ICLR 2026

TECHNION | TEL AVIV UNIVERSITY

DeLeaker

Dynamic Inference-Time Reweighting For Semantic Leakage Mitigation in Text-to-Image Models

Mor Ventura*, Michael Toker*, Or Patashnik, Yonatan Belinkov, Roi Reichart





So, what exactly is *semantic leakage*?

“The unintended cross-feature transfer of attributes between semantically related entities”.

“A **raccoon** climbing a tree and a **possum** walking below”

Text-to-Image Model (TTI)

How To Measure Semantic Leakage Mitigation?

Semantic Leakage in IMages (SLIM) Dataset

1,130 human-verified examples of semantic leakage

Semantic Leakage in IMages (SLIM) Dataset: Curation Process

Dataset Design

Number of Entities

"A **bat** is flying high in the night sky while an **owl** is perched on a tree branch and a **moth** flutters"

Multiple Entities (animals)

Triggers Semantic Leakage

Semantic Leakage in IMages (SLIM) Dataset: Curation Process

Dataset Design

↑ Number of Entities

↑ **Fine-grained Categories**

"A **cow** and a **horse** in a farm"

A **strawberry**, a **radish**, and a **cherry** sit together.

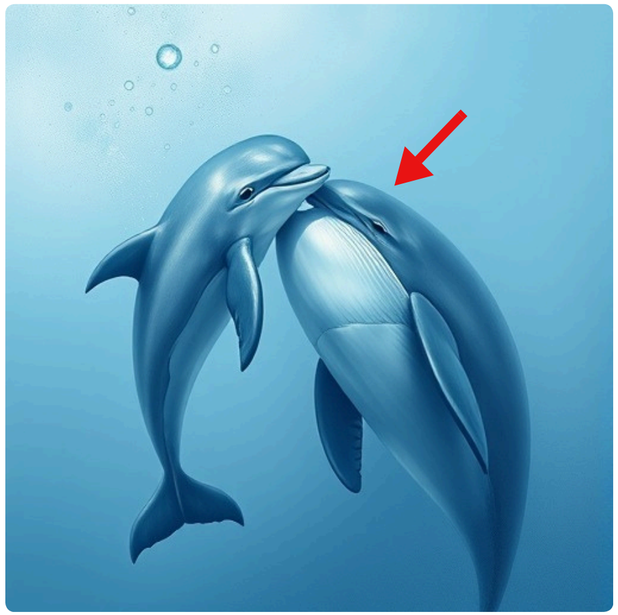
↑ Triggers Semantic Leakage

Semantic Leakage in IMages (SLIM) Dataset: Curation Process

Dataset Design

- ⬆ Number of Entities
- ⬆ Fine-grained Categories
- ⬆ **Shared Interactions**

"The **dolphin** and the **whale** are cuddling together"



Interactions

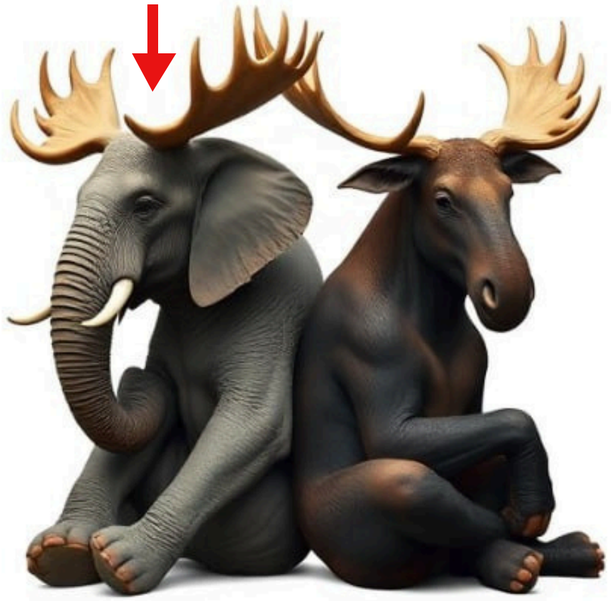
⬆ Triggers Semantic Leakage

Semantic Leakage in IMages (SLIM) Dataset: Curation Process

Dataset Design

- ↑ Number of Entities
- ↑ Fine-grained Categories
- ↑ ? Shared Interactions
- ↑ **Shared Visual Style**

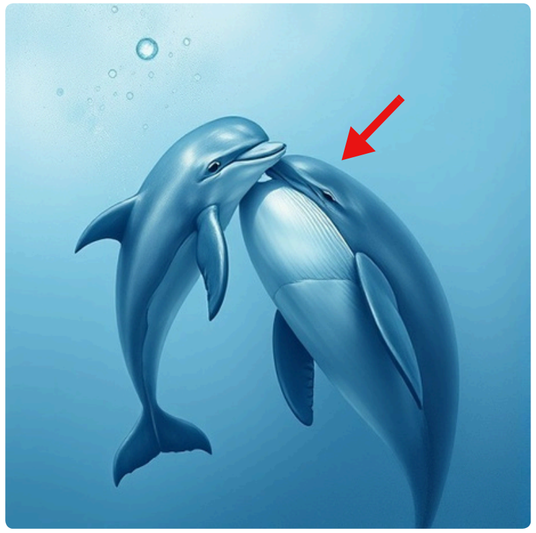
"An **elephant** and a **moose** are sitting back to back *in a 3D render* "



