SAM2 AND SAA

October 2024

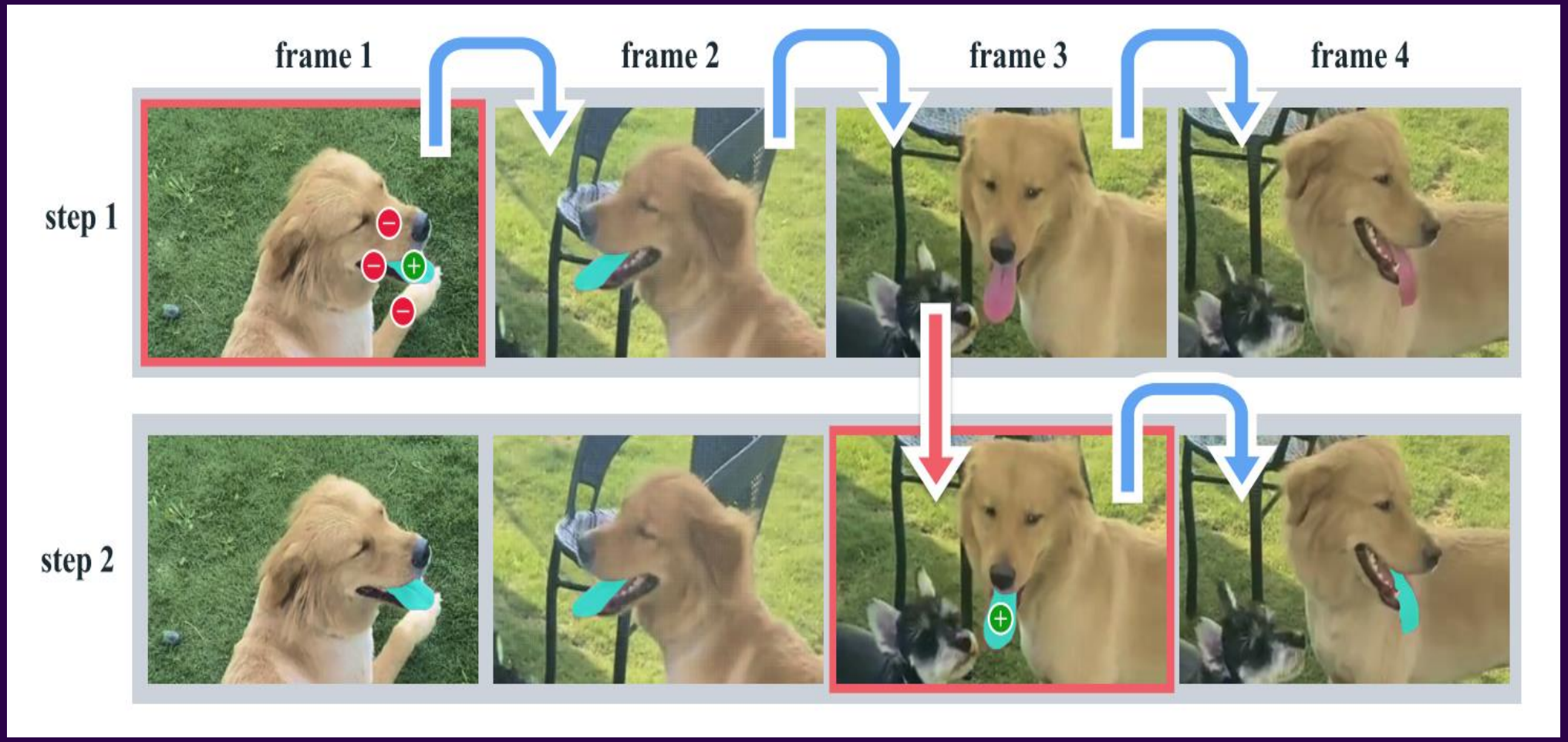
SAM2

- SAM2 works on image AND video
- Focus on Promptable Visual Segmentation
  - Points (positive or negative)
  - Bounding boxes
  - Mask

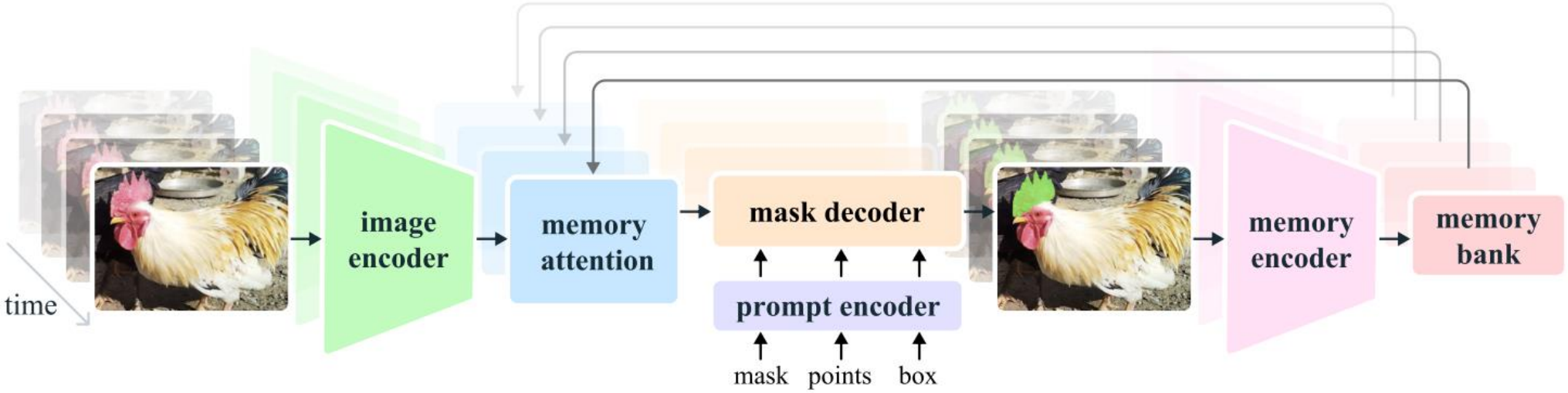
CHALLENGES

- Video frames have lower quality (motion blur, light) compared to images.
- processing of multiple frames for real-time applications.
- Objects in videos can change appearance due to motion, occlusion, lighting variations, and deformation.

