DINO(Swin-T)† [58] | Swin-T | O365 | 46.2 | 56.9 / - |
| Grounding-DINO-T (Ours) | Swin-T | O365 | 46.7 | 56.9 / - |
| Grounding-DINO-L (Ours) | Swin-L | O365,OI,GoldG,Cap4M,COCO,RefC | **60.7** | **62.6** / - |

# Introduction

## Limitations & Outlooks

Limitations：

（1） MDETR: Under full fine-tuning on the whole training set, the performance on rare objects drops significantly, likely due to the extreme class imbalance. We expect that common techniques such as Repeat Factor Sampling will improve the situation in future work.

（2） Grounding DINO: Although the great performance on openset object detection setting, Grounding DINO cannot be used for segmentation tasks like GLIPv2. Moreover, our training data is less than the largest GLIP model, which may limit our final performance.

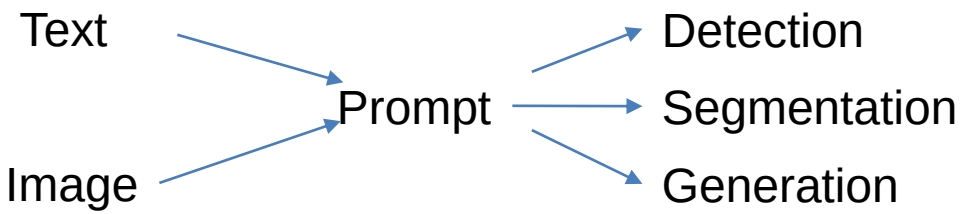
Outlooks：

（1） GLIP: We leave a detailed study of how GLIP scales with text-image data size to future work.

（2） Grounding DINO: A larger-scale training will be left as our future work

# Introduction

- **Summary**
  - Grounding methods: Text -> Detection/Segmentation
  - Visual-Language Pre-training: Image -> Text
  - Segment Anything: Prompt -> Segmentation/New Task



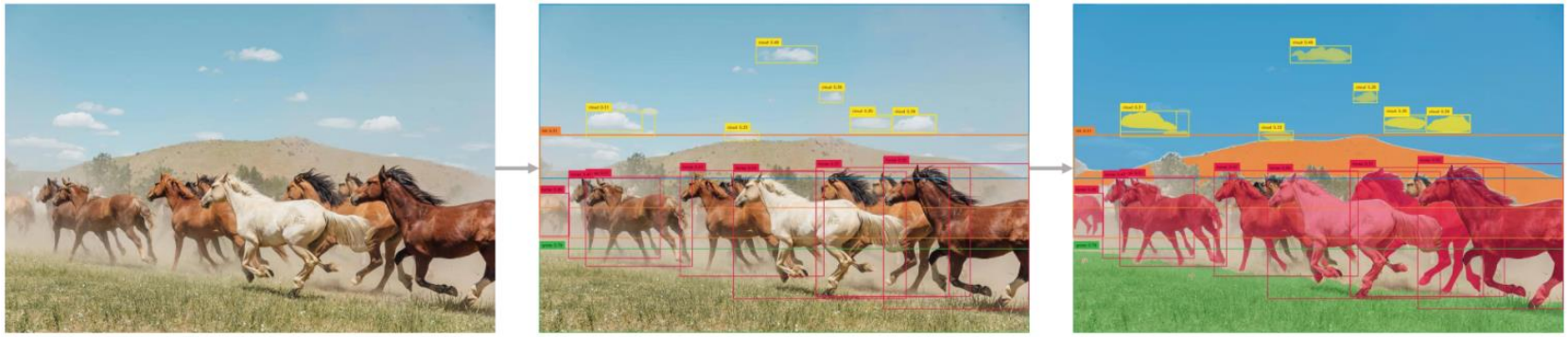

  - Promptable
  - Text - Image Alignment
  - Utilize Language Model for Visual Task

# Introduction

- **Develop**
  - 下游任务结合通用模型提升性能：提供初始化/蒸馏/额外信息
  - 弥补通用模型的不足：下游任务的专门设计、微调、新的学习方式
  - 通过模型组合可以实现新的通用视觉应用。

# Application

- **Grounded SAM**
  - **Grounding DINO：text-based open-set detector**
  - **SAM：segment in the bbox**
  - **Stable Dufusion：inpaint the masked image**
  - **SAM + Grounding DINO + Stable Difussion = Grounded Segment Anything**



