From Specialist to Generalist: Towards General Vision Understanding Interface

Jianwei Yang
Microsoft Research
06/19/2023
Q1: how to learn image representations?
Q2: how to extend vision models with more flexible, promptable interfaces?
Q3: how to do image generation?
Q4: how to train multimodal LLM?
Q5: how to chain multimodal experts with LLM?
Q1: how to learn image representations?

Q2: how to extend vision models with more flexible, promptable interfaces?

Q3: how to do image generation?

Q4: how to train multimodal LLM?

Q5: how to chain multimodal experts with LLM?
A Lesson from LLMs

NLP

Translation
Semantic Parsing
Question Answering
Summary Writing
A Lesson from LLMs

Decoder LLMs (e.g., GPT-3)

NLP

Unification
A Lesson from LLMs

Decoder LLMs (e.g., GPT-3) → Human-AI Interaction → Conversational AI (e.g., ChatGPT)

NLP

Translation
Semantic Parsing
Question Answering
Summarizing
A Lesson from LLMs

Decoder LLMs (e.g., GPT-3) → Human-AI Interaction → Conversational AI (e.g., ChatGPT)

NLP: Translation, Semantic Parsing, Question Answering, Summarization
A Lesson from LLMs

NLP
- Translation
- Semantic Parsing
- Question Answering
- Summarization

Vision
- Image captioning
- Classification
- Object detection
- Visual question answering
- Segmentation

Unification → Decoder LLMs (e.g., GPT-3) → Human-AI Interaction

2018-2022

Conversational AI (e.g., ChatGPT)
A Lesson from LLMs

NLP

- Translation
- Semantic Parsing
- Question Answering
- Summarization

Vision

- Image captioning
- Classification
- Object answering
- Visual question answering
- Segmentation

Decoder LLMs (e.g., GPT-3) → Human-AI Interaction → Conversational AI (e.g., ChatGPT)

Unification

2018-2022
Unique Challenges in Vision: Modeling
Unique Challenges in Vision: Modeling

a) Different types of inputs:
Temporality: static image, video sequence
Multi-modality: w/text, w/audio, etc.
Unique Challenges in Vision: Modeling

**a) Different types of inputs:**
- **Temporalinity**: static image, video sequence
- **Multi-modality**: w/text, w/audio, etc.

**b) Different granularities of tasks:**
- **Image-level**: classification, captioning, etc.
- **Region-level**: object detection, grounding, etc.
- **Pixel-level**: segmentation, depth, SR, etc.
Unique Challenges in Vision: Modeling

a) Different types of inputs:
   Temporality: static image, video sequence
   Multi-modality: w/text, w/audio, etc.

b) Different granularities of tasks:
   Image-level: classification, captioning, etc.
   Region-level: object detection, grounding, etc.
   Pixel-level: segmentation, depth, SR, etc.

c) Different types of outputs:
   Spatial: edges, boxes, masks, etc.
   Semantic: class labels, descriptions, etc.
Unique Challenges in Vision: Data

- Mask Annotation (COCO, LVSI)
- Box Annotation (COCO, O365)
- Image Annotation (ImageNet, LAION)

From poor to richer semantics
From coarse to finer grain

Scales differ significantly across different types of annotations
Clear Attempts towards General Vision
Clear Attempts towards General Vision

Closed-set Classification

AlexNet\[1\], ResNet\[2\], ViT\[3\]

Open-world Recognition

CLIP\[4\], ALIGN\[5\], FLORENCE\[6\]

Clear Attempts towards General Vision

Closed-set Classification  Open-world Recognition

Specialist Models  Generalist Models

Detection[1], Segmentation[2], VQA[3]

Pixel2Seqv2[4], UniTAB[5], OFA[6], Unified-IO[7], X-Decoder[8]

Clear Attempts towards General Vision

Closed-set Classification → Open-world Recognition

Specialist Models → Generalist Models

Representation Learning

BEIT[1], MAE[2], DINO[3]

Promptable Interface

SAM[4], SegGPT[5], SEEM[6]