OVERVIEW



MAIN MODEL(SAM2)



SAM 1

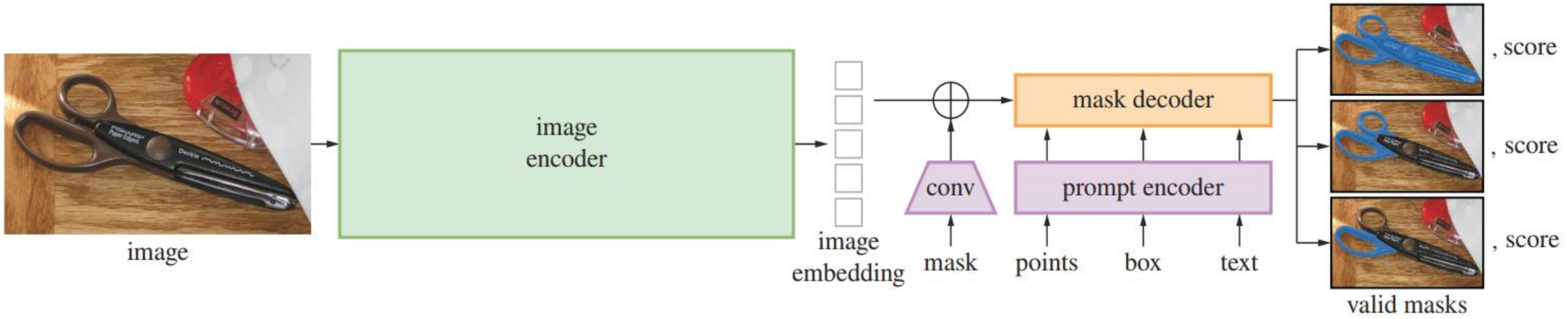
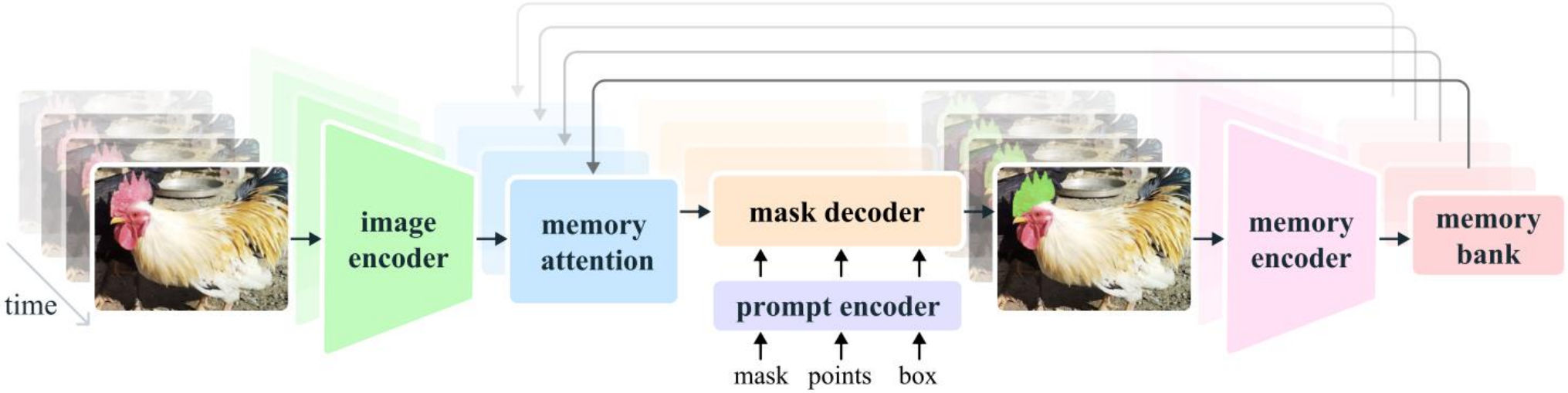


image
image encoder
image embedding
conv
mask
mask decoder
prompt encoder
points
box
text
, score
, score
, score
valid masks

IMAGE ENCODER

- Use Hiera image encoder for feature extraction
    - Hiera use hierarchical structure
- allowing to use multiscale features
- Do not use low level feature in memory attention(stage 1 and 2)

