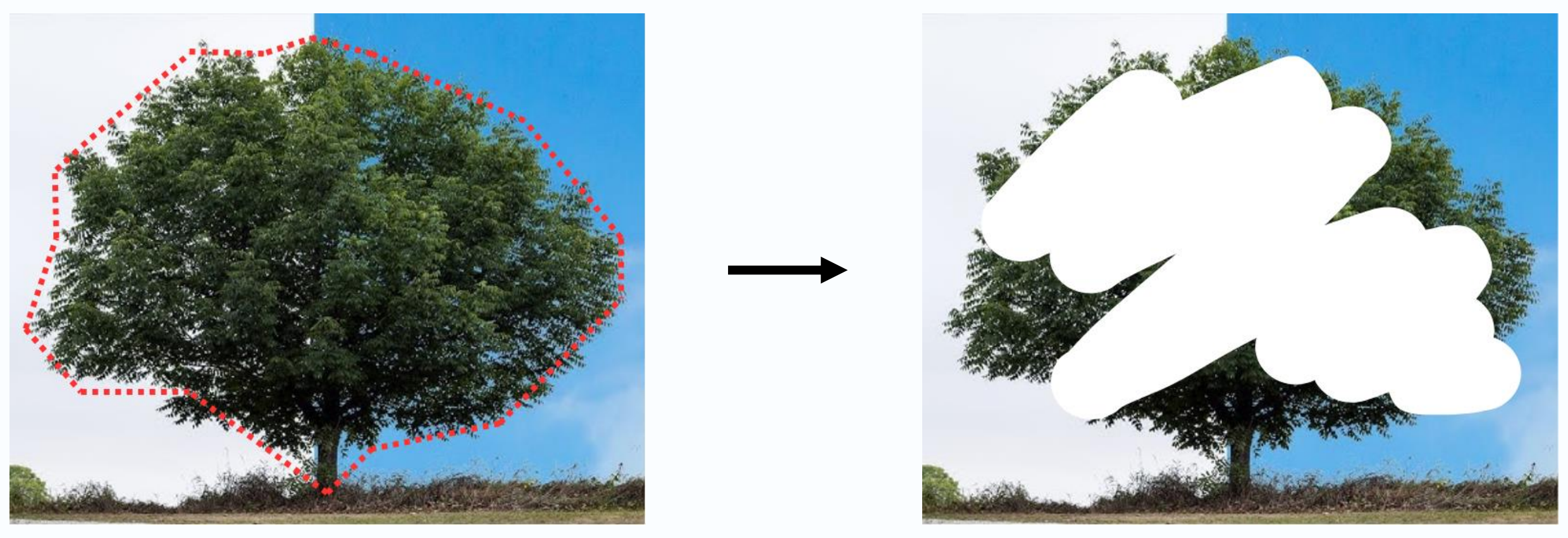
InpaLLa : All Inpainting Start From Your Thoughts By mLLm Architecture

생성 모델

Team InpaLLA | 김창현, 김용진, 신기섭, 강현구, 김근하

Introduction

- Most people have difficulty creating the designs they desire themselves.
- Photo editing software is typically designed for professionals, requiring users to manually draw outlines when editing objects.
- When multiple objects are involved, the complexity and difficulty of the task increase significantly.

