NEURAL INFORMATION PROCESSING SYSTEMS

CVIT

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Adobe

# PartConcepts

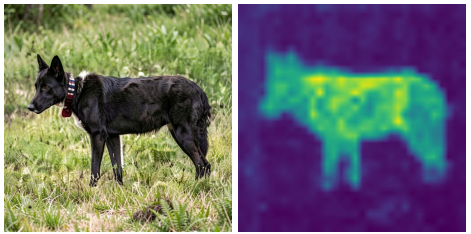
A Unified Mechanism for Fine-Grained Part **Localization** and **Generation**

🌐 https://visionpartconcepts.github.io

2026

# Motivation

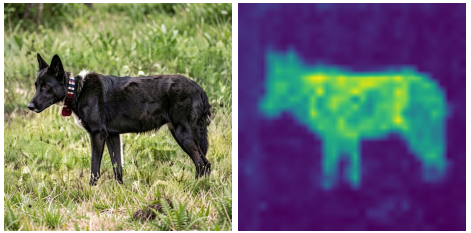
Diffusion latents have strong semantics at the object-level, e.g., **dog**.



*Cross-attention map with the dog token.*

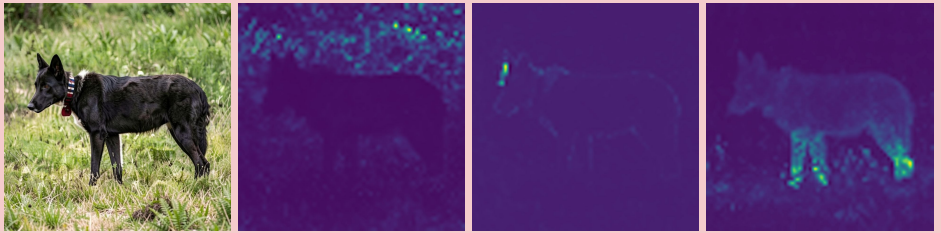
# Motivation

Diffusion latents have strong semantics at the object-level, e.g., **dog**.



*Cross-attention map with the* `dog` *token.*

But this fails to extend to the part-level granularity, e.g., **left front leg**.



*Cross-attention map with the* `left`, `front` *and* `leg` *tokens.*

# Motivation

The poor part-level semantics correlate with several capability gaps in existing T2I diffusion models.

Poor part-level prompt following



Nano Banana 2 FLUX.2

"A man in an art studio with right upper arm with green stains and left lower arm with purple stains"

# Motivation

The poor part-level semantics correlate with several capability gaps in existing T2I diffusion models.

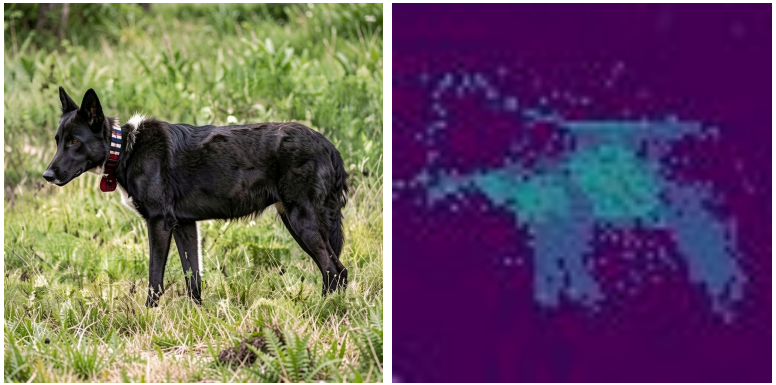
**Poor part-level prompt following**



Nano Banana 2 FLUX.2

"A man in an art studio with right upper arm with green stains and left lower arm with purple stains"

**Poor concept segmentation**



Input image ConceptAttention (ICML-2025)

(segmenting `left front leg`)

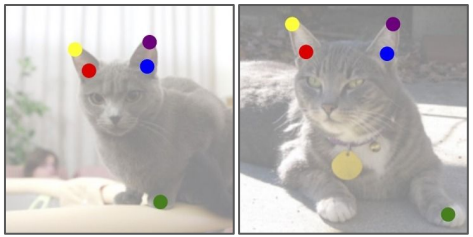
# Key motivating observations

We make two key observations about text-to-image model's features.

# Key motivating observations

We make two key observations about text-to-image model's features.

Diffusion models enable dense **semantic correspondence** [1]; hence they already possess the **required feature resolution** for part-level discrimination.