Clear Attempts towards General Vision

Closed-set Classification

Open-world Recognition

Specialist Models

Generalist Models

Representation Learning

Promptable Interface
Clear Attempts towards General Vision

Open-world Recognition

Generalist Models

Promptable Interface
In this talk
In this talk

**Intuition:** language as the common space to share information
**Benefit:** Zero-shot transfer to novel vocabularies

**Intuition:** language, spatial prompts and beyond
**Benefit:** Reduce the ambiguity of expressing human intents

**Intuition:** vision is multi-task, multi-granularity
**Benefit:** Build synergy across task granularities

Bridge vision with language

Unify different granularities

Take various prompts
I. Bridge Vision with Language

Bridge vision with language

Unify different granularities

Take various prompts
Bridge Vision with Language

[Image Classification] e.g., CLIP [1] [Semantic]
[Object Detection] e.g., GLIP [2] [Label]
[Segmentation] e.g., MaskCLIP [3] [Granularity]

Language

Label

Image Classification Object Detection Segmentation

---

1. Radford et al. "Learning transferable visual models from natural language supervision." ICM, PMLR, 2021
2. Li et al. "Grounded language-image pre-training." CVPR, 2022
3. Zhou et al. "Extract Free Dense Labels from CLIP." ECCV, 2022
4. Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale.” ICLR, 2021
Bridge Vision with Language

(a) Converting labels to language is agnostic to granularity
(b) Coarse-grained knowledge can be transferred to fine-grained tasks

1. Radford et al. "Learning transferable visual models from natural language supervision." ICLR, PMLR, 2021
2. Li et al. "Grounded language-image pre-training." CVPR, 2022
3. Zhou et al. "Extract Free Dense Labels from CLIP." ECCV, 2022
4. Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale.” ICLR, 2021
Bridge Vision with Language

Bridge Vision with Language

Replace labels with concept names, and use text encoder to encode all concepts as they are language tokens

Image Classification  Object Detection  Segmentation

Language

Label

Image  Region  Pixel

semantic

granularity

e.g., CLIP [1]  e.g., GLIP [2]  e.g., MaskCLIP [3]
e.g., ViT [4]  e.g., DETR [5]  e.g., Mask2Former [6]

Bridge Vision with Language

CVPR2022 Tutorial: Recent Advanced in Vision-and-Language Pre-training (vlp-tutorial.github.io)

Replace labels with concept names, and use text encoder to encode all concepts as they are language tokens

Bridge Vision with Language for Segmentation

- **Segmentation tasks:**
  - Generic segmentation (semantic/instance/panoptic segmentation)
  - Referring segmentation (segment image with specific text phrase)

- **Methodologies:**
  - Initialize from CLIP v.s. train from scratch
  - Weakly supervised training v.s. supervised training
  - Two-stage v.s. end-to-end training
**Bridge Vision with Language for Segmentation**

- **GroupViT**: Learn to group semantic similar regions by learning from image-text pairs from scratch:
  - Bottom-up grouping using a novel grouping block
  - Top-down image-text supervision for visual-semantic alignment
Bridge Vision with Language for Segmentation

- **MaskCLIP**: Extract free dense label from CLIP
  - Change attention pooling to a new adaptation strategy
  - Pseudo-label masks using CLIP as the teacher model

Bridge Vision with Language for Segmentation

- **MaskCLIP**: Extract free dense label from CLIP
- Change attention pooling to a new adaptation strategy
- Pseudo-label masks using CLIP as the teacher model

Bridge Vision with Language for Segmentation

- **OpenSeg**: Weakly supervised learning by enforcing fine-grained alignment between textual features and mask-pooled features.
  - Learn from image-text pairs and local narrations.
  - A pretrained mask proposal network is used.
Bridge Vision with Language for Segmentation

- **OpenSeg**: Weakly supervised learning by enforcing fine-grained alignment between textual features and mask-pooled features.
  - Learn from image-text pairs and local narrations.
  - A pretrained mask proposal network is used.
Bridge Vision with Language for Segmentation

- **OpenSeg**: Weakly supervised learning by enforcing fine-grained alignment between textual features and mask-pooled features.
  - Learn from image-text pairs and local narrations.
  - A pretrained mask proposal network is used.

