Università di Catania

PeRCeiVe.AI Lab

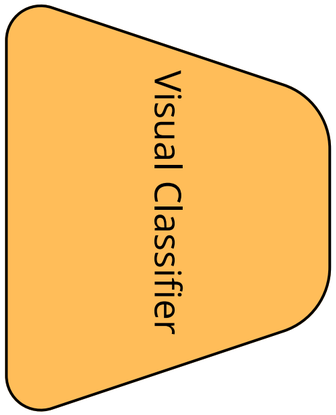

DEXTER

Diffusion-guided EXplanation through TExtual Reasoning

Simone Carnemolla, Matteo Pennisi, Sarinda Dhanesh Samarasinghe, Giovanni Bellitto, Simone Palazzo, Daniela Giordano, Mubarak Shah, Concetto Spampinato

Motivation

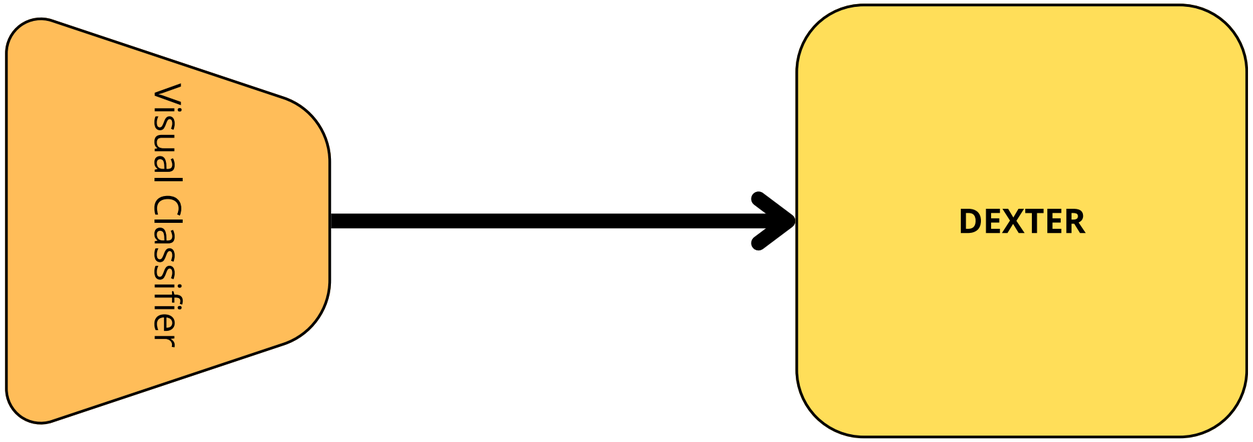
Is it possible to explain a visual classifier decision-making process with text?



1. Investigating model reasoning for **“Howler Monkey”** classification

Motivation