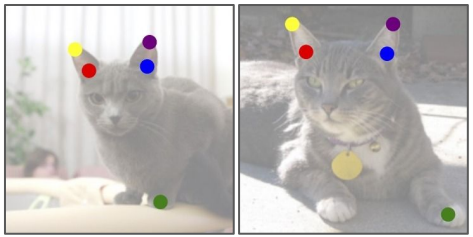


*Tang, Luming, et al. "Emergent correspondence from image diffusion." Advances in neural information processing systems 36 (2023): 1363-1389.*

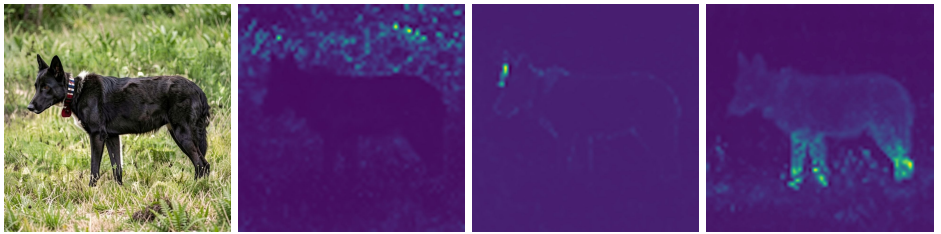
# Key motivating observations

We make two key observations about text-to-image model's features.

Diffusion models enable dense **semantic correspondence** [1]; hence they already possess the **required feature resolution** for part-level discrimination.



The dispersed cross-attention maps imply that **text-image interaction is a bottleneck**; thus failing to leverage the diffusion latents!



*Cross-attention map with the* `left`, `front` *and* `leg` *tokens.*

*Tang, Luming, et al. "Emergent correspondence from image diffusion." Advances in neural information processing systems 36 (2023): 1363-1389.*

# Research question

Our observations suggest that part phrases (e.g., left ear, right eye) are **not well represented on the text-image manifold**.

## Research question

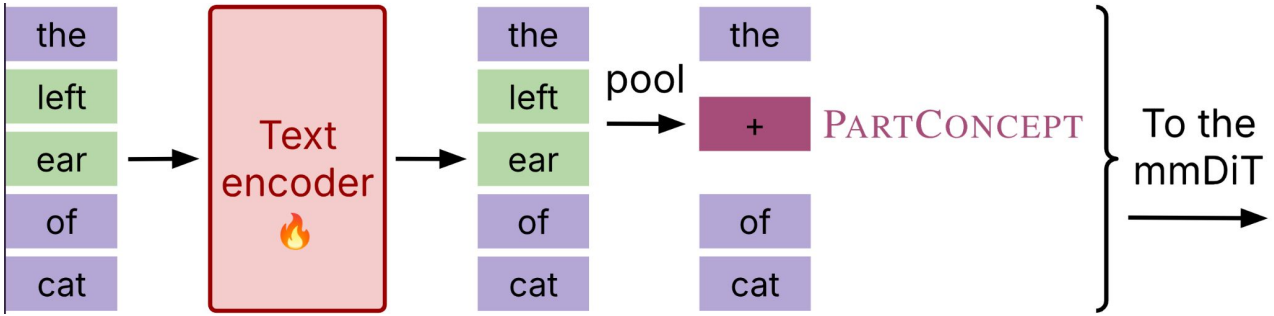
Our observations suggest that part phrases (e.g., left ear, right eye) are **not well represented on the text-image manifold**.

Can we introduce **textual part concepts** into the T2I model, while preserving its base capabilities?

# Method

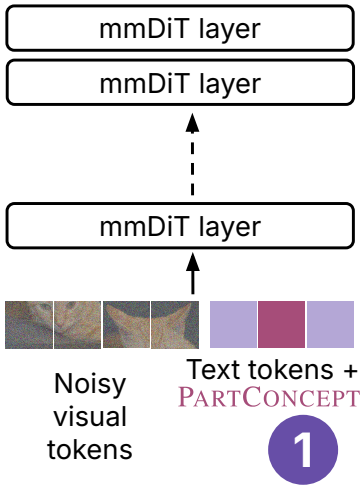
# **1.** Encoding a textual part description into a PartConcept token

- Each part-phrase (e.g., *left ear*) is mean pooled into a compact PartConcept token, which represents a **coherent semantic signal** rather than disparate tokens.
- Selective layers of the text-encoder are LoRA finetuned; this enables the PartConcept token to learn part-level semantics.