Interactions with *style*

↑ Triggers Semantic Leakage

Teaser: DeLeaker (SLIM)

	"A **cow** and a **horse** in a farm"	"The **dolphin** and the **whale** are cuddling together"	"An **elephant** and a **moose** are sitting back to back *in a 3D render* "	"A **bat** is flying high in the night sky while an **owl** is perched on a tree branch and a **moth** flutters"
Original (Flux)				
	Pairs	Interactions	Interactions with *style*	Multiple Entities (animals)

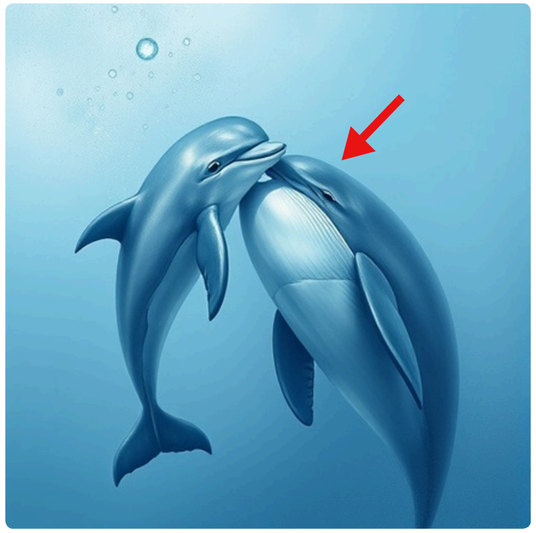
SLIM: 1,130 (prompt, image with semantic leakage) samples

Teaser: DeLeaker (SLIM)

"A **cow** and a **horse** in a farm"

Pairs

"The **dolphin** and the **whale** are cuddling together"



Interactions

"An **elephant** and a **moose** are sitting back to back *in a 3D render* "

Interactions with *style*

"A **bat** is flying high in the night sky while an **owl** is perched on a tree branch and a **moth** flutters"

Multiple Entities (animals)

DeLeaker (Ours)

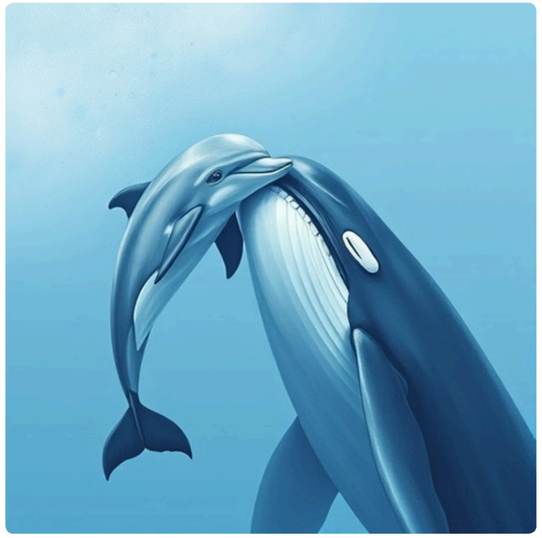
Teaser: DeLeaker (SLIM)

"A **cow** and a **horse** in a farm"

Pairs

DeLeaker (Ours)

"The **dolphin** and the **whale** are cuddling together"



Interactions

DeLeaker (Ours)

"An **elephant** and a **moose** are sitting back to back *in a 3D render* "

Interactions with *style*

"A **bat** is flying high in the night sky while an **owl** is perched on a tree branch and a **moth** flutters"

Multiple Entities (animals)

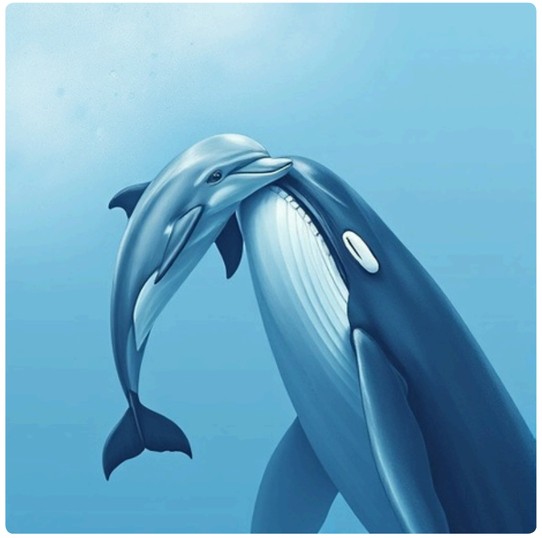
Teaser: DeLeaker (SLIM)

"A **cow** and a **horse** in a farm"

Pairs

DeLeaker (Ours)

"The **dolphin** and the **whale** are cuddling together"



Interactions

DeLeaker (Ours)

"An **elephant** and a **moose** are sitting back to back *in a 3D render* "

Interactions with *style*

DeLeaker (Ours)

"A **bat** is flying high in the night sky while an **owl** is perched on a tree branch and a **moth** flutters"

Multiple Entities (animals)