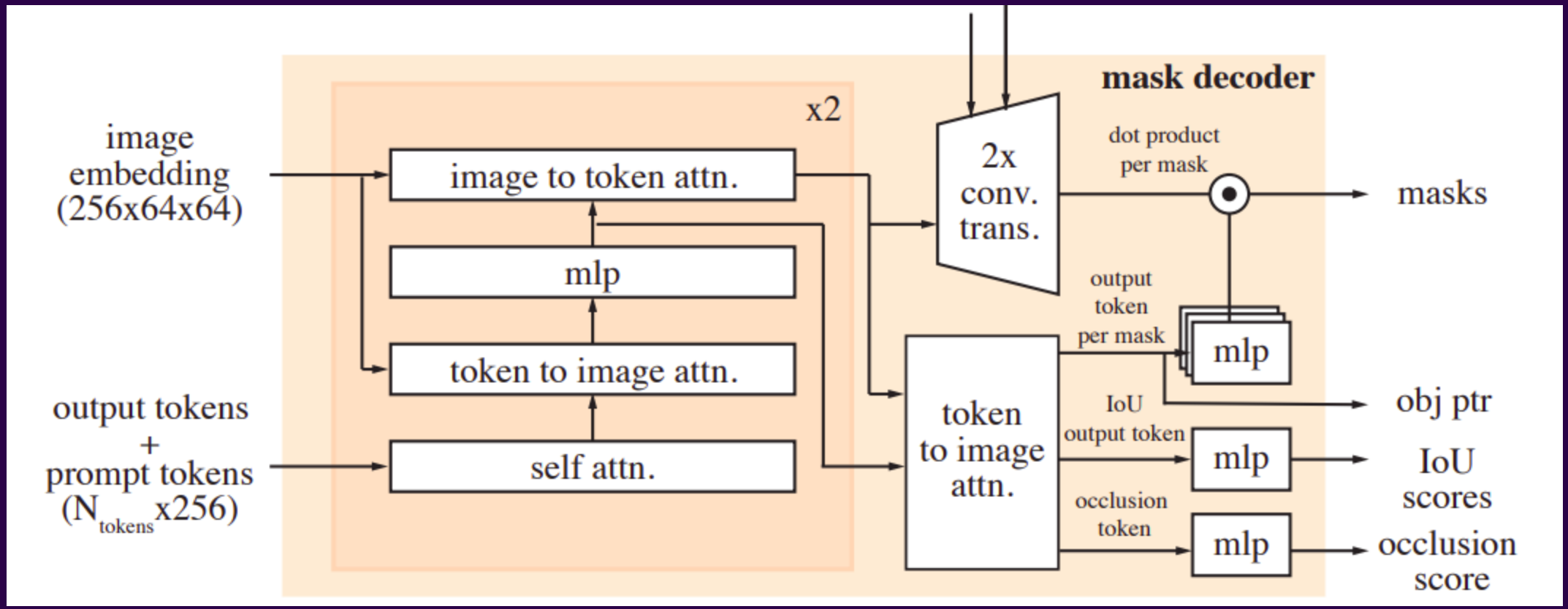
MAIN MODEL(SAM2)

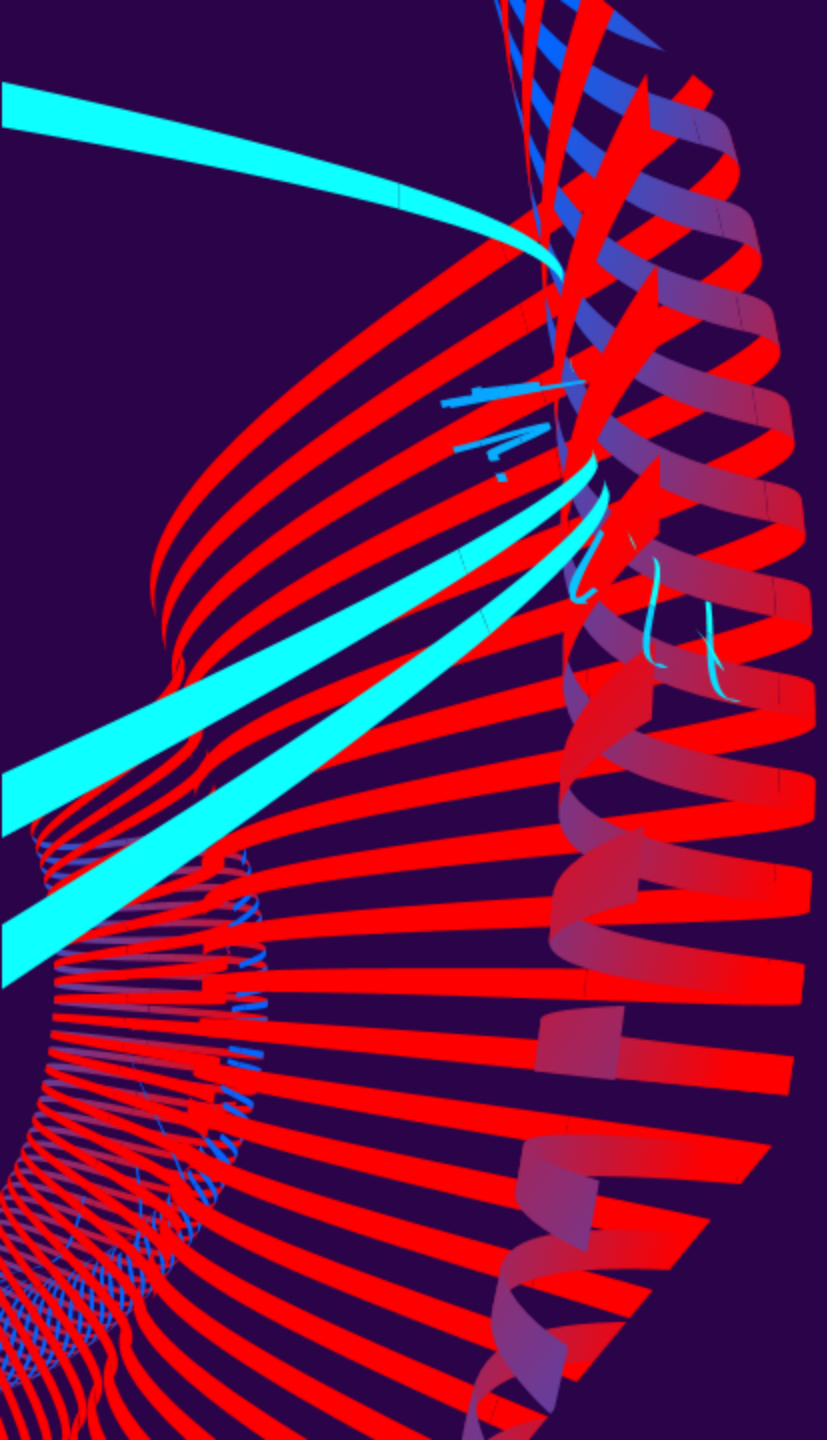


MEMORY ATTENTION

- Input:
  - Current frame
  - memories of past frames
- model use of information from previous frames
- Self-attention apply on current frame features, help model focus on different parts of the current frame.
- Cross-attention is then apply between the current frame and the stored memory
- Use high level feature