**Text Prompt:**
**"Horse. Clouds. Grasses. Sky. Hill."**

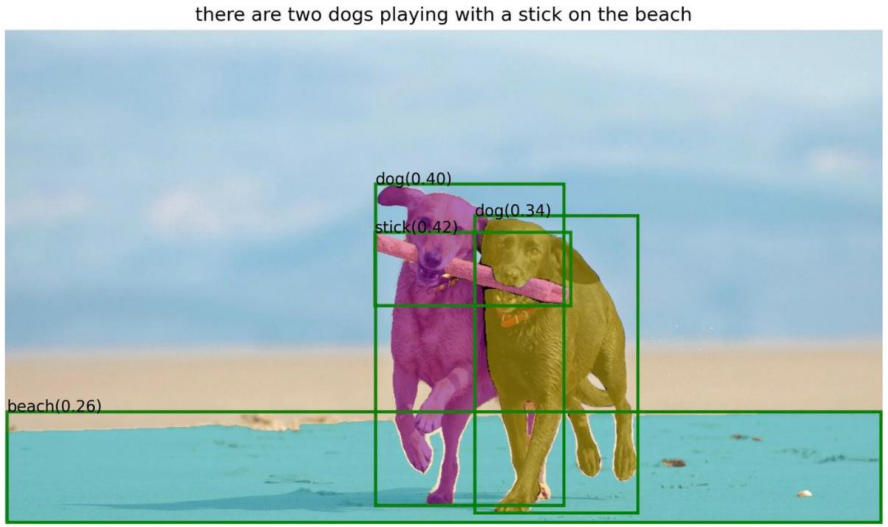
**Grounding DINO:**
**Detect Everything**

**Grounded-SAM:**
**Detect and Segment Everything**

https://github.com/IDEA-Research/Grounded-Segment-Anything

# Application

- **Grounded SAM**
  - **BLIP：Image -> caption**
  - **Grounded SAM：caption -> detection & masks**
  - **BLIP + Grounded SAM = Automatic Labeling**




https://github.com/IDEA-Research/Grounded-Segment-Anything

# Application

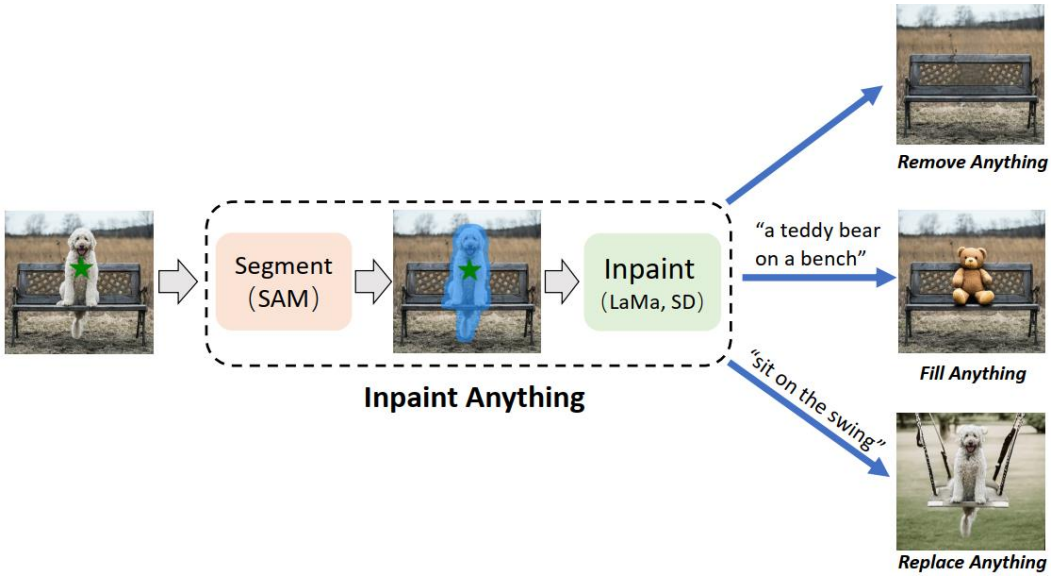
- **Tracking Anything**
  - **SAM：click -> target mask**
  - **VOS：track and segment the target**
  - **SAM + VOS = Tracking and Segment Anything**

https://github.com/gaomingqi/Track-Anything

Yang J, Gao M, Li Z, et al. Track Anything: Segment Anything Meets Videos[J]. arXiv preprint arXiv:2304.11968, 2023.