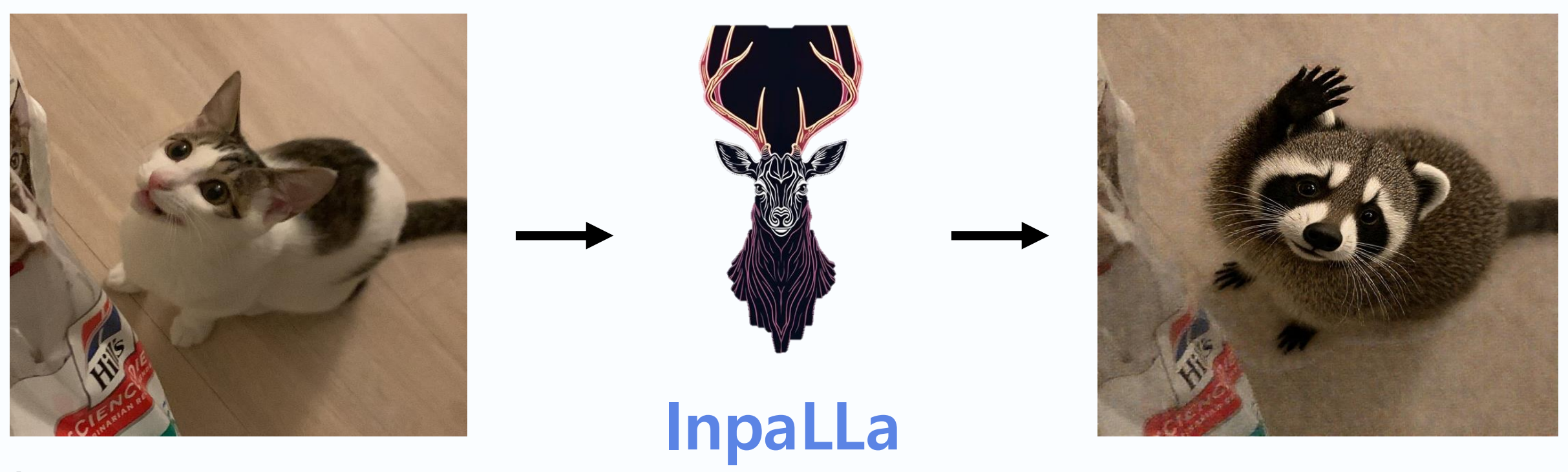


Need to draw outline

Ask : “Remove this tree”

Figure 1. Replacing traditional manual tasks with task performed by asking.

- Therefore, we introduce **InpaLLa** : **Inpa**inting with mLLm **A**rchitecture which enables users to easily replace specific object in images
- Our model allows seamless replacement of objects in an image based on only user provided text requirements and image.
- **InpaLLa** helps users by providing an AI agent that assists in generating desired designs efficiently and intuitively.