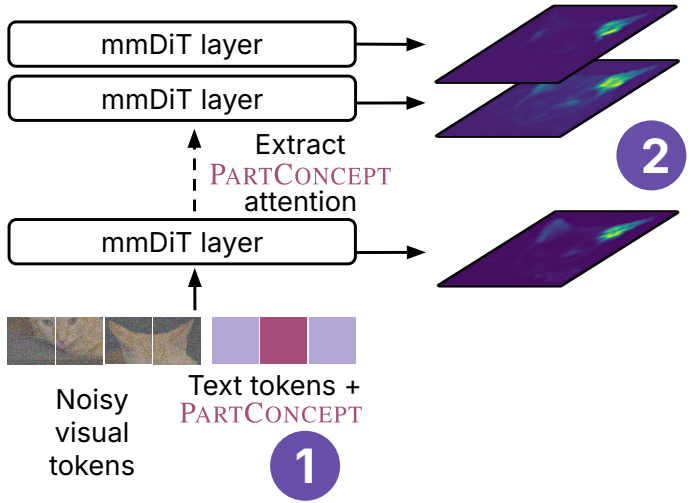
# **2.** The training objective

1. Noisy visual latents, and text + PARTCONCEPT tokens are input to the mmDiT.



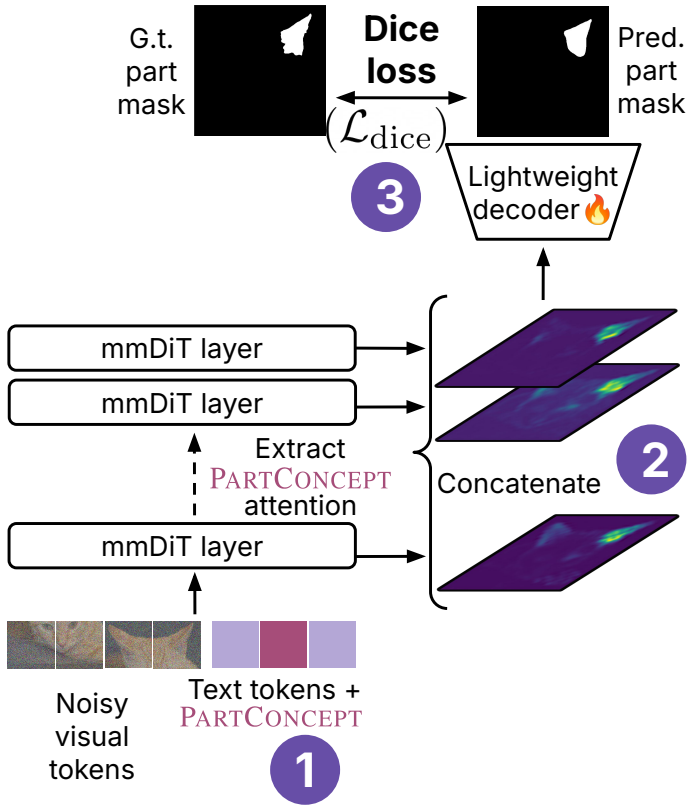
# **2.** The training objective

1. Noisy visual latents, and text + PARTCONCEPT tokens are input to the mmDiT.
2. Layer-wise cross attention maps are extracted.



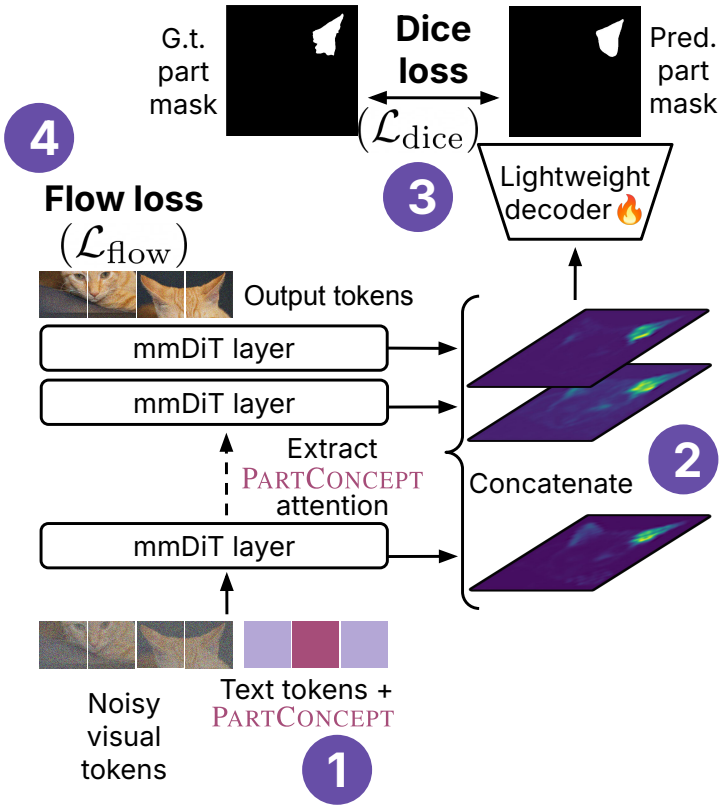
# **2.** The training objective

1. Noisy visual latents, and text + PARTCONCEPT tokens are input to the mmDiT.
2. Layer-wise cross attention maps are extracted.
3. Layer-wise concatenated attention maps are decoded to segmentation masks, and penalized using **Dice loss.**



