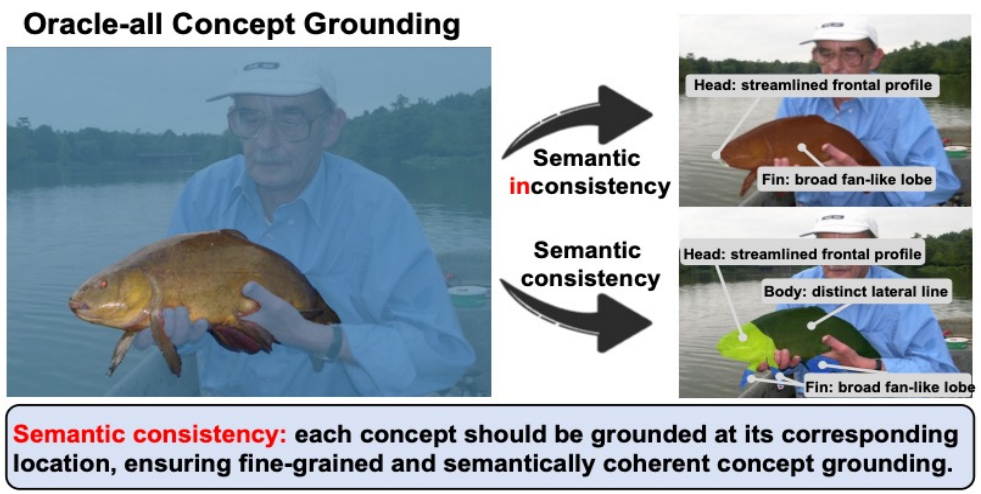
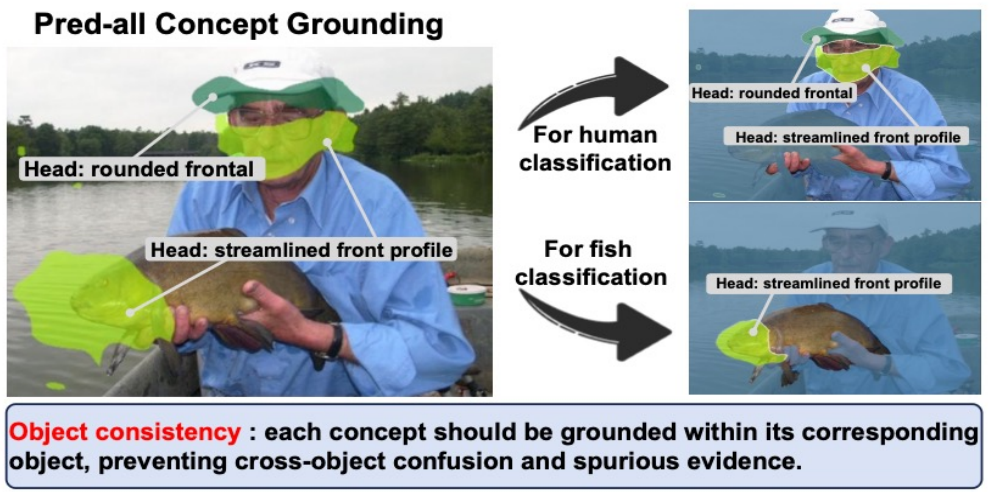
Rounded or Streamlined Head? Bridging Concept Bottleneck Models and Attribute-Described Object Parts

Yang Liu[1*†], Jiajin Zhang[1,4,5*†], Yaojun Hu[1,2], Bingguang Hao[2],
Xin Cao[3], Yingda Xia[1✉], Danyang Tu[1,4,5†✉], Shi Gu[2✉], Ling Zhang[1]
[1]DAMO Academy, Alibaba Group [2]Zhejiang University [3]UNSW Sydney [4]Hupan Lab
[5]Shanghai Jiao Tong University

[*]Equal contribution [†]Project lead [✉]Corresponding author

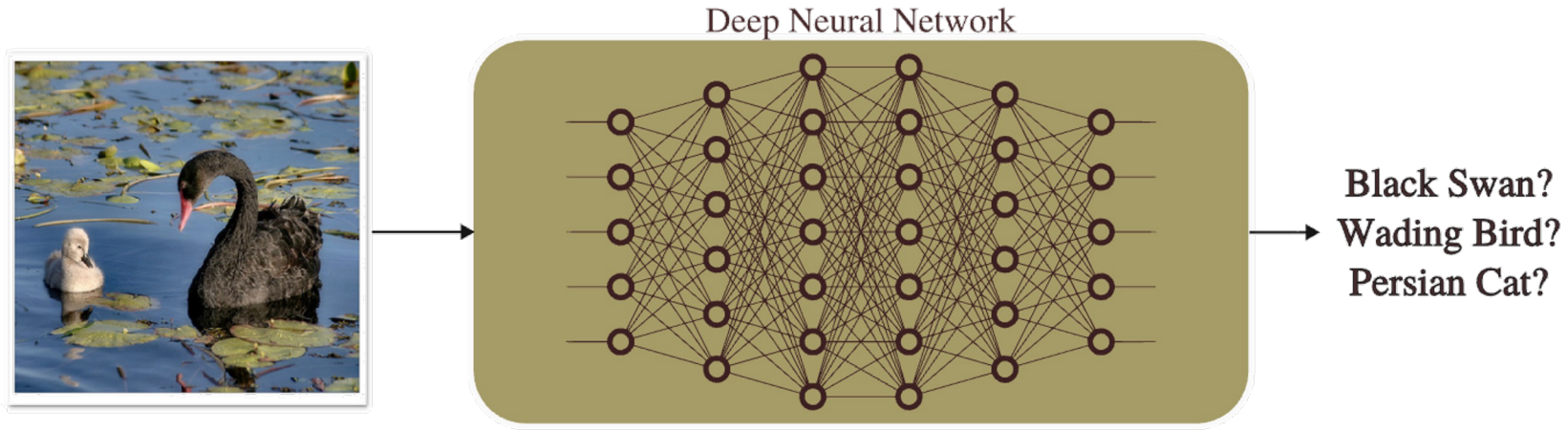
{yangye.ly, tudanyang.tdy}@alibaba-inc.com, gus@zju.edu.cn