Pipeline

InpaLLA Pipeline

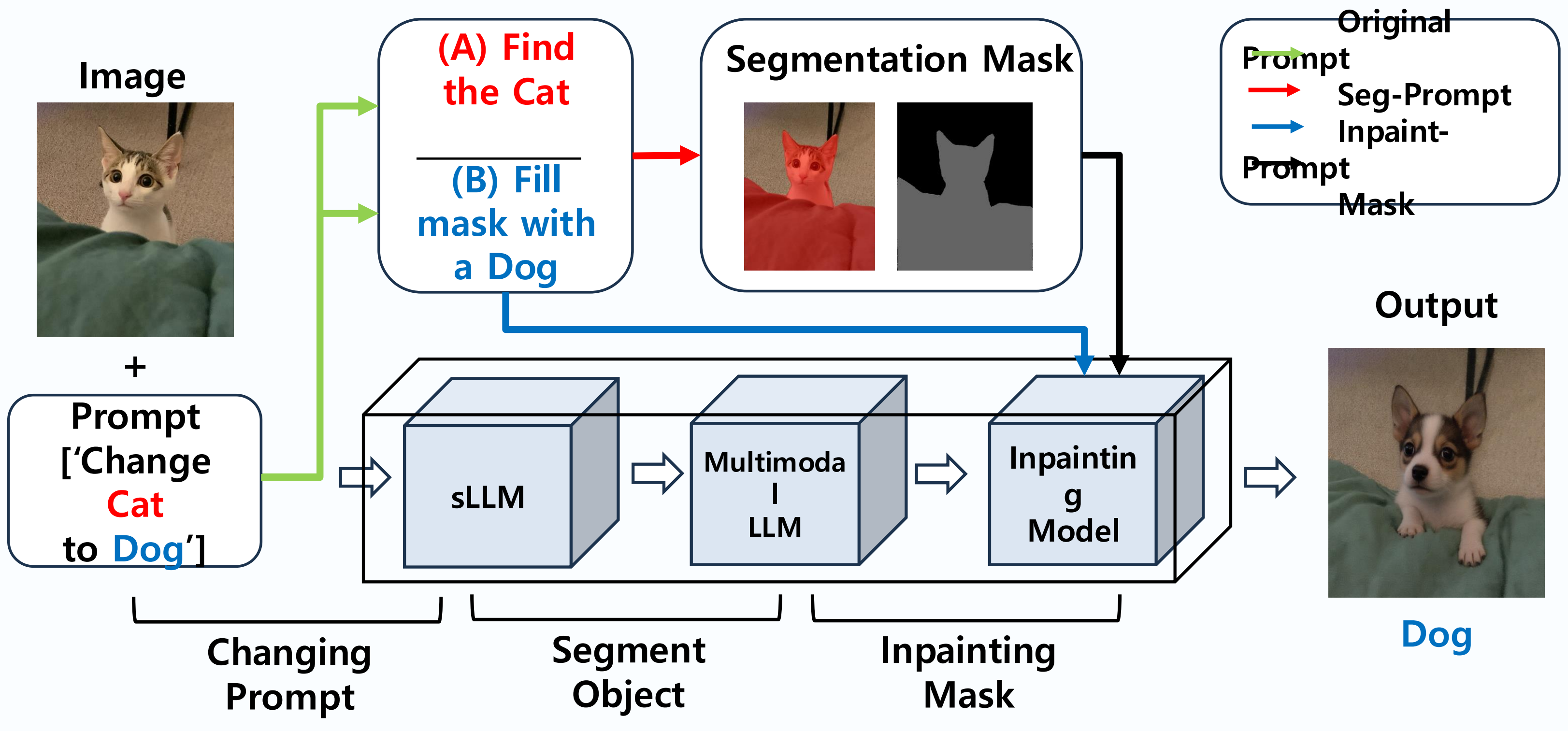


Figure 2. Pipeline of InpaLLa

Process of InpaLLa

1. **Generating User Prompts for Multimodal-LLM and Inpainting**:

$$p_{seg}, p_{inp} = \mathcal{F}_{sLLM}(x_{txt}, x_{img})$$

sLLM, $\mathcal{F}_{sLLM}(\cdot)$ generates text prompts p_{inp} for inpainting model $\mathcal{F}_{inpaint}(\cdot)$ and p_{seg} for Multimodal-LLM $\mathcal{F}_{MLLM}(\cdot)$ with image x_{img} and prompt x_{txt} .

2. **Creating Segmentation Mask**: Segmentation mask m is generated by the Multimodal-LLM:

$$m = \mathcal{F}_{MLLM}(p_{seg}, x_{img})$$

3. **Applying Inpainting**: The inpainting model $\mathcal{F}_{inpaint}$ generates output image $\hat{I}$ by using the input image x_{img}, mask m, and p_{inp}:

$$\hat{I} = \mathcal{F}_{inpaint}(x_{img}, m, p_{inp})$$

~~**We use LISA for MLLM model $\mathcal{F}_{MLLM}$, FLUX for inpainting model $\mathcal{F}_{inpaint}$.**~~