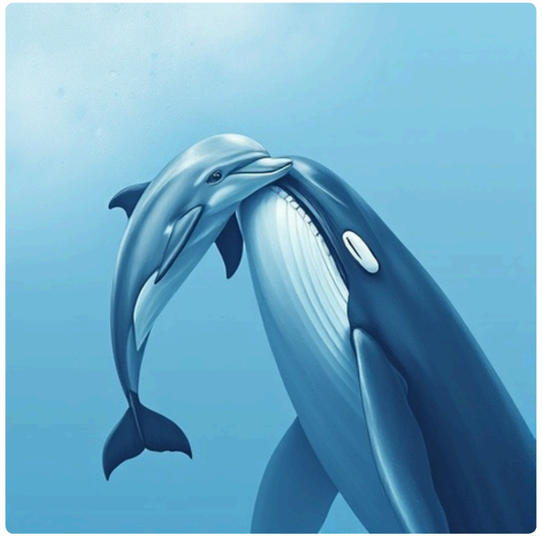
Teaser: DeLeaker (SLIM)

"A **cow** and a **horse** in a farm"

Pairs

DeLeaker (Ours)

"The **dolphin** and the **whale** are cuddling together"



Interactions

DeLeaker (Ours)

"An **elephant** and a **moose** are sitting back to back *in a 3D render* "

Interactions with *style*

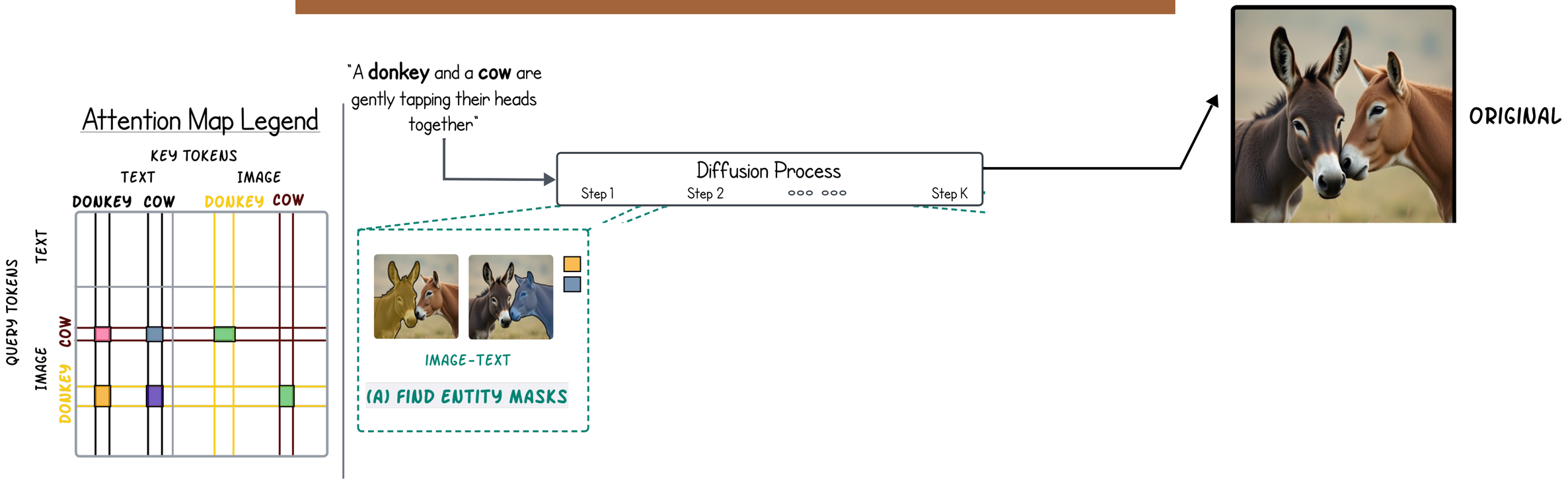
DeLeaker (Ours)

"A **bat** is flying high in the night sky while an **owl** is perched on a tree branch and a **moth** flutters"

Multiple Entities (animals)

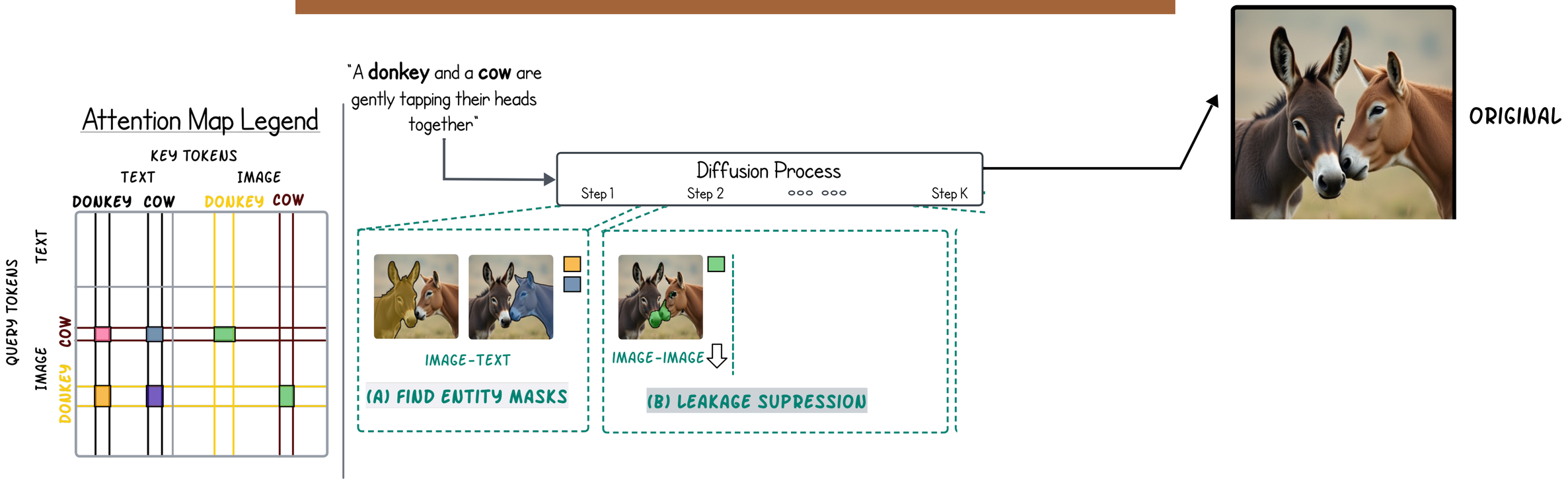
DeLeaker (Ours)