Un-interpretable Decision Process

- DNNs are treated like *black-box models*:
  - Given an input, the model takes a decision via an **un-interpretable decision process.**
  - The model complexity, generally, **hinders any potential examination** of the underlying process.



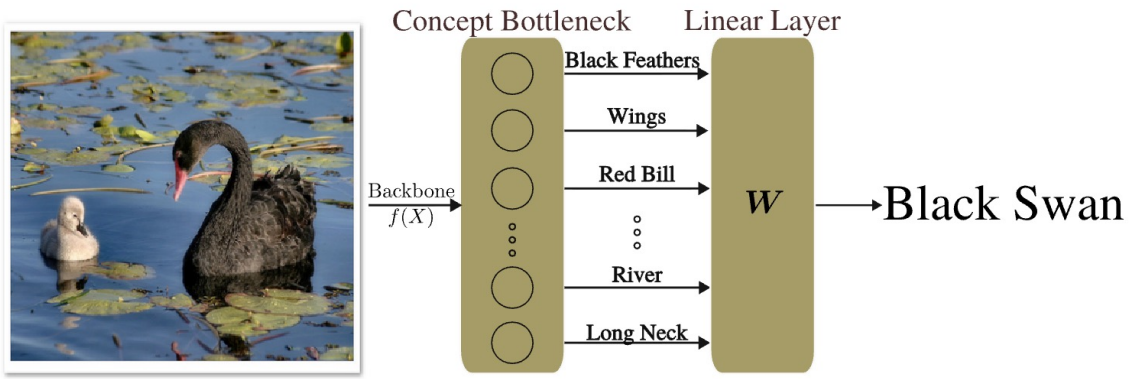
Lack of interpretability

Undesired property, especially in safety- or bias- aware applications ⇒ Crucial research and societal challenge

Textual vs Visual Explanations

達摩院 DAMO ACADEMY

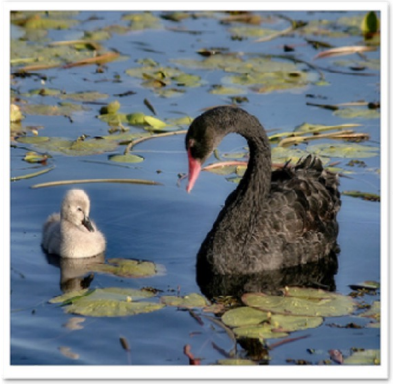
Textual Explanation : How ?

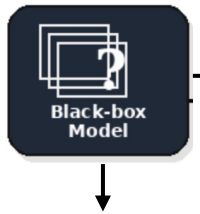


Concept Bottleneck Models

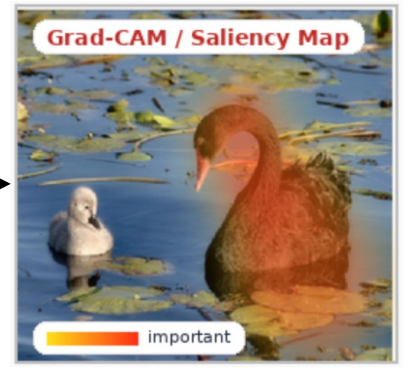
Textual explanation explains "what concepts there are and how they affect predictions"

Visual Explanation : Where ?





gradient



Grad-CAM / Saliency map / Attention map

Where dose the model look?

Visual explanation explains "where the evidence is in the image."