Process of LISA : Reasoning Segmentation via LLM

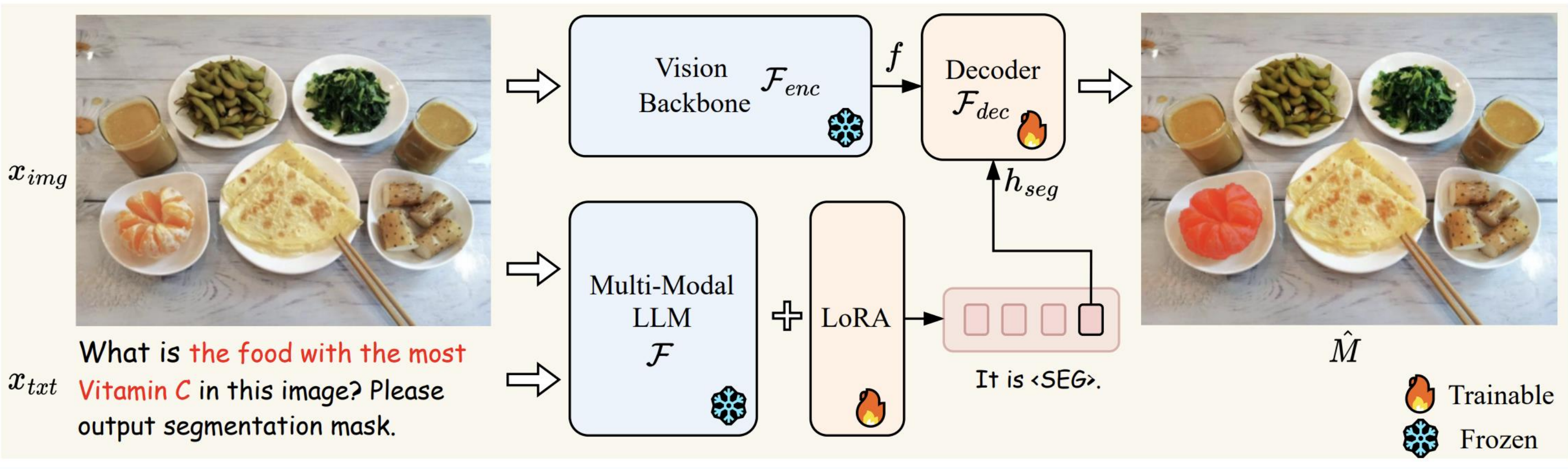

x_{img}
x_{txt}
What is the food with the most Vitamin C in this image? Please output segmentation mask.
Vision Backbone $\mathcal{F}_{enc}$
f
Decoder $\mathcal{F}_{dec}$
h_{seg}
Multi-Modal LLM $\mathcal{F}$
LoRA
It is <SEG>.
$\hat{M}$
Trainable
Frozen

Figure 3. Pipeline of LISA

- **Embedding-as-Mask Paradigm :** A framework where a segmentation mask is generated by decoding the embedding of a specific token $<seg>$. This allows a multimodal LLM to integrate segmentation capabilities seamlessly.
- **Mask generation:** Represent Embedding-as-Mask Paradigm

$$\widehat{M} = \mathcal{F}_{dec}(\gamma, (\tilde{h}_{seg})\, \mathcal{F}_{enc}(x_{img}))$$

$\widehat{M}$ is mask that generated by model, $\tilde{h}_{seg}$ is the hidden embedding of $<seg>$ token and $\mathcal{F}_{enc}$ is encoder and $\mathcal{F}_{dec}$ is decoder.

Process of FLUX

- **Rectified flow : More efficient approach to make image**

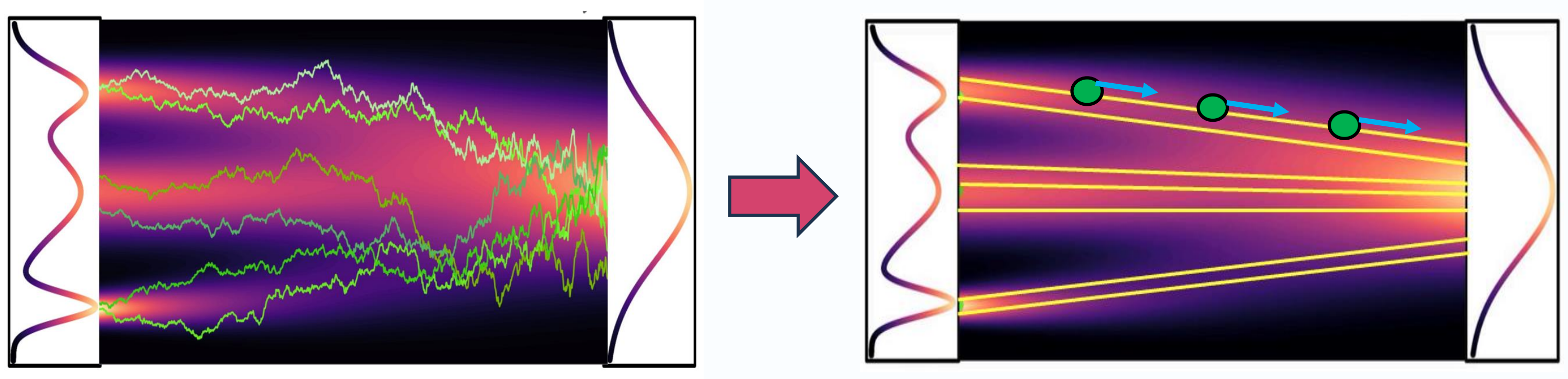


Figure 4. Rectified flow

Rectified flow is based on ODE and is trained to ensure that data points follow the straight path as much as possible.