Method: DeLeaker



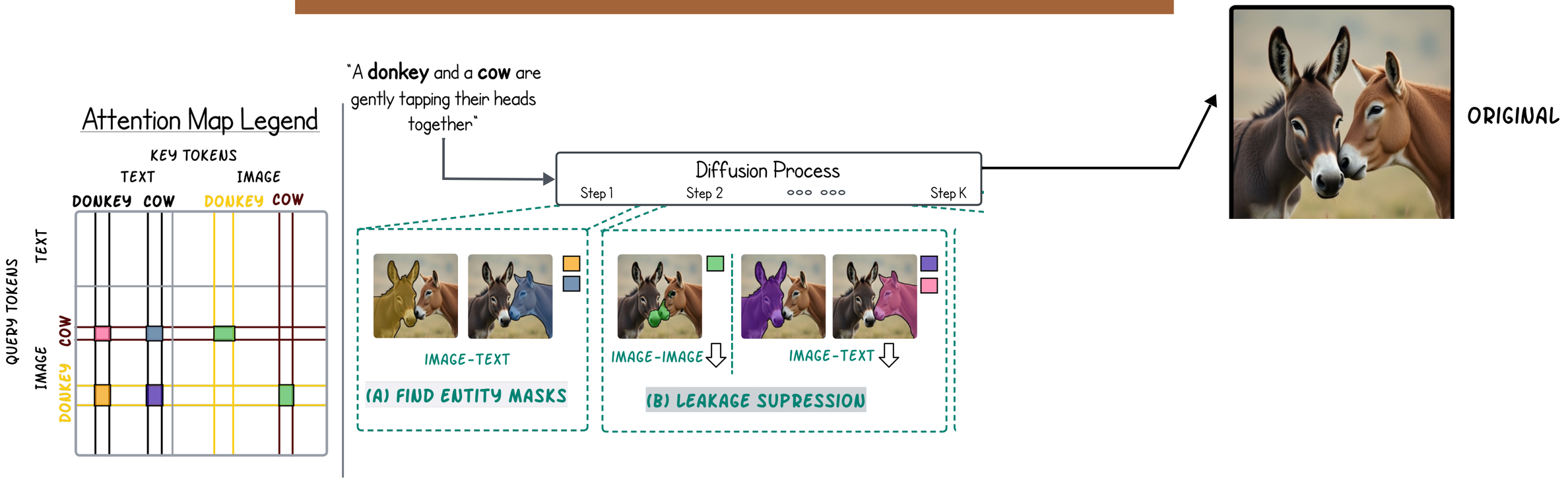
1) $$\mathcal{E}_i^{\text{img}} = \{q \in \mathcal{I} \mid Attn_{qk} > \mu_i + \beta_1 \cdot \sigma_i, k \in (\mathcal{E}_i^{\text{txt}} \cap \mathcal{I})\}$$

Method: DeLeaker



2) $$\mathrm{H}_{ij}^{\text{img-img}} = \{(q,k) \mid Attn_{qk} > \mu_{ij} + \beta_2 \cdot \sigma_{ij}, q,k \in \mathcal{I}\}$$