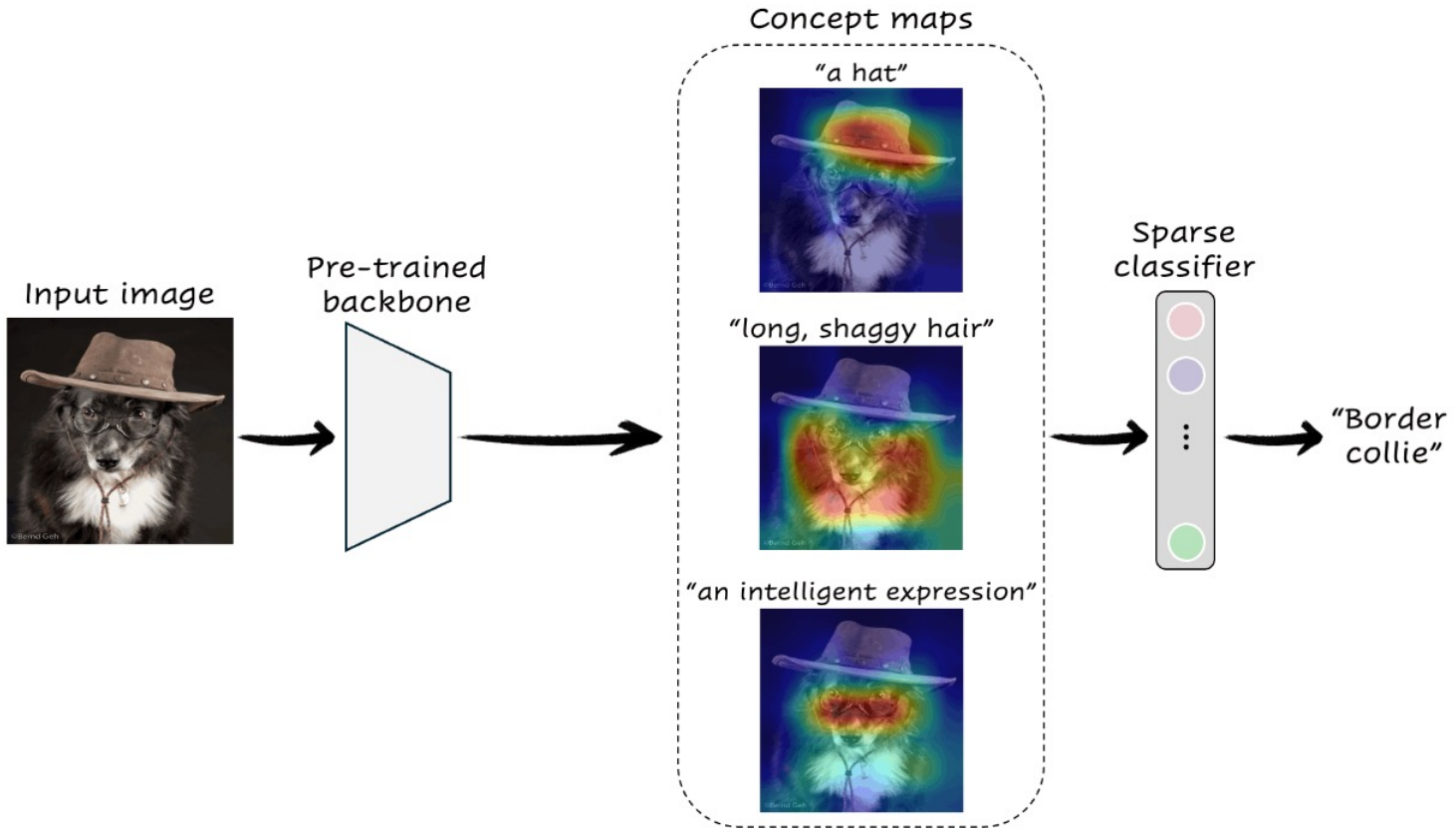
The motivation:
to connect the two, allowing each concept to be clearly articulated while also being associated with the correct location and object.

Decision Process: Where and How?

- The decision-making process cannot rely solely on heat maps: it requires verifiable conceptual evidence.

- CBM breaks down the prediction into "concept → category," but the concept score must come from the correct spatial area.

- VLM has the capability of open vocabulary image-text alignment; CBM provides editable semantic reasoning chains.



Connect the spatial positioning of VLM with the conceptual reasoning of CBM in a stable manner.

Grounding Consistency Gaps

The current explanation of VLM-CBM may "seem correct," but the evidence source is unstable.

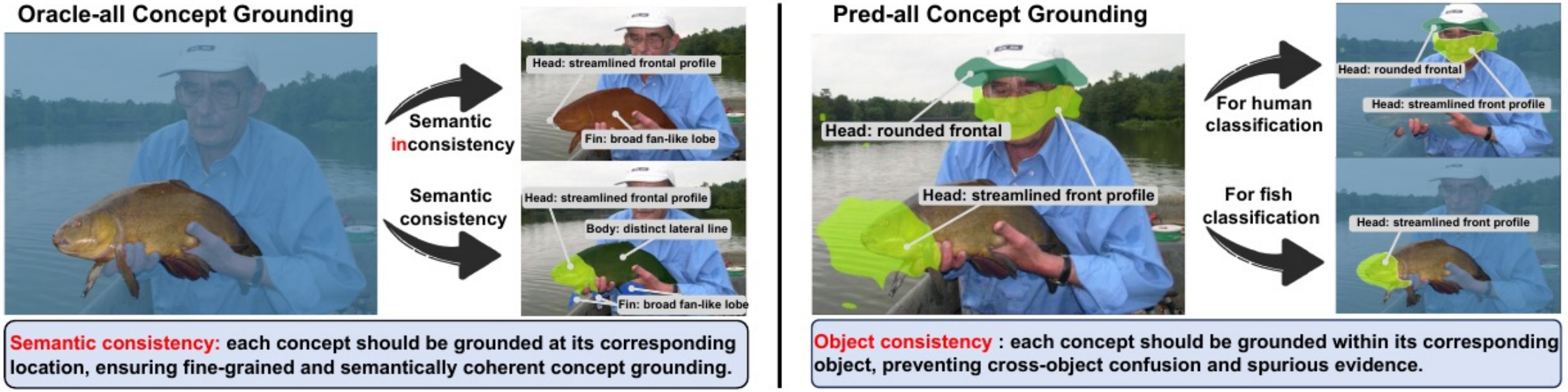


Figure 1: semantic consistency vs. object consistency

Concept activation not only affects interpretation but also transmits false evidence to the classification head.

Method

First obtain spatial evidence consistent with object, and carry out concept activation that is semantically consistent.