<table>
<thead>
<tr>
<th>Model</th>
<th>COCO Train label</th>
<th>mask cap.</th>
<th>mIoU A-847</th>
<th>PC-459</th>
<th>A-150</th>
<th>PC-59</th>
<th>COCO</th>
<th>mIoU A-847</th>
<th>PC-459</th>
<th>A-150</th>
<th>PC-59</th>
<th>COCO</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALIGN</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>4.8</td>
<td>3.6</td>
<td>9.7</td>
<td>18.5</td>
<td>15.6</td>
<td>17.8</td>
<td>21.8</td>
<td>25.7</td>
<td>34.2</td>
</tr>
<tr>
<td>ALIGN w/proposal</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>5.8</td>
<td>4.8</td>
<td>12.9</td>
<td>22.4</td>
<td>17.9</td>
<td>17.3</td>
<td>19.7</td>
<td>25.3</td>
<td>32.0</td>
</tr>
<tr>
<td>LSeg†</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>3.8</td>
<td>7.8</td>
<td>18.0</td>
<td>46.5</td>
<td>55.1</td>
<td>10.5</td>
<td>17.1</td>
<td>30.8</td>
<td>56.7</td>
</tr>
<tr>
<td>OpenSeg</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>6.3</td>
<td>9.0</td>
<td>21.1</td>
<td>42.1</td>
<td>36.1</td>
<td>21.8</td>
<td>32.1</td>
<td>41.0</td>
<td>57.2</td>
</tr>
<tr>
<td>OpenSeg w/L. Narr.</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>6.8</td>
<td>11.2</td>
<td>24.8</td>
<td>45.9</td>
<td>38.1</td>
<td>25.4</td>
<td>39.0</td>
<td>45.5</td>
<td>61.5</td>
</tr>
</tbody>
</table>
Bridge Vision with Language for Segmentation

- **MaskCLIP (UCSD):** Supervised training for panoptic segmentation with COCO using CLIP as the initialization
  - Two-stage training: 1) mask proposal network training; 2) CLIP model adaptation
Bridge Vision with Language for Segmentation

- **MaskCLIP (UCSD):** Supervised training for panoptic segmentation with COCO using CLIP as the initialization
  - Two-stage training: 1) mask proposal network training; 2) CLIP model adaptation

<table>
<thead>
<tr>
<th>Method</th>
<th>COCO Training Data</th>
<th>A-150↑</th>
<th>A-847↑</th>
<th>P-459↑</th>
<th>P-59↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALIGN (Jia et al., 2021)</td>
<td>None</td>
<td>10.7</td>
<td>4.1</td>
<td>3.7</td>
<td>15.7</td>
</tr>
<tr>
<td>ALIGN w/ proposals (Jia et al., 2021)</td>
<td>Masks</td>
<td>12.9</td>
<td>5.8</td>
<td>4.8</td>
<td>22.4</td>
</tr>
<tr>
<td>LSeg+ (Li et al., 2022a)</td>
<td>Masks + Labels</td>
<td>18.0</td>
<td>3.8</td>
<td>7.8</td>
<td>46.5</td>
</tr>
<tr>
<td>OpenSeg (Ghiasi et al., 2022)</td>
<td>Masks + Captions</td>
<td>21.1</td>
<td>6.3</td>
<td>9.0</td>
<td>42.1</td>
</tr>
<tr>
<td>SimSeg (Xu et al., 2022)</td>
<td>Masks + Labels</td>
<td>20.5</td>
<td>7.0</td>
<td>-</td>
<td>47.7</td>
</tr>
<tr>
<td><strong>CLIP Baseline</strong></td>
<td>Masks</td>
<td>13.8</td>
<td>5.2</td>
<td>5.2</td>
<td>25.3</td>
</tr>
<tr>
<td>MaskCLIP w/o RMA</td>
<td>Masks</td>
<td>14.9</td>
<td>5.6</td>
<td>5.3</td>
<td>26.1</td>
</tr>
<tr>
<td>MaskCLIP (MaskRCNN)</td>
<td>Masks + Labels</td>
<td>22.4</td>
<td>6.8</td>
<td>9.1</td>
<td>41.3</td>
</tr>
<tr>
<td><strong>MaskCLIP</strong></td>
<td>Masks + Labels</td>
<td>23.7</td>
<td>8.2</td>
<td>10.0</td>
<td>45.9</td>
</tr>
</tbody>
</table>

CLIP baseline works and mask proposals help slightly.
Bridge Vision with Language for Segmentation

- **MaskCLIP (UCSD):** Supervised training for panoptic segmentation with COCO using CLIP as the initialization
  - Two-stage training: 1) mask proposal network training; 2) CLIP model adaptation