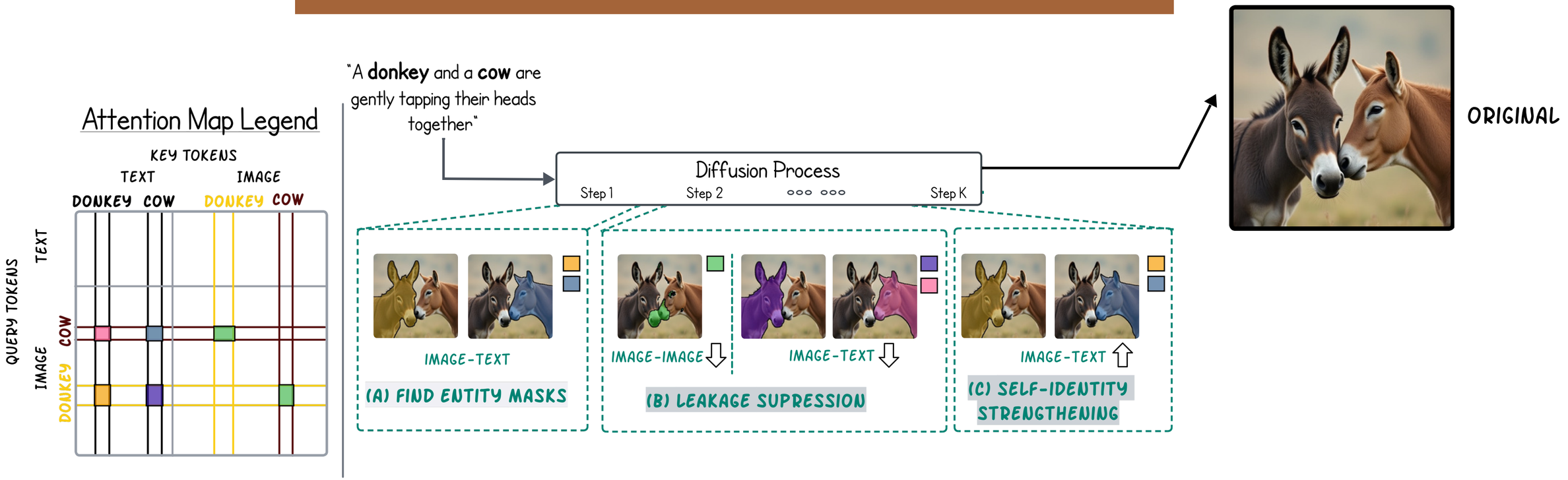
Method: DeLeaker



3)

$$Attn'_{qk} = \begin{cases} -\infty & \text{if } q \in \mathcal{E}_i^{\text{img}}, k \in \mathcal{E}_j^{\text{img}}, \text{ and } (q,k) \in \mathrm{H}_{ij}^{\text{img-img}} \\ -\infty & \text{if } q \in \mathcal{E}_i^{\text{img}}, k \in \mathcal{E}_j^{\text{txt}} \end{cases}$$

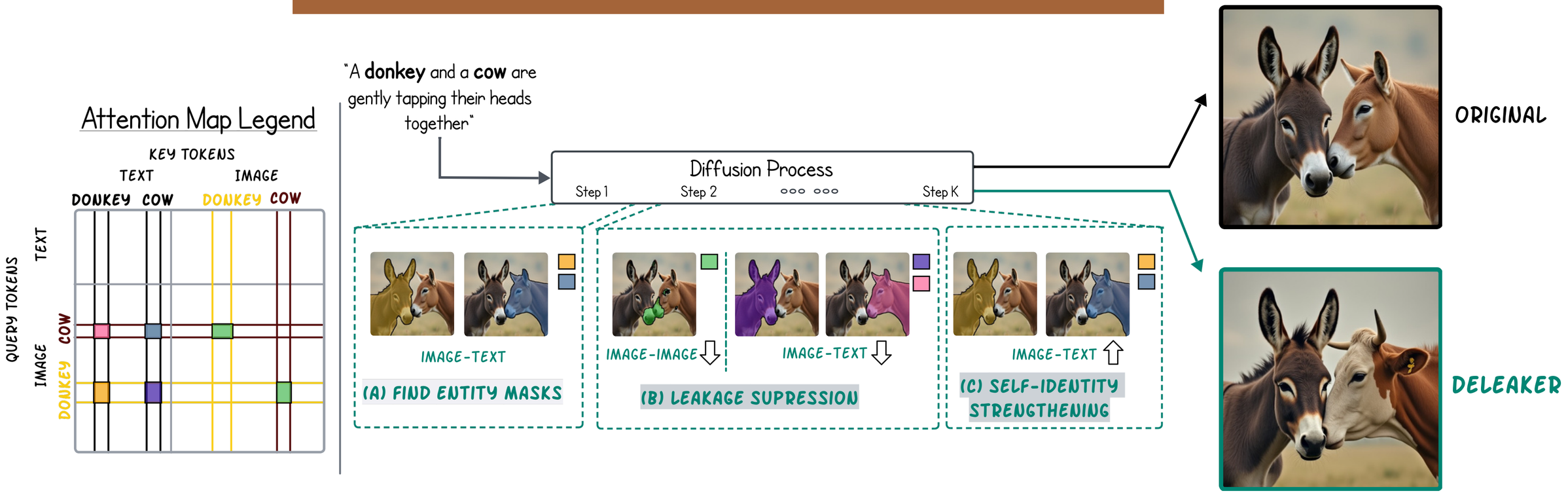
Method: DeLeaker



3)

$$Attn'_{qk} = \begin{cases} -\infty & \text{if } q \in \mathcal{E}_i^{\text{img}}, k \in \mathcal{E}_j^{\text{img}}, \text{ and } (q,k) \in \mathrm{H}_{ij}^{\text{img-img}} \\ -\infty & \text{if } q \in \mathcal{E}_i^{\text{img}}, k \in \mathcal{E}_j^{\text{txt}} \\ \alpha \cdot Attn_{qk} & \text{if } q \in \mathcal{E}_i^{\text{img}}, k \in \mathcal{E}_i^{\text{txt}} \\ Attn_{qk} & \text{else} \end{cases}$$

Method: DeLeaker



3)

$$Attn'_{qk} = \begin{cases} -\infty & \text{if } q \in \mathcal{E}_i^{\text{img}}, k \in \mathcal{E}_j^{\text{img}}, \text{ and } (q,k) \in \mathrm{H}_{ij}^{\text{img-img}} \\ -\infty & \text{if } q \in \mathcal{E}_i^{\text{img}}, k \in \mathcal{E}_j^{\text{txt}} \\ \alpha \cdot Attn_{qk} & \text{if } q \in \mathcal{E}_i^{\text{img}}, k \in \mathcal{E}_i^{\text{txt}} \\ Attn_{qk} & \text{else} \end{cases}$$

How To Measure Semantic Leakage Mitigation?