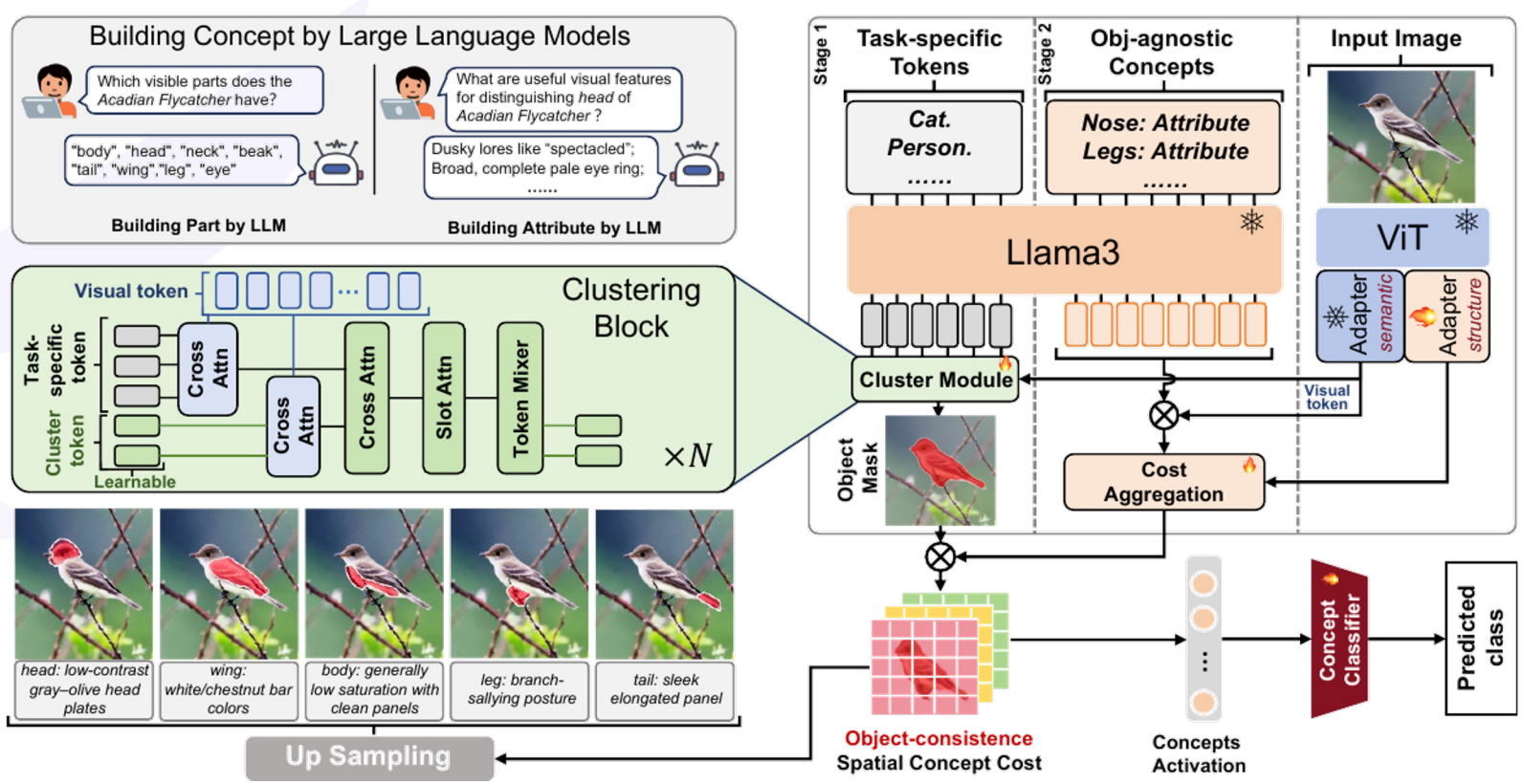


Figure 2: OA-CBM pipeline

Construction of part-attr concepts :
Object → Visible Part → Distinctive Attributes

HC output object mask:
Suppress non-target object responses

CA cost aggregation:
Stable pixel-concept correspondence

The cls head looks at concept activation
Maintain the interpretability of CBM

The final activation comes from $\bar{C} \odot M$: the concept cost map multiplied by the object mask and then pooled.

Data Construction： Part-Attribute Concept

To ensure fine-grained consistency, two sets of segmentation data have been organized.