<table>
<thead>
<tr>
<th>Method</th>
<th>COCO Training Data</th>
<th>A-150 ↑</th>
<th>A-847 ↑</th>
<th>P-459 ↑</th>
<th>P-59 ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALIGN (Jia et al., 2021)</td>
<td>None</td>
<td>10.7</td>
<td>4.1</td>
<td>3.7</td>
<td>15.7</td>
</tr>
<tr>
<td>ALIGN w/ proposals</td>
<td>Masks</td>
<td>12.9</td>
<td>5.8</td>
<td>4.8</td>
<td>22.4</td>
</tr>
<tr>
<td>LSeg+ (Li et al., 2022a)</td>
<td>Masks + Labels</td>
<td>18.0</td>
<td>3.8</td>
<td>7.8</td>
<td>46.5</td>
</tr>
<tr>
<td>OpenSeg (Ghiasi et al., 2022)</td>
<td>Masks + Captions</td>
<td>21.1</td>
<td>6.3</td>
<td>9.0</td>
<td>42.1</td>
</tr>
<tr>
<td>SimSeg (Xu et al., 2022)</td>
<td>Masks + Labels</td>
<td>20.5</td>
<td>7.0</td>
<td>-</td>
<td>47.7</td>
</tr>
<tr>
<td>CLIP Baseline</td>
<td>Masks</td>
<td>13.8</td>
<td>5.2</td>
<td>5.2</td>
<td>25.3</td>
</tr>
<tr>
<td>MaskCLIP w/o RMA</td>
<td>Masks</td>
<td>14.9</td>
<td>5.6</td>
<td>5.3</td>
<td>26.1</td>
</tr>
<tr>
<td>MaskCLIP (MaskRCNN)</td>
<td>Masks + Labels</td>
<td>22.4</td>
<td>6.8</td>
<td>9.1</td>
<td>41.3</td>
</tr>
<tr>
<td>MaskCLIP</td>
<td>Masks + Labels</td>
<td>23.7</td>
<td>8.2</td>
<td>10.0</td>
<td>45.9</td>
</tr>
</tbody>
</table>

Label information significantly boost open-vocabulary performance.
Put all together

Data

mIoU

Image-text pairs

Pseudo mask + Self-training

CLIP or others

Image-text pairs

CLIP or others

Labels

CLIP or others

Masks

Labels

CLIP or others

Masks

Image-text pairs

GroupViT (Pascal 52.3)

MaskCLIP (Pascal 86.1)

OpenSeg (Pascal 72.2)

Zegformer (Pascal 80.7)

X-Decoder (Pascal 97.9)
Put all together

CLIP as the foundation helps a lot for open-vocabulary seg.
Put all together

GroupViT (Pascal 52.3)

MaskCLIP (Pascal 86.1)

OpenSeg (Pascal 72.2)

Zegformer (Pascal 80.7)

X-Decoder (Pascal 97.9)

Combine weak annotations with golden ones for better performance
Bridge Vision with Language for Core Vision

**Language**
- e.g., CLIP [1]
- e.g., GLIP [2]
- e.g., MaskCLIP [3]
- e.g., ViT [4]
- e.g., DETR [5]
- e.g., Mask2Former [6]

**Granularity**
- Image
- Region
- Pixel

**Semantic**
- Image Classification
- Object Detection
- Segmentation
These models are unleashed to recognize open-world concepts but still mostly task-specific.
Connect tasks horizontally across different granularities
II. Unify Different Granularities

- Bridge vision with language
- Unify different granularities
- Take various prompts
Unify Different Granularities

- **Image Annotation** (ImageNet, LAION)
  - Label
    - e.g., ViT [4]
    - e.g., DETR [5]
    - e.g., Mask2Former [6]
  - Language
  - Image
  - Region
  - Pixel
  - **granularity**

- **Box Annotation** (COCO, O365)
  - Mask Annotation (COCO, LVIS)
  - e.g., CLIP [1]
  - e.g., GLIP [2]
  - e.g., MaskCLIP [3]

- **Semantic**
  - Image Classification
  - Object Detection
  - Segmentation
Unify Different Granularities

From coarse-grain to fine-grain: rich semantics
From fine-grain to coarse-train: better grounding

Granularity

Image Annotation
(ImageNet, LAION)

Language

Label

Image
Region
Pixel

ViT [4]
DETR [5]
Mask2Former [6]

Mask Annotation
(COCO, LVSI)

Image Classification
Object Detection
Segmentation

semantic
Unify Different Granularities

• Tasks we are considering:
  • **Image-level**: image recognition, image-text retrieval, image captioning, visual question answering, etc.
  • **Region-level**: object detection, dense caption, phrase grounding, etc.
  • **Pixel-level**: generic segmentation, referring segmentation, etc.