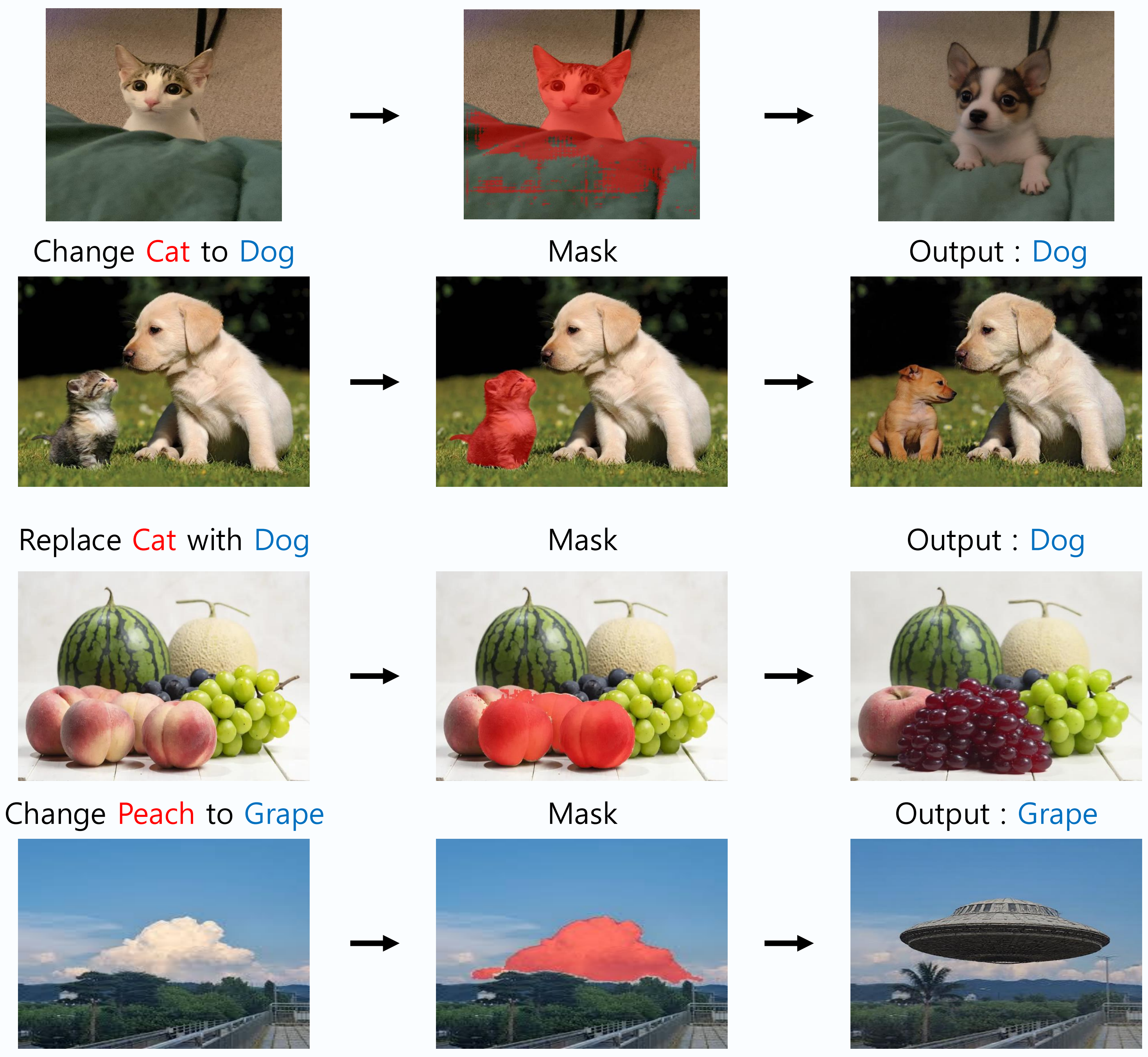
Training: To learn the drift force $v(Z_t, t)$ of the ODE, a nonlinear least squares optimization problem is solved.

$$\min_{v} \int_0^1 E[\| (X_1 - X_0) - v(X_t, t) \|^2]dt$$

Here, $X_t = tX_1 + (1-t)X_0$ represents the linear interpolation from X_0 to X_1.

Result

Result : Example of InpaLLa



Change Cat to Dog | Mask | Output : Dog

Replace Cat with Dog | Mask | Output : Dog

Change Peach to Grape | Mask | Output : Grape

Switch Cloud to UFO | Mask | Output : UFO