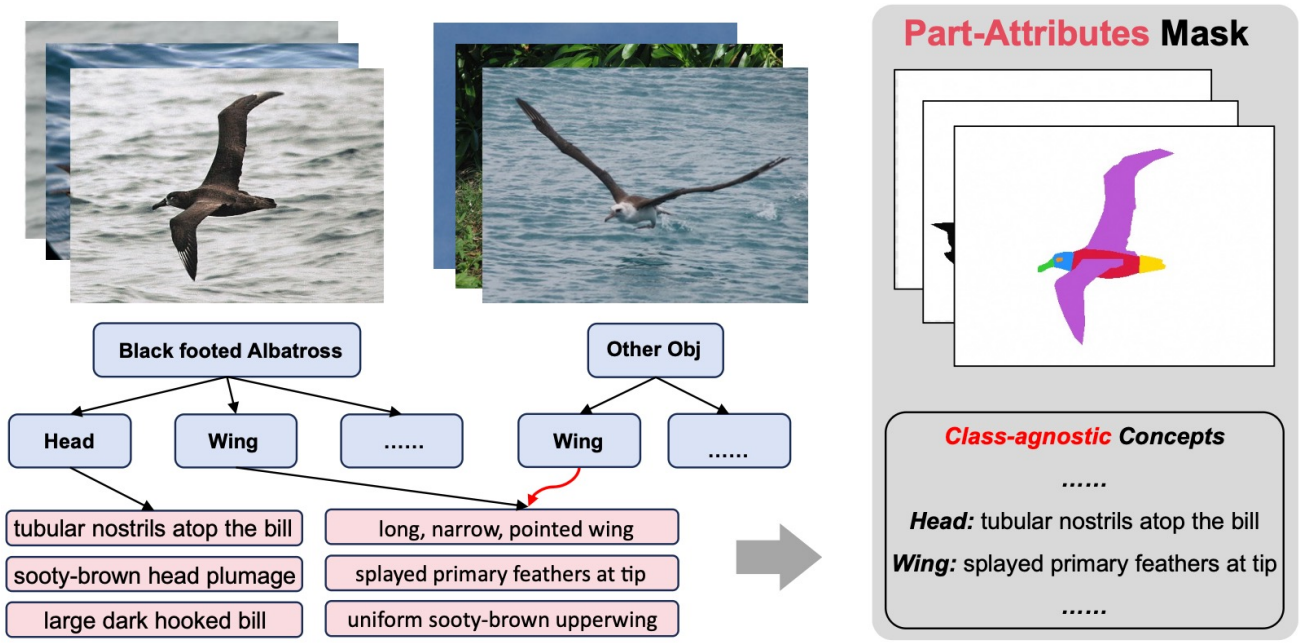


Figure 3: object-part-attribute tree → masks

PartAttrCUB

PartAttrImageNet

Three-step process

1) Collection & decomposition
PartCUB-70 + PartImageNet, LLM decomposes objects into visible parts and attributes.

2) Consolidation
Merge synonymous/near-synonymous attributes using SemHash, and then perform part-wise object clustering.

3) Mask alignment
Map the attributes back to pixel-level part masks, and perform manual sampling and conflict resolution.

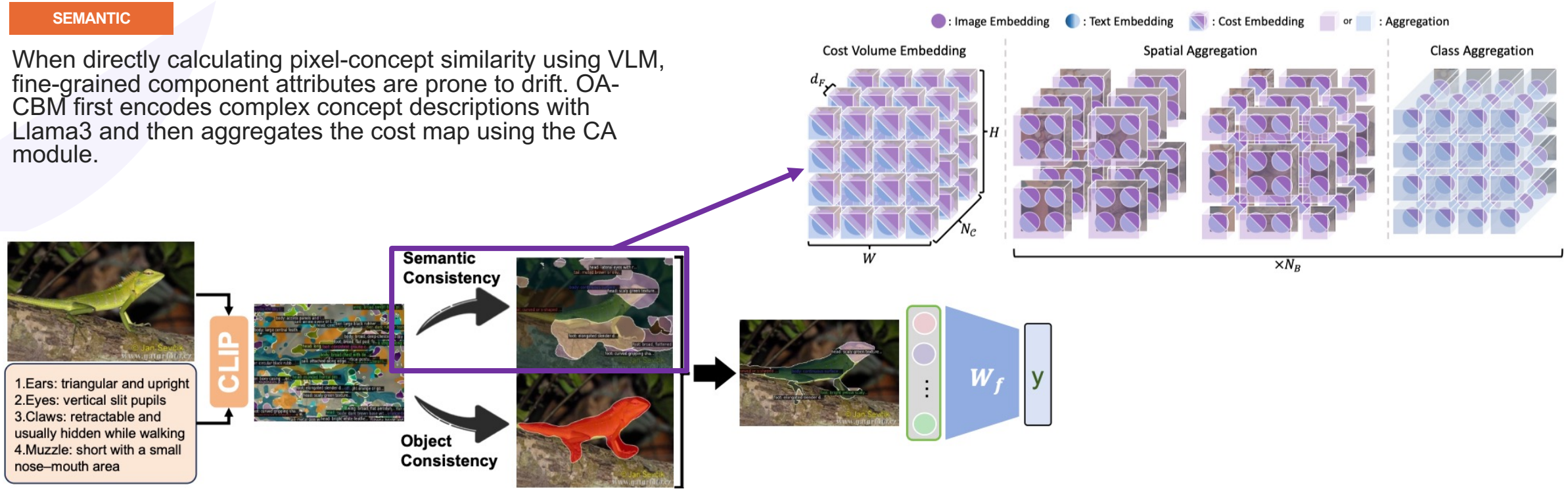
New Task Open Concept Grounding (OCG):
Determine whether the concept appears and locate where the concept is.

Semantic Consistency? Cost Aggregation (CA)

Let "head: streamlined frontal profile" other long descriptions consistently fall to the correct components.

SEMANTIC

When directly calculating pixel-concept similarity using VLM, fine-grained component attributes are prone to drift. OA-CBM first encodes complex concept descriptions with Llama3 and then aggregates the cost map using the CA module.



Spatial aggregation: Utilizing local continuity to suppress background noise.