OVERVIEW OF MASK DECODER





MASK DECODER

- Unlike SAM, in SAM2 there is possibility for no valid object to exist on some frames

- Unlike SAM, it use skip connections to incorporate high-resolution information

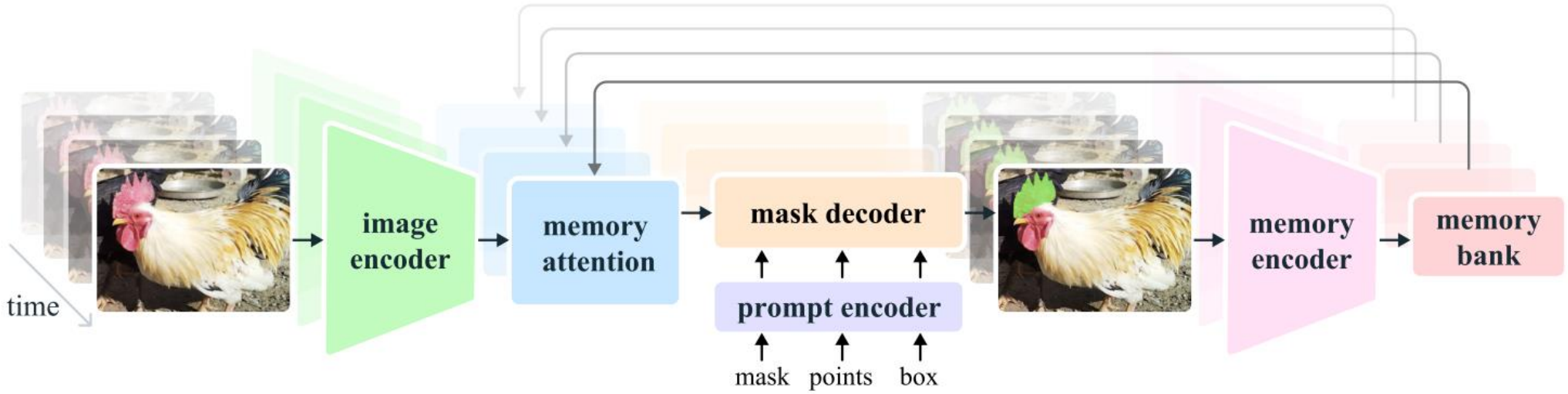
MASK DECODER

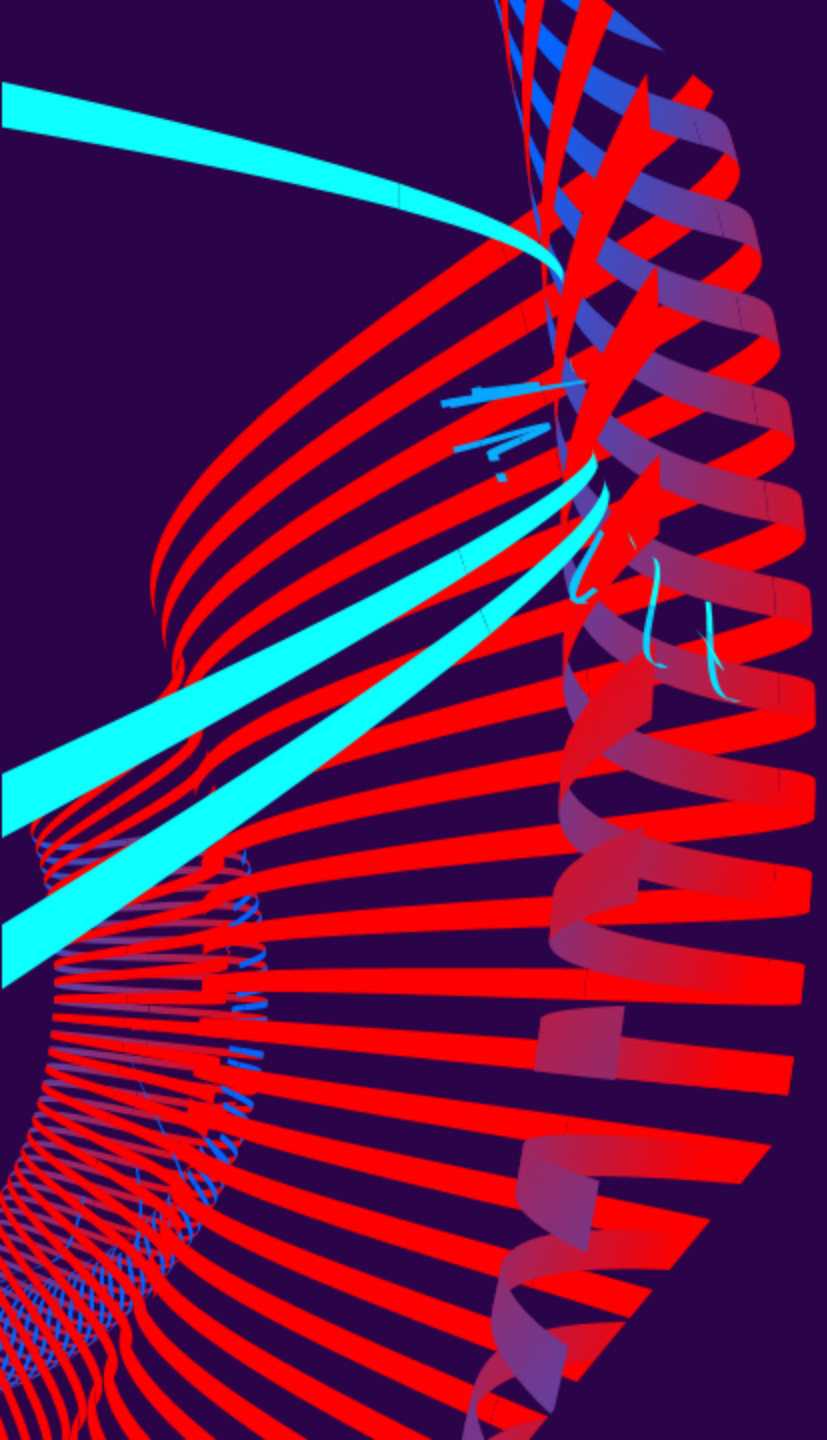
- Input
  - Token from the prompt encoder
  - Image embeddings form image encoder
- Transformer Blocks
  - Self-attention: help to process in frame
  - Cross-attention: allows model to relate the current frame to past frames
- token to image Attention:
  - allows the image features interact with the tokens