• Two types of unifications:
  • **Output unification**: convert all outputs into sequence.
  • **Functionality unification**: share the commons maximally but with respect to the differences.
Unify Different Granularities

Convert all outputs into sequence and `decode` to corresponding outputs

Predict shared output types and `combine` one or more to produce the final outputs
Outputs Unification

- Convert both inputs and outputs into sequences:
  - **Inputs**: Text as it is or add some prefixes; Image into a sequence of tokens (not necessarily)
  - **Outputs**: Boxes: a sequence of coordinates (top left + bottom right); Masks: a sequence of polygon coordinates encompassing mask; Key points: a sequence of coordinates.
Outputs Unification

- **UniTab and Pix2Seqv2**: Unify text and box outputs with no specific modules

- **Common vocabulary**: text and coordinates are both tokenized and put into the same vocabulary
- **Task prefix**: requires a task prefix to determine which task the model is coping with

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B@4</td>
<td>M</td>
</tr>
<tr>
<td>NBT [49]</td>
<td>27.1</td>
<td>21.7</td>
</tr>
<tr>
<td>GVD [86]</td>
<td>27.3</td>
<td>22.5</td>
</tr>
<tr>
<td>Cyclical [50]</td>
<td>26.8</td>
<td>22.4</td>
</tr>
<tr>
<td>POS-SCAN [88]</td>
<td>30.1</td>
<td>22.6</td>
</tr>
<tr>
<td>Chen et al. [9]</td>
<td>27.2</td>
<td>22.5</td>
</tr>
<tr>
<td>UniTAB</td>
<td><strong>30.1</strong></td>
<td><strong>23.7</strong></td>
</tr>
</tbody>
</table>
Outputs Unification

• **Unified-IO**: unify a wide range of understanding tasks including segmentation
  • Output Quantization: VQVAE for different types of tasks, such as mask, depth, image.
    (shared by UViM and OFA to some extent)
  • Two-stage pretraining: 1) pretraining VQVAE; 2) jointly pretraining on multiple tasks in a seq-to-seq manner
Outputs Unification

• Other works like VisionLLM use LLM as the output interface
• It unifies a wide range of vision tasks so that an encoder-decoder can be trained end-to-end
• It also:
  • needs task-specific decoder to decode the sequence to final outputs:
    • E.g., extract coordinates and translate into a box, convert polygon/color map into mask
  • might be hard to interpret the interactions across different tasks of different granularities
  • may not be able to build a strong cross-task synergy as we expect
Functionality Unification

• Vision tasks are not fully isolated:
  • **Box outputs**: shared by generic object detection, phrase grounding, regional captioning
  • **Mask outputs**: shared by instance/semantic/panoptic segmentation, referring segmentation, exemplar-based segmentation, etc.
  • **Semantic outputs**: shared by image classification, image captioning, regional captioning, detection, segmentation, visual question answering, image-text retrieval, etc.
Functionality Unification

- **UniPerceiver-v2**: a unified decoder is exploited for many vision understanding tasks.

**Our Generalist Model – Uni-Perceiver v2**

**General Task Adaptation**

Image Classification
Object Detection
Instance Segmentation
Image-Text Retrieval
Image Captioning

\[ \#P_{\text{total}} = \#P_{E^I} + \#P_{E^T} + \#P_{D_{\text{general}}} \]
Functionality Unification

• **X-Decoder**: Generalized Decoding for Pixels, Images, and Language
Functionality Unification

• **X-Decoder**: Generalized Decoding for Pixels, Images, and Language

Query: Zebra, antelope, giraffe, ostrich, sky, water, grass, sand, tree

Query: Owl on the left

Query: The tangerine on the plate.

Cap: river in the mountains near the town
Unify Different Granularities

![Diagram showing the unification of different granularities in computer vision tasks like image classification, object detection, and segmentation.](image)
Computer Vision in the Wild

Image Classification in the Wild

Example of knowledge sources

- **Concept name**: risotto
- **Def. wild**: An Italian savoury dish made with rice and other ingredients
- **Def. con**: rice cooked with broth and sprinkled with grated cheese
- **Path. con**: risotto, dish, nutriment, food, substance, matter, physical, entity, entity
- **GPT3**: “A rice dish made with Arborio rice and typically served with meat or fish.”
- **GPT3**: “A rice dish made by stirring rice into a simmering broth”