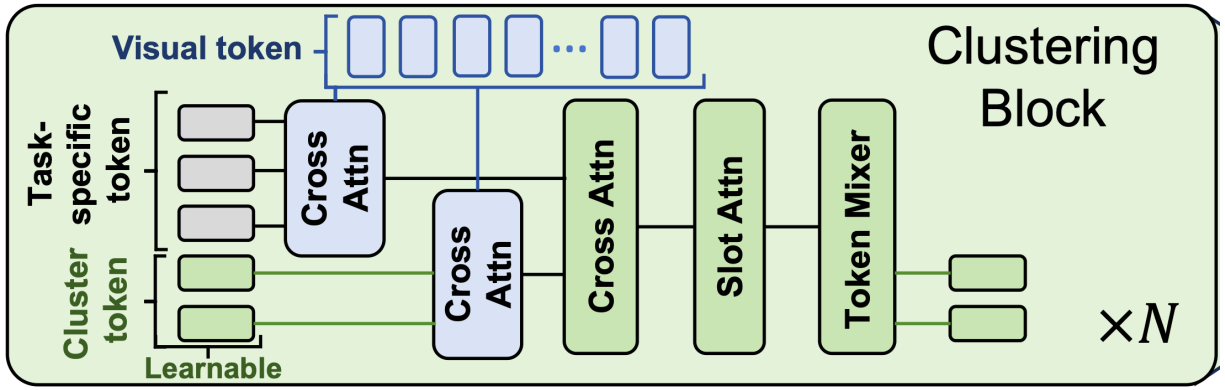
Concept aggregation: further understanding the between different concepts.

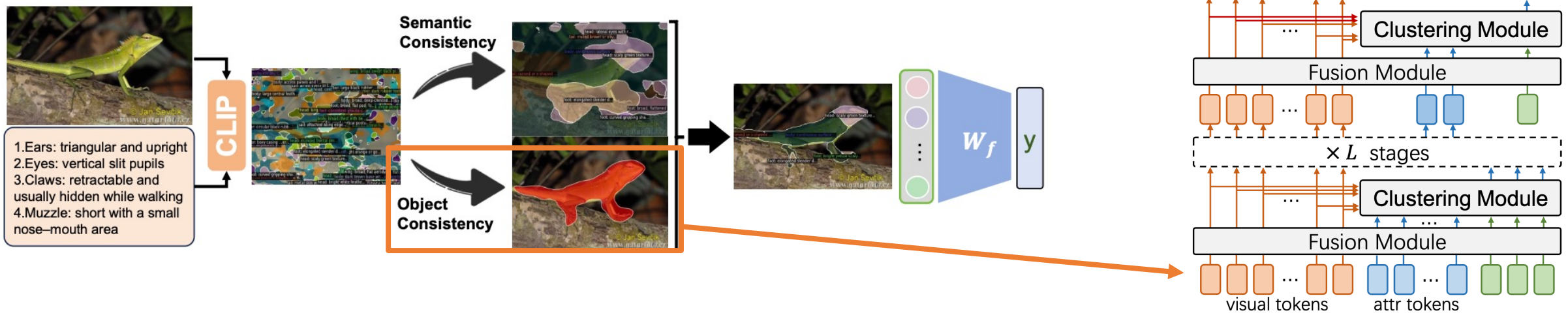
Object Consistency： Hierarchical Clustering (HC)

While keeping the concept class-agnostic, filter out erroneous objects.

OBJECT

Key contradiction: The concept should remain class-agn (for example, "head: rounded frontal"), but classifying, it is necessary to know whether this concept belongs to the target or the background/other objects.





Open Concept Grounding (OCG)

OCG simultaneously examines "which concepts exist" and "where they are."

方法	prior for the model	衡量的能力
Oracle-Obj	Object mask, category label	Concept positioning under the supervision of an ideal object
Oracle-All	Object mask, no category label	Concept generalization after removing category priors
Pred-All	No object mask, no category label	Fully self-predictive capability in a real open scene.

Grounding Metrics
- **mIoU**
- **h-IoU：harmonic mean of seen/unseen**

Classification Metrics
CUB-200 / ImageNet / CIFAR-100 / Food-101 / Flower-102
Top-1 Accuracy

Baseline：Labo、SALF-CBM、DOT-CBM、SC-CLIP/SC-CBM。

The most critical protocol is Pred-All: without oracle object information, it best reflects whether the object consistency is effective.

Main Result: Pred-All

When there are no oracle cues, object-aware grounding brings the greatest benefits.

Model	Concept Grounding on PartAttrImageNet									Concept Grounding on PartAttrCUB								
	Oracle-Obj			Oracle-All			Pred-All			Oracle-Obj			Oracle-All			Pred-All		
	Seen	Unseen	h-IoU	Seen	Unseen	h-IoU	Seen	Unseen	h-IoU	Seen	Unseen	h-IoU	Seen	Unseen	h-IoU	Seen	Unseen	h-IoU
CLIP (Labo)	16.1	18.1	17.0	2.2	3.6	2.8	1.4	2.7	1.9	14.8	14.0	14.4	3.8	3.8	3.8	2.5	0.9	1.3
DINO(DOT-CBM)	35.4	34.1	34.7	11.8	9.2	10.4	5.4	4.8	5.1	24.1	35.9	28.8	4.2	7.6	5.4	1.9	3.0	2.3
OCLIP(SALF-CBM)	28.3	26.8	27.6	16.9	13.5	15.0	7.4	6.8	7.1	16.5	40.7	23.5	11.2	22.0	14.9	2.9	9.6	4.4
SC-CLIP(SC-CBM)	30.8	31.3	31.1	15.1	11.9	13.3	10.1	9.5	9.8	14.7	12.0	13.2	6.4	2.0	3.1	3.8	0.4	0.6
Ours (w/o obj)	60.3	44.8	51.4	46.2	30.8	37.0	14.5	11.7	12.9	48.7	26.0	33.9	28.0	22.0	24.7	8.2	5.5	6.6
Ours (w/ obj)	**60.3**	**44.8**	**51.4**	**46.2**	**30.8**	**37.0**	**44.1**	**29.9**	**35.7**	**48.7**	**26.0**	**33.9**	**28.0**	**22.0**	**24.7**	**19.5**	**22.9**	**20.8**

Table 1. Comparison of zero-shot performance with state-of-the-art methods on our concept segmentation datasets. mIoU (%) under Oracle-Obj, Oracle-All and Pred-All protocols. h-IoU is the harmonic mean of Seen/Unseen.