MASK DECODER

- Output:
  - Occlusion Prediction: predicting object of interest is visible in the current frame or not
  - IoU scores: evaluate the quality of each predicted mask
  - Mask

MAIN MODEL(SAM2)

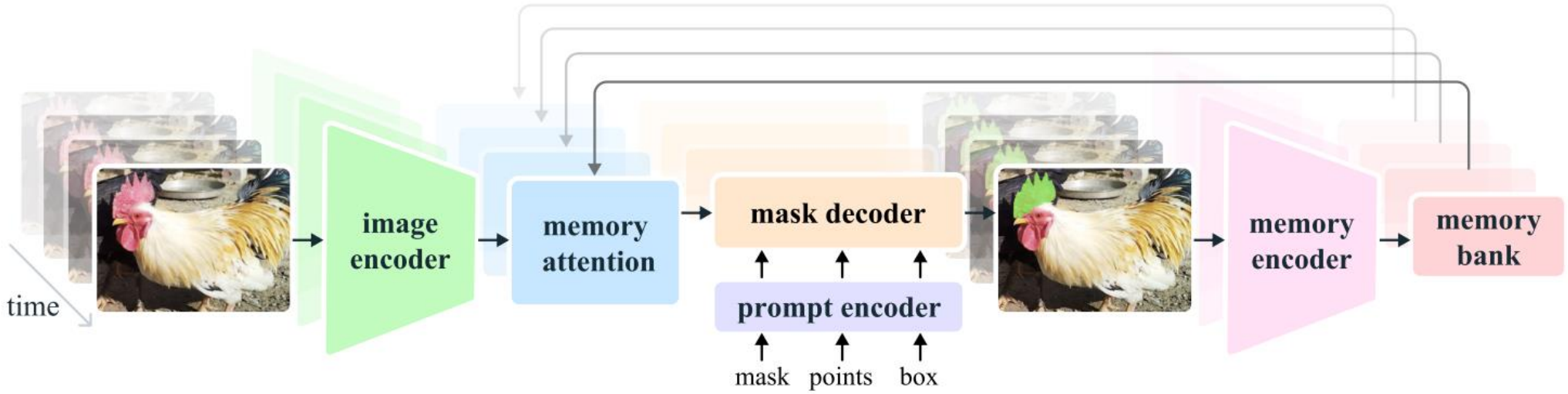


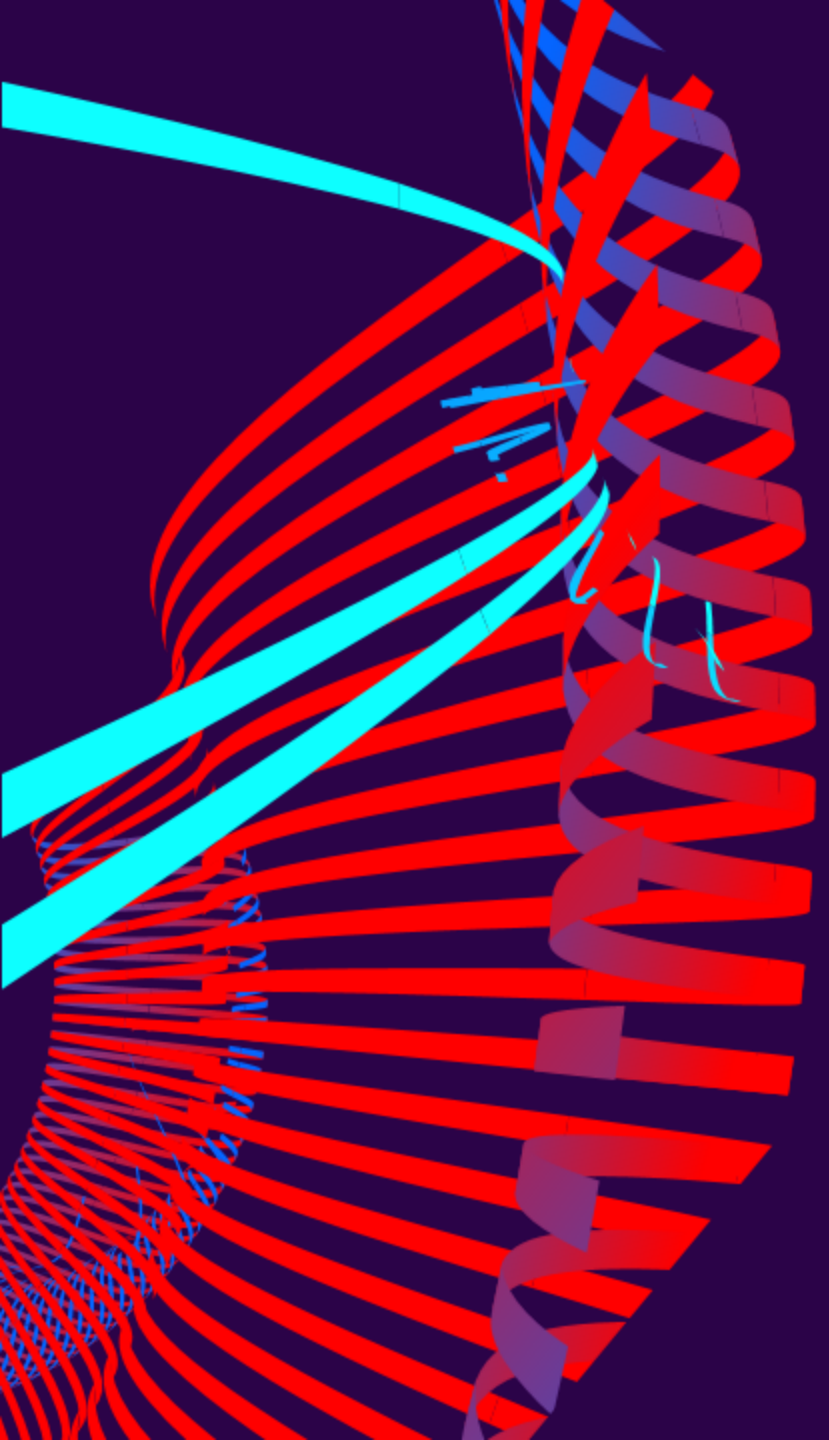


MEMORY ENCODER

- downsampling the output of the mask
- Input:
  - Predicted mask: segmented object
  - Frame embeddings: Features from the image encoder
- fuses the predicted segmentation mask and the frame embeddings into memory feature map
- Store memory feature maps and object pointers into memory bank

MAIN MODEL(SAM2)





MEMORY BANK

- Save information about past predictions for the target object
- Limit in number of frame and prompt
- Based on FIFO
- Memories stored as spatial feature maps
  - Segmented object
  - visual features
  - location