# **2.** The training objective

1. Noisy visual latents, and text + PARTCONCEPT tokens are input to the mmDiT.
2. Layer-wise cross attention maps are extracted.
3. Layer-wise concatenated attention maps are decoded to segmentation masks, and penalized using **Dice loss.**
4. **Flow loss** on the velocity prediction; this preserves the generative capabilities.



## **3.** Balancing localization and generation

- Both the dice loss (localization) and flow loss (generation) are balanced during training:

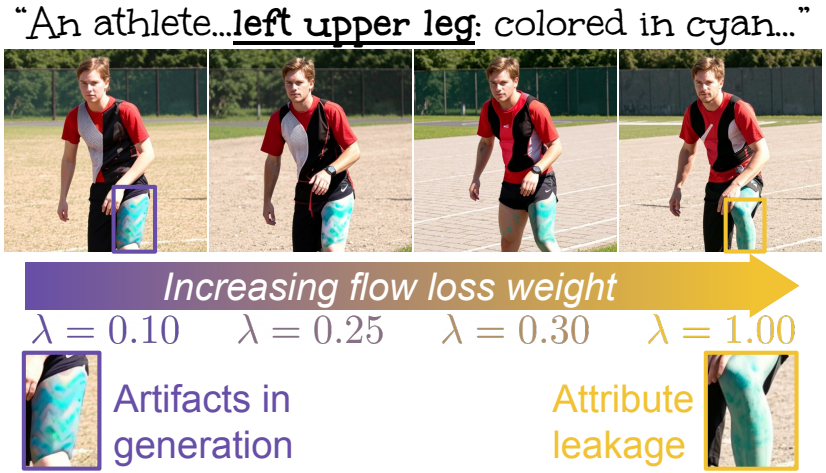
$$\mathcal{L} = \mathcal{L}_{\text{dice}} + \lambda \mathcal{L}_{\text{flow}}$$

# **3.** Balancing localization and generation

- Both the dice loss (localization) and flow loss (generation) are balanced during training:

$$\mathcal{L} = \mathcal{L}_{\text{dice}} + \lambda \mathcal{L}_{\text{flow}}$$

- λ governs a tradeoff between localization accuracy and generation quality, as shown

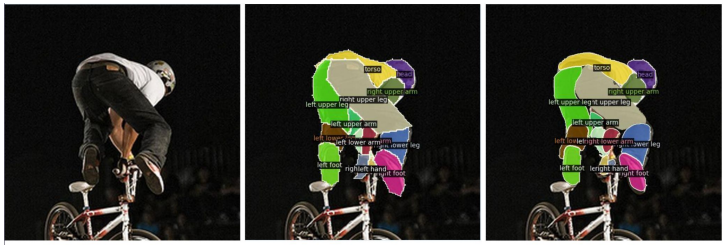


# Evaluating PartConcepts

# Evaluating PartConcepts

- We evaluate our method PartConcepts for localization and generation capabilities
  - Part instance segmentation
  - Part-level instruction following in text-to-image generation