Part-only: Still able to locate components after removing attributes

Even if only the part names such as "head / wing / tail" are given, OA-CBM still maintains a stronger correspondence.

Model (m-IoU)	Part Grounding			Obj
	Oracle-Obj	Oracle-All	Pred-All	
CLIP (Labo)	15.7	8.6	3.8	44.6
DINO (DOT-CBM)	22.7	11.8	4.0	45.4
CLIP (SALF-CBM)	16.1	13.2	5.1	35.4
SC-CLIP(SC-CBM)	16.8	18.7	6.3	13.1
Ours (w/o obj)	50.5	48.3	17.5	N/A
Ours (w/ obj)	**50.5**	**48.3**	**46.5**	**83.1**

Table 2. Comparison of zero-shot part-only performance with various methods on PartAttrImageNet datasets. mIoU (%) under Oracle-Obj, Oracle-All and Pred-All protocols.

PartAttrImageNet

Model (m-IoU)	Part Grounding			Obj
	Oracle-Obj	Oracle-All	Pred-All	
CLIP (Labo)	9.3	8.4	3.9	53.3
DINO (DOT-CBM)	15.3	15.3	3.6	46.1
CLIP (SALF-CBM)	15.0	15.1	4.7	43.2
SC-CLIP(SC-CBM)	3.6	3.6	2.1	34.8
Ours (w/o obj)	62.5	58.8	10.3	N/A
Ours (w/ obj)	**62.5**	**58.8**	**30.9**	**72.1**

Table 3. Comparison of zero-shot part-only performance with various methods on PartAttrCUB datasets. mIoU (%) under Oracle-Obj, Oracle-All and Pred-All protocols.

PartAttrCUB

Part-only Pred-All
ImageNet: 46.5 vs 6.3
CUB: 30.9 vs 4.7

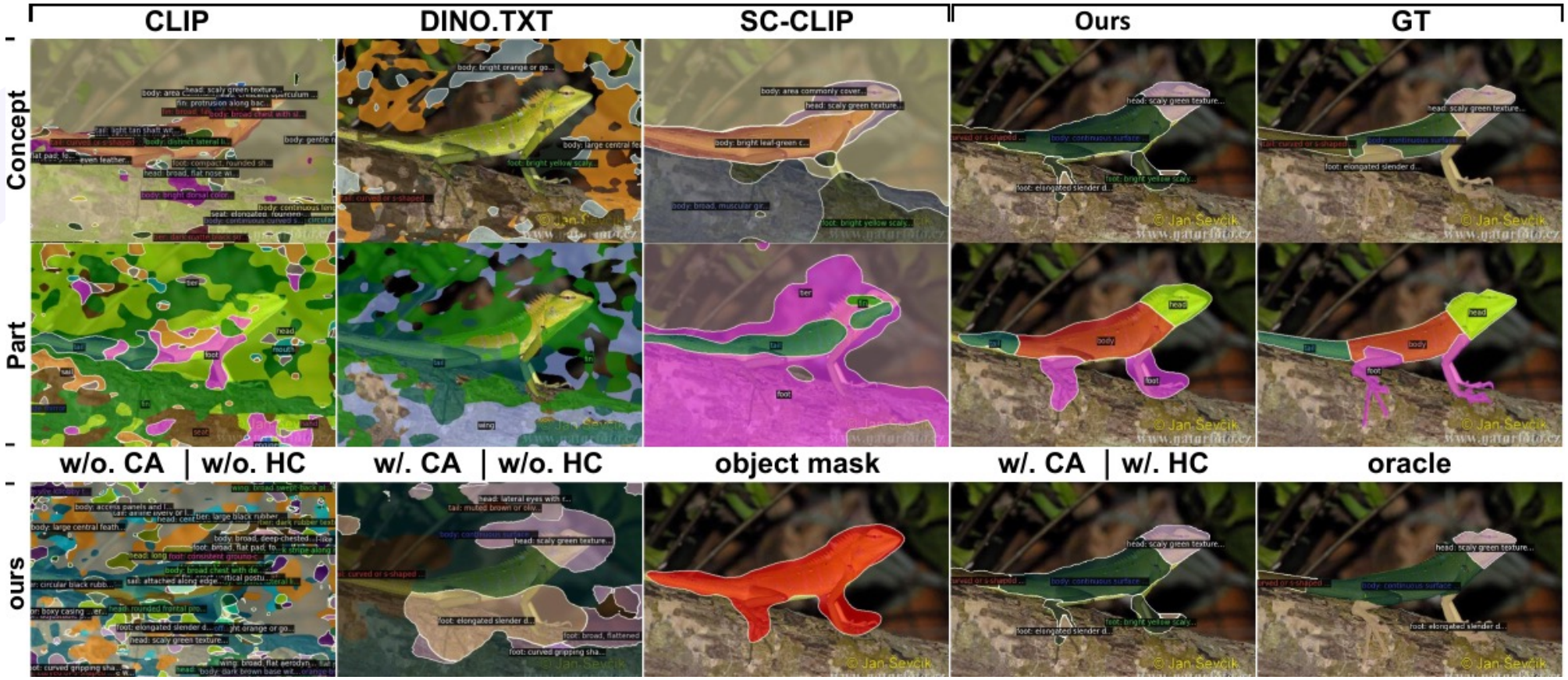
Object mask mIoU
ImageNet: 83.1
CUB: 72.1

Component-attribute training relies not only on attribute words; the learned object consistency and spatial correspondence can be transferred to a more simplified component concept.

Qualitative Results: From Noise Masks to Concepts Within Objects

達摩院 DAMO ACADEMY

Visualization of CA and HC progressively cleaning semantic noise and object confusion



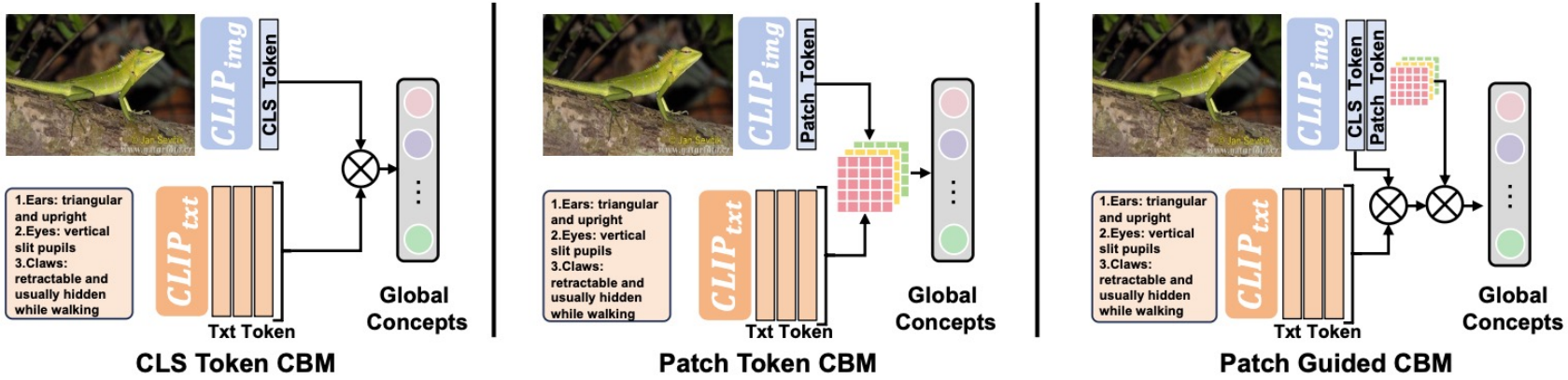
Observation 1: The concept masks of CLIP/DINO/SC-CLIP have background noise and component misalignment.

Observation 2: CA makes the boundaries of the same component more continuous and the semantic labeling more focused.

Observation 3: The HC-generated object mask is closer to the GT in the end.

Classification performance: interpretability without sacrificing accuracy

While maintaining evidence of spatial concepts, OA-CBM maintains competitive Top-1 accuracy.



Model	Classification Accuracy (CLS Token)				
	CUB200	ImageNet	CIFAR100	FOOD101	FLOWER
Labo	73.5	70.0	87.2	93.2	97.1
SC-CBM	78.3	74.2	88.0	94.2	97.8
DOT-CBM	77.3	78.7	**97.3**	93.0	99.4
SALF-CBM	73.1	69.9	86.4	94.1	97.0
Ours	**80.7**	**83.1**	97.2	**94.5**	**98.4**

Table 4. Performance comparison of classification accuracy

Model	Classification Accuracy (Patch Token)			
	CUB200	CIFAR100	FOOD101	FLOWER
Labo	1.7	4.5	1.3	58.5
SC-CBM	36.6	84.6	**91.7**	88.9
DOT-CBM	58.6	92.7	88.4	88.4
Ours	**63.3**	**96.1**	87.3	**91.8**

Table 5. Performance comparison of classification accuracy

Model	Classification Accuracy (Patch Guided)			
	CUB200	CIFAR100	FOOD101	FLOWER
Labo	71.7	74.5	89.2	96.5
SC-CBM	69.2	82.4	91.8	98.1
DOT-CBM	74.5	91.8	90.5	**99.2**
Ours	**80.6**	**96.8**	**94.2**	97.8

Table 6. Performance comparison of classification accuracy

Ablation & Contribution

The functions of the two modules are complementary: CA brings a usable concept cost map, while HC provides object scope in Pred-All.

CA For semantics, HC For objects.

CA	HC	Classification (ACC)				Grounding (h-IoU)	
		CUB200	CIFAR100	FOOD101	FLOWER	PartAttr CUB	PartAttr ImageNet
✗	✗	33.8	86.9	92.0	85.3	1.3	1.9
✗	✔	0.01	86.1	0.1	25.6	0.6	2.2
✔	✗	63.2	96.1	87.7	92.8	6.6	12.9
✔	✔	63.3	96.2	87.3	91.8	20.8	35.7

Table 7. Ablation studies on classification and concept grounding.

Contribution

1. **Concept Definition**: Change from object/category concepts to part-attribute pairs.

2. **Model design**: CA ensures semantic consistency, HC ensures object consistency.

3. **Data and evaluation**: PartAttrCUB, PartAttrImage, and OCG.