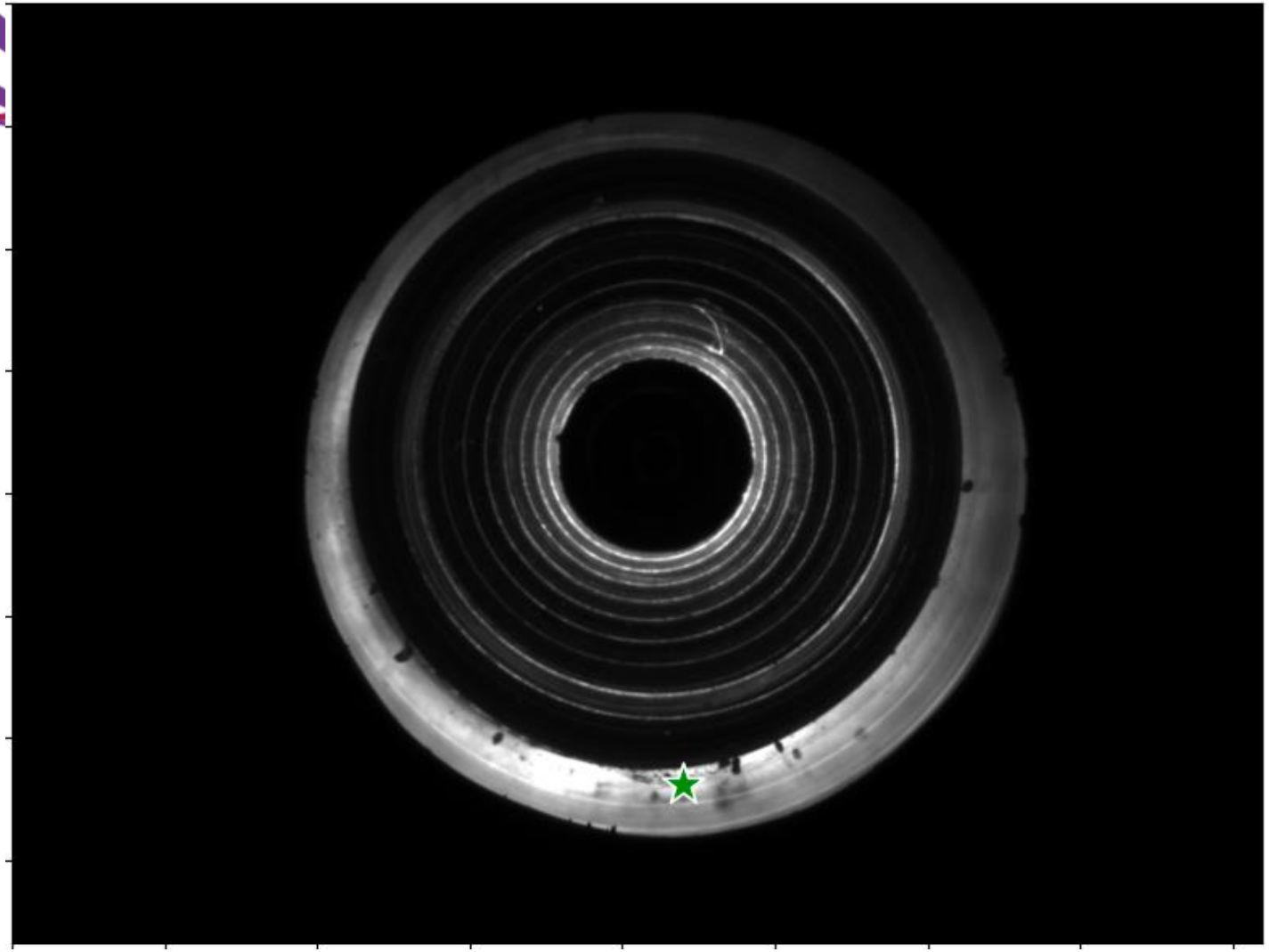
KEY IMPROVMENT

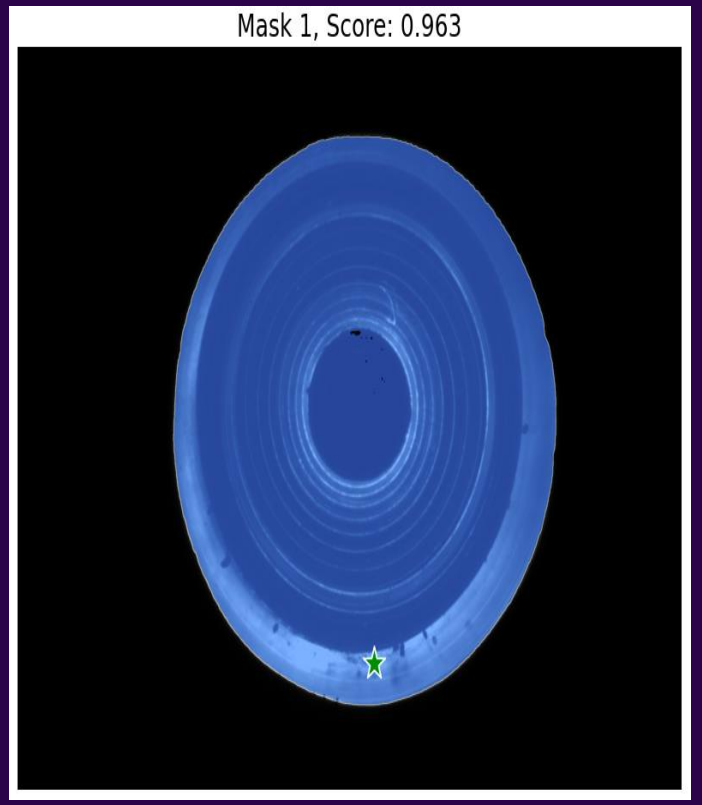
- embedding used by decoder is not