Task preview

- **Flowers102**
- **DTD**
- **Food101**
- **Country211**
- **RESISC45**
- **FGVCAircraft**
- **Caltech101**
- **FER2013KiwiDistance**
- **EuroSat**
- **VOC2007**
- **StanfordCars**
- **MNIST**
- **GTSRB**
- **OxfordPets CIFAR100 CIFAR10**

Object Detection in the Wild

Example of knowledge sources

- **Concept name**: starfish
- **Def. wild**: Any of various asteroids or other echinoderms (not in fact fish) with usually five arms, many of which eat bivalves or corals by everting their stomach.
- **Def. con**: echinoderms characterized by five arms extending from a central disk
- **Path. con**: starfish, echinoderm, invertebrate, animal, organism, living_thing, whole, object, physical, entity, entity
- **GPT3**: A marine animal of class Asteroidea, typically having a central disk and five arms.

Task preview

- **ChessPieces**
- **BrackishUnderwater**
- **Plantdoc**
- **AmericanSignLanguageLetters**
- **UnoCards**
- **SelfDrivingCar**
- **AerialMaritimeDrone**
- **EgoHands(specific)**
- **AerialMaritimeDrone(tiled) eaglesAirplaneGen2**

Segmentation in the Wild

Exemplar images in SginW Benchmark

Task preview

- **Nutterfly**
- **Metal**
- **Fruits**
- **Squirrel**
- **Garbage**
- **Airplane**
- **Puppies**
- **ElectricShaver**
- **HouseHold_Items**
- **Brain**
- **Hand**
- **Hand**
- **Elephants**
- **Strawberry**
- **Trash**
- **Watermelon**

2nd Workshop on Computer Vision in the Wild, East Ballroom B, June 19th full day
Promptable Interface

- Bridge vision with language
- Unify different granularities
- Take various prompts
How to Enable Vision Model to “Chat”

Decoder LLMs (e.g., GPT)  Human-AI Interaction  Conversational AI (e.g., ChatGPT)

Generalist Vision Models  Human-AI Interaction  ?
How to Enable Vision Model to “Chat”

• We need to build a promptable interface with two important properties:

  • **Promptable for in-context learning**: Instead of finetuning the model parameters, simply providing some contexts will make the model predict

  • **Interactive for user-friendly interface**: multi-round of interaction between human and AI is important to finish complicated tasks.
In-Context Learning for Vision

- **Visual Prompting via Image Inpainting:**
  - Concatenate in-context sample with query into a single image
  - Ask model to inpaint the missed part of the image grid

![Diagram](image.png)
In-Context Learning for Vision

- **SegGPT**: Segment Everything as in-context learning
Interactive Interface for Vision

- **SAM**: Segment Anything
  - Promptable segmentation
Interactive Interface for Vision

- **SAM**: Segment Anything
Interactive Interface for Vision

- **SEEM**: Segment Everything Everywhere all at Once
Interactive Interface for Vision

- **SEEM**: Segment Everything Everywhere all at Once
A quick recap

**Intuition:** Human use language as the common space to share information

**Benefit:** Zero-shot transfer to novel vocabularies

**Intuition:** Human vision is for multi-task, multi-granularity

**Benefit:** Build synergy across task granularities

**Intuition:** Human uses both language, spatial prompts and beyond for vision.

**Benefit:** Reduce the ambiguity of expressing human intents
Sparks of Artificial General Intelligence (AGI)

Sparks of Artificial General Intelligence:
Early experiments with GPT-4

Sébastien Bubeck    Varun Chandrasekaran    Ronen Eldan    Johannes Gehrke
Eric Horvitz       Ece Kamar         Peter Lee      Yin Tat Lee     Yuanzhi Li     Scott Lundberg
Harsha Nori         Hamid Palangi     Marco Tulio Ribeiro   Yi Zhang

Microsoft Research
Artificial General Intelligence (AGI)

- Natural Language Processing
- Computer Vision
- Auditory sensation - Speech
- Motor control - Action
- ...
Drawing dots for generalist vision to

Enable an intimate cooperation with LLMs for physic world task
  Give GPT, ChatGPT, BioGPT the eyes!

Empower more grounded image/video manipulation
  Let DALLE-1/2 not only imaging things but grounding to the realistic!

Achieve multi-sensory general intelligent agent!
  A real agent that can see, talk, act!

We are fortunate to have a lot of imagination space!!!
Thanks for your attention!