## Part instance segmentation



Input image | Ground truth | **Ours**

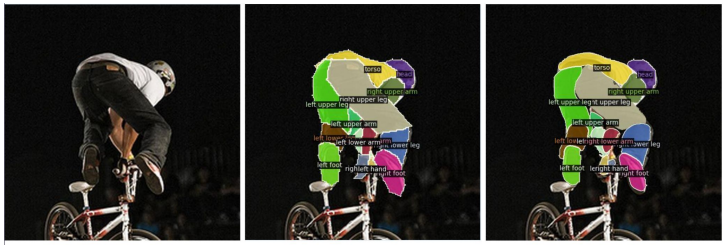
## Part-level instruction following



# Evaluating PartConcepts

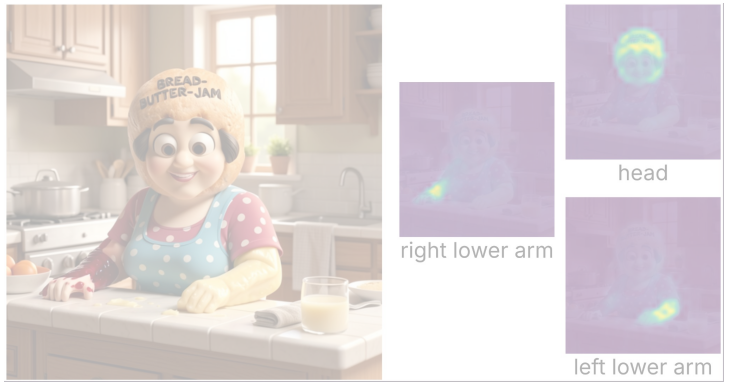
- We evaluate our method PartConcepts for localization and generation capabilities
  - **Part instance segmentation**
  - Part-level instruction following in text-to-image generation

Part instance segmentation



Input image | Ground truth | **Ours**

Part-level instruction following



# Evaluating PartConcepts for part-instance segmentation

(A) Ground Truth

**(B) PartConcepts**

(C) PartCATSeg (CVPR-25)

(D) LangHOPS(NeurIPS-25)



# Evaluating PartConcepts

- We evaluate our method PartConcepts for localization and generation capabilities
  - Part instance segmentation
  - **Part-level instruction following in text-to-image generation**

Part instance segmentation



Input image | Ground truth | **Ours**

Part-level instruction following



# Evaluating PartConcepts for part-level prompt following



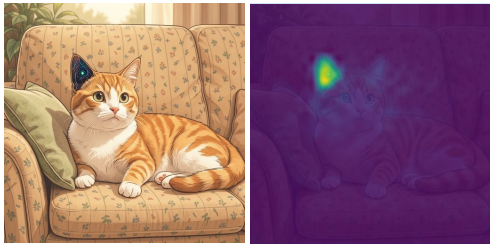
"... **left lower leg**: red glowing robotic ..."



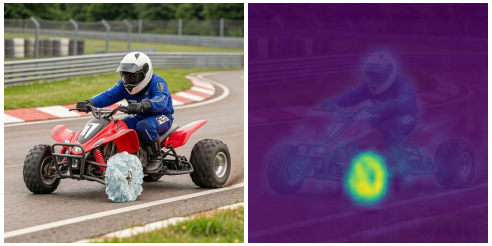
"... **right wrist**: wearing red band ..."



"... **left back leg**: made of spaghetti ..."



"... **right ear**: with cyber-blue veins ..."



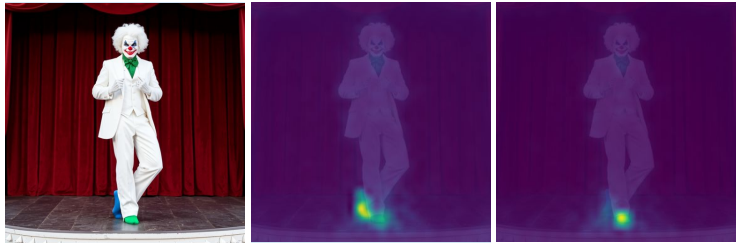
"... **left front wheel**: made of ice ..."



" ... **left armrest**: plated with gold ..."



"... **left arm**: made of fire ... **right arm**: made of ice ..."



" ...**left foot**: wearing blue sock ... **right foot**: wearing green sock ..."

# Summary of experiments

- PartConcepts enables state-of-the-art part instance segmentation
  - respects spatial organization of part instances (*left front leg* vs. *right back leg*)

# Summary of experiments

- PartConcepts enables state-of-the-art part instance segmentation
  - respects spatial organization of part instances (*left front leg* vs. *right back leg*)
- The textual attributes in the prompt naturally bind to the PartConcept token
  - e.g., *purple* binds to *left lower arm*
  - without need for explicit architectural mechanisms for attribute binding

"A man in an art studio with right upper arm with green stains and left lower arm with purple stains"



# Summary of experiments

- PartConcepts enables state-of-the-art part instance segmentation
  - respects spatial organization of part instances (*left front leg* vs. *right back leg*)
- The textual attributes in the prompt naturally bind to the PartConcept token
  - e.g., *purple* binds to *left lower arm*
  - without need for explicit architectural mechanisms for attribute binding
- **The PartConcept token behaves similar to a standard text embedding, but enriched with part-level semantics!**

"A man in an art studio with right upper arm with green stains and left lower arm with purple stains"



# Explore **PartConcepts**!

- See you at NeurIPS 2026!



https://visionpartconcepts.github.io