directly from an image encoder

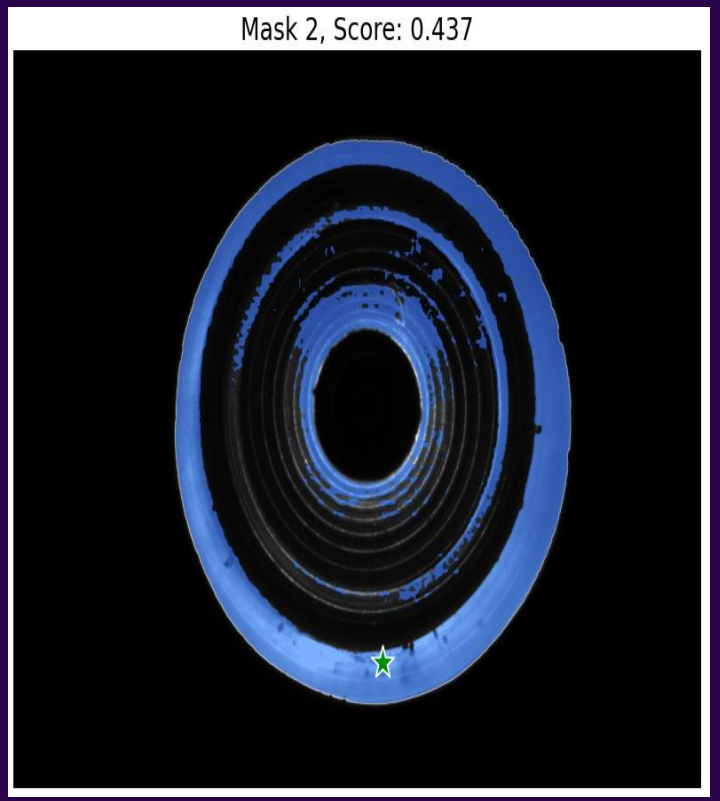
- Decoder use memories of past predictions and prompted frames.