Semantic Leakage in IMages (SLIM) Dataset

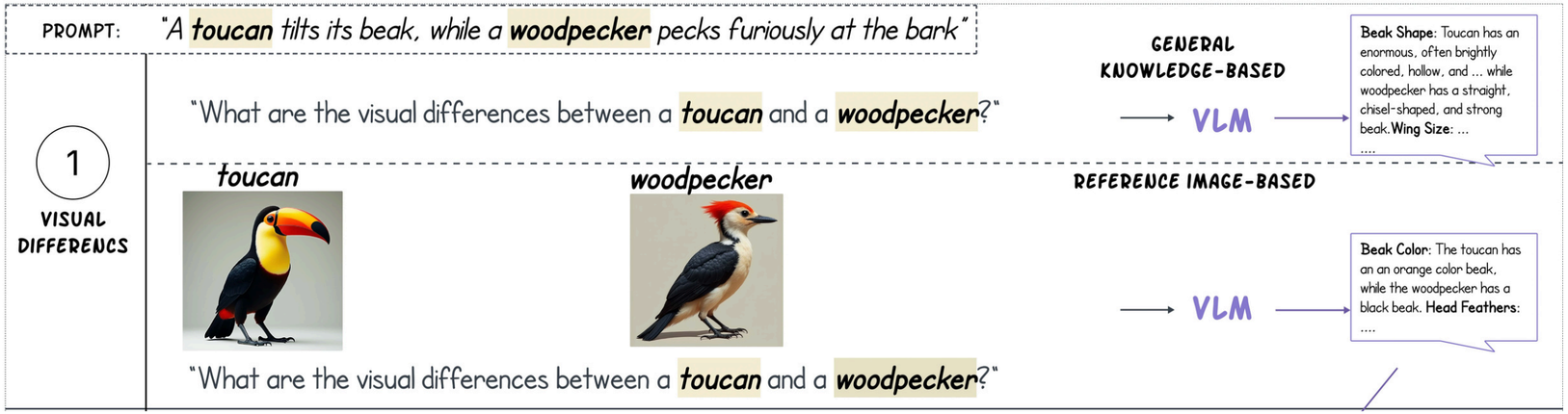
1,130 human-verified examples of semantic leakage

Automatic Evaluation Framework

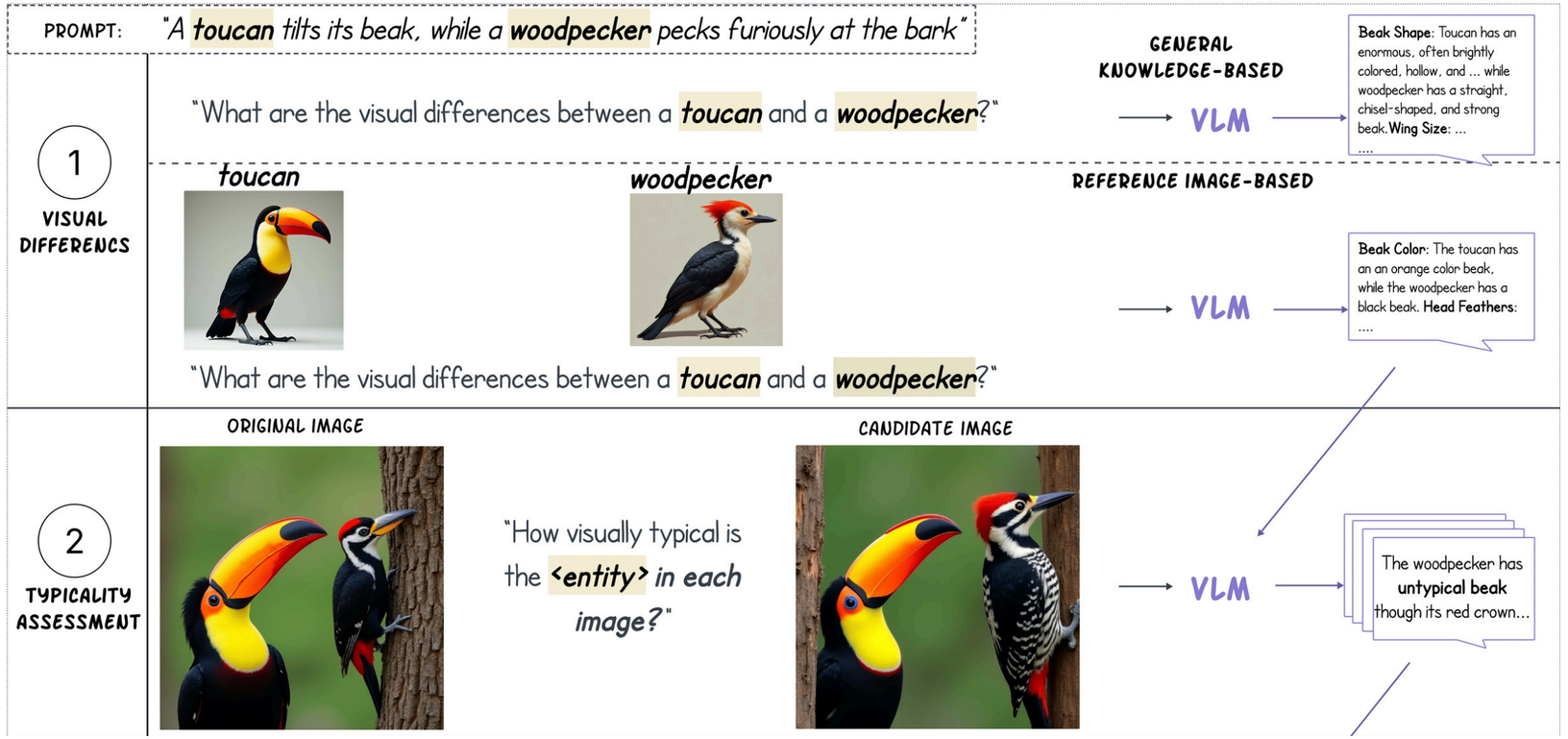
Automated pipeline for reproducible leakage mitigation scoring