# Application

- **Inpaint Anything**
  - **SAM：click -> target mask**
  - **Stable Diffusion： inpaint the masked image**
  - **SAM + Stable Diffusion = Remove/Fill/Replace Anything**

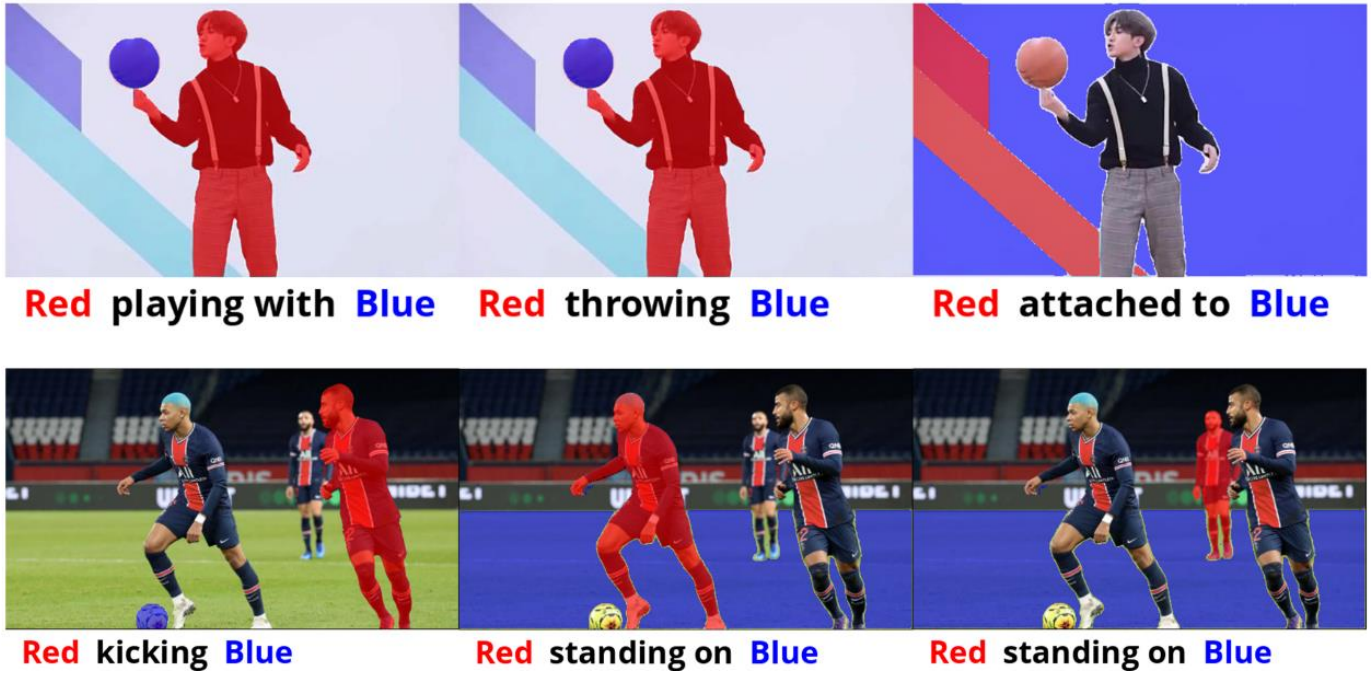



https://github.com/ geekyutao/Inpaint-Anything.

Yu T, Feng R, Feng R, et al. Inpaint anything: Segment anything meets image inpainting[J]. arXiv preprint arXiv:2304.06790, 2023.

# Application

- **Relate Anything**
  - **SAM： click -> target mask**
  - **Panoptic Scene Graph Generation（PSG）: relationship modeling**
  - **SAM + PSG = Relate Anything**




https://github.com/Luodian/RelateAnything

# Limitations

- Trade-off： Performance vs Usability

| Method | Venue | Initialization | Evaluation | DAVIS-2016-val | | | DAVIS-2017-test-dev | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | *J&F* | *J* | *F* | *J&F* | *J* | *F* |
| STM [12] | ICCV2019 | Mask | One Pass | 89.3 | 88.7 | 89.9 | 72.2 | 69.3 | 75.2 |
| AOT [15] | NeurIPS2021 | Mask | One Pass | 91.1 | 90.1 | 92.1 | 79.6 | 75.9 | 83.3 |
| XMem [1] | NeurIPS2022 | Mask | One Pass | 92.0 | 90.7 | 93.2 | 81.2 | 77.6 | 84.7 |
| SiamMask [14] | CVPR2019 | Box | One Pass | 69.8 | 71.7 | 67.8 | - | - | - |
| MiVOS [2] | CVPR2021 | Scribble | 8 Rounds | 91.0 | 89.6 | 92.4 | 78.6 | 74.9 | 82.2 |
| TAM (Proposed) | - | Click | One Pass | 88.4 | 87.5 | 89.4 | 73.1 | 69.8 | 76.4 |

- Substitutability
  - Grounded SAM： text -> mask
    - How about Inference Seg / SAM + CLIP?
  - How to form papers？

# Outlooks

- New scenes, new requirements.
  - light-weight、 interpretability 、 reasoning ...
  - adaption、combination、distillation ...
  - long tail、fine-grained ...
- Do something that not easy.
- Avoid domains covered by LFM.
- Take action now.