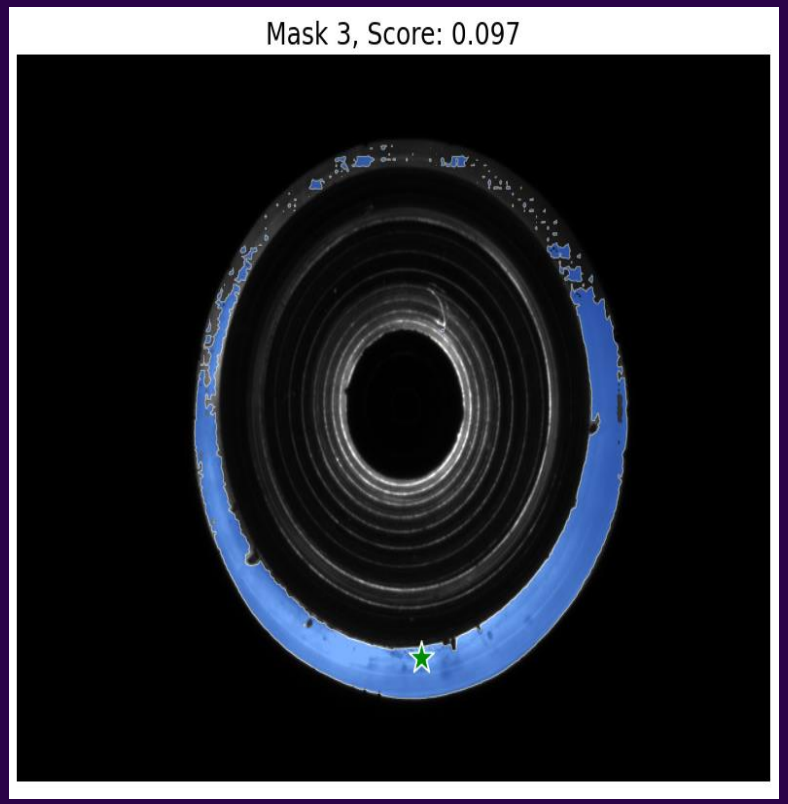
• Saved memory encoder data into memory bank for each frame

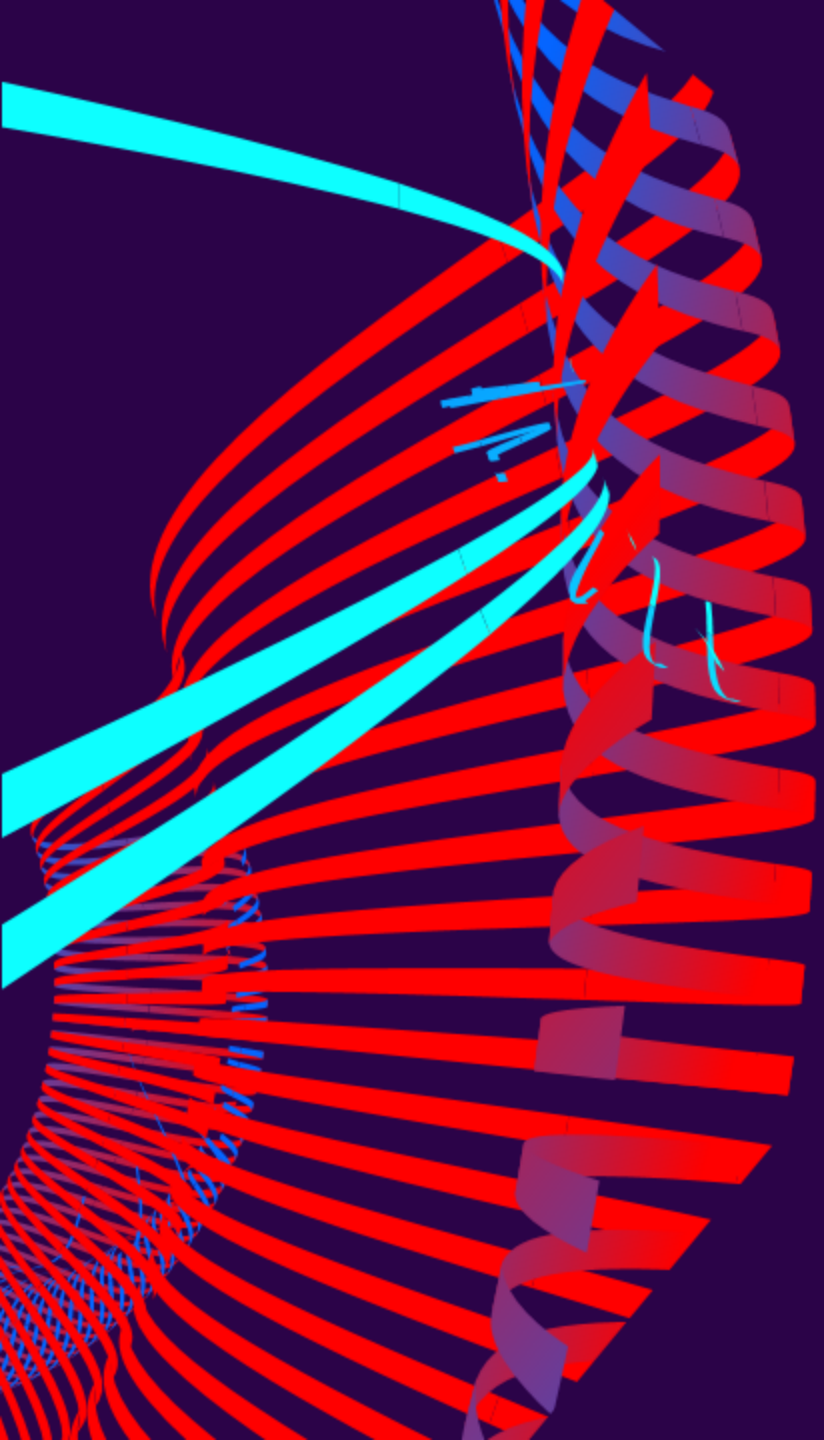
EXAMPLE




Mask 1, Score: 0.963


Mask 2, Score: 0.437


Mask 3, Score: 0.097



SAA+

Prompt :

- textual_prompts
  - ['dark spots.', 'outer ring defects']
  - ['black marks.', 'outer ring']

- property_text_prompts
  - 'the image of pipe have 1 dissimilar pipe, with a maximum of 10 anomaly. The anomaly would not exceed 1. object area.'