Automatic Evaluation

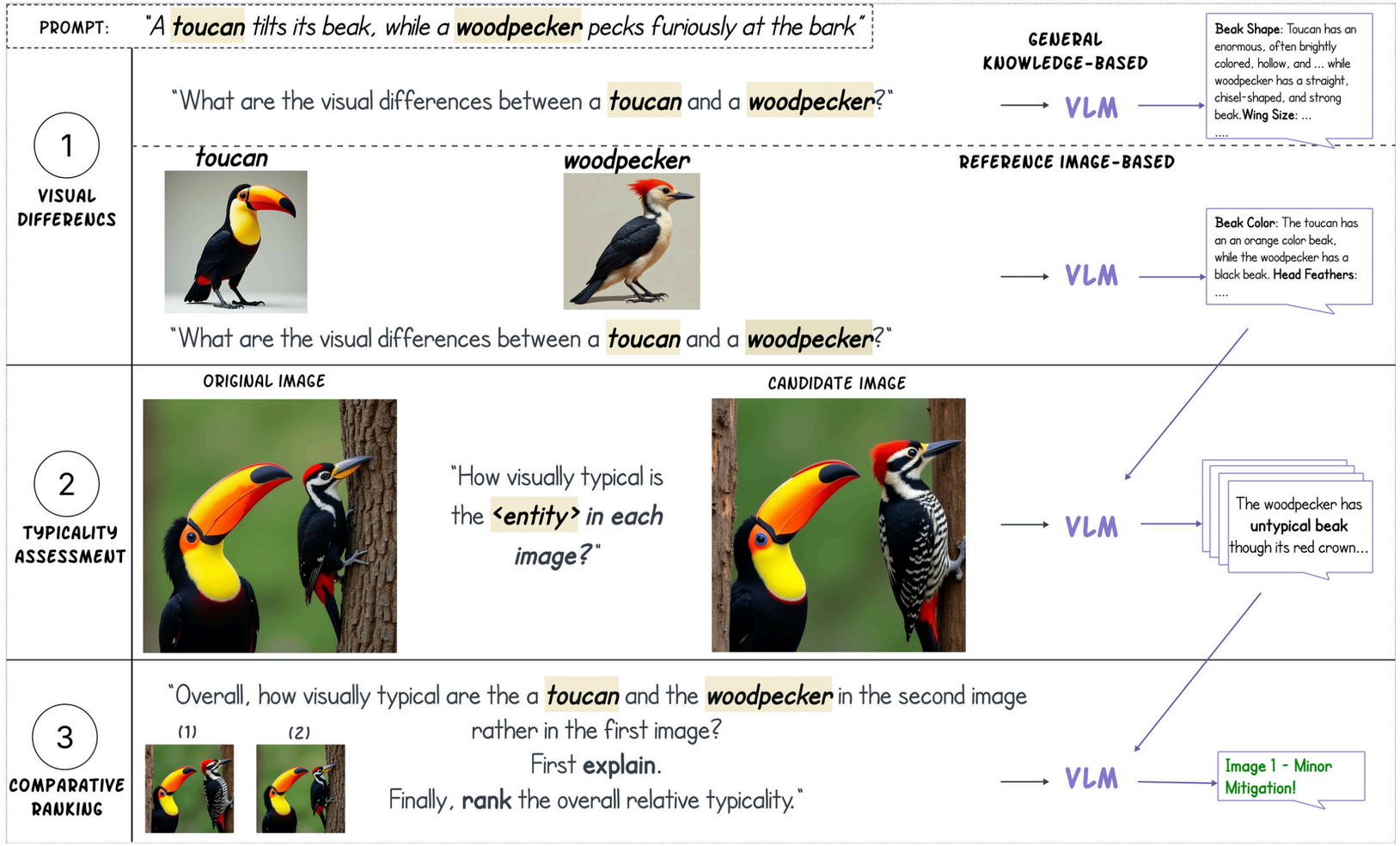
Automatic Evaluation



Automatic Evaluation



Automatic Evaluation

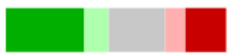


Results

Layout-based

Prompt-based

DeLeaker

Method	Semantic Leakage (Automatic)					
	Visualization	Mitigation ↑			Degradation ↓	
		Major	Minor	*No Change*	Minor	Major
RAG-Diffusion		17.55%	4.17%	5.03%	8.34%	64.91%
RPF		20.74%	9.06%	16.57%	15.26%	38.38%
3DIS		29.08%	8.10%	7.63%	10.13%	45.05%
QwenFLUX		17.28%	7.51%	15.85%	12.75%	46.60%
Instruction Prompt		23.92%	11.54%	35.35%	9.28%	19.88%
Entity Description Prompt		35.60%	11.07%	25.71%	9.17%	18.45%
DeLeaker		46.07%	9.76%	25.36%	5.83%	12.98%
DeLeaker + Description		53.57%	8.57%	15.95%	6.55%	15.36%

Results

Layout-based

Prompt-based

DeLeaker

Method	Visualization	Semantic Leakage (Automatic)				
		Mitigation ↑			Degradation ↓	
		Major	Minor	*No Change*	Minor	Major
RAG-Diffusion		17.55%	4.17%	5.03%	8.34%	64.91%
RPF		20.74%	9.06%	16.57%	15.26%	38.38%
3DIS		29.08%	8.10%	7.63%	10.13%	45.05%
QwenFLUX		17.28%	7.51%	15.85%	12.75%	46.60%
Instruction Prompt		23.92%	11.54%	35.35%	9.28%	19.88%
Entity Description Prompt		35.60%	11.07%	25.71%	9.17%	18.45%
DeLeaker		46.07%	9.76%	25.36%	5.83%	12.98%
DeLeaker + Description		53.57%	8.57%	15.95%	6.55%	15.36%

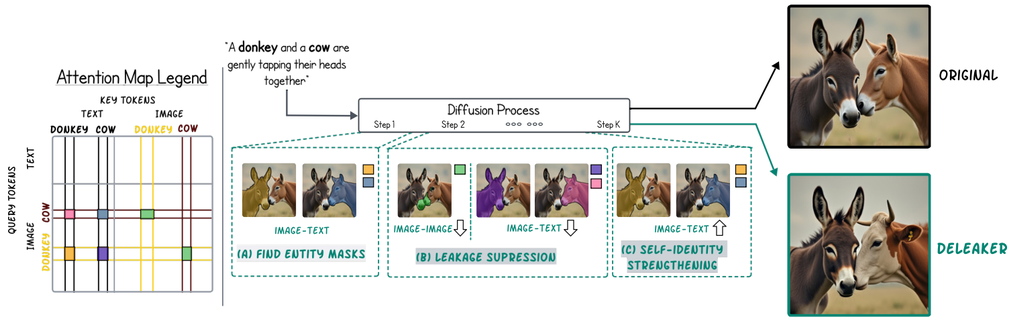
Results

	ORIGINAL	DELEAKER (OURS)	ENTITY DESC. PROMPT	IMAGE CONDITION PROMPT	RAG-DIFFUSION	3DIS	RPF
In a pop art style, A **hippopotamus** and a **rhinoceros** are one gently pushing the other with its nose.							
A **zebra** and a **horse** are rolling in the sand together, covering themselves in dust.							

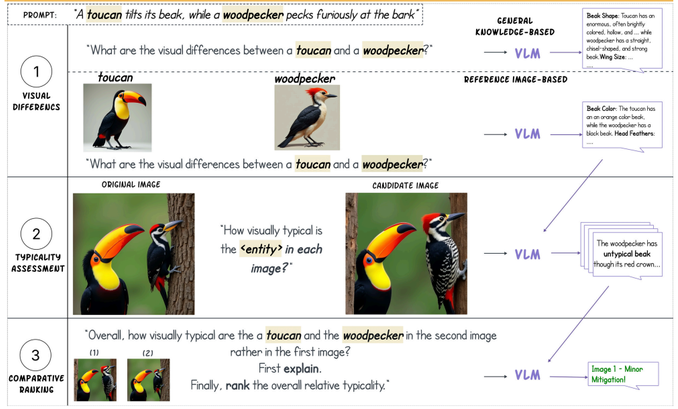
ICLR 2026

DeLeaker

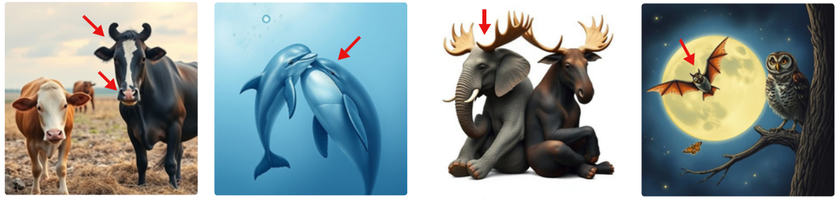
Method: DeLeaker



Automatic Evaluation



Semantic Leakage in IMages Dataset



Pairs | Interactions | Interactions with *style* | Multiple Entities (animals)

Dynamic Inference-Time Reweighting For Semantic Leakage Mitigation in Text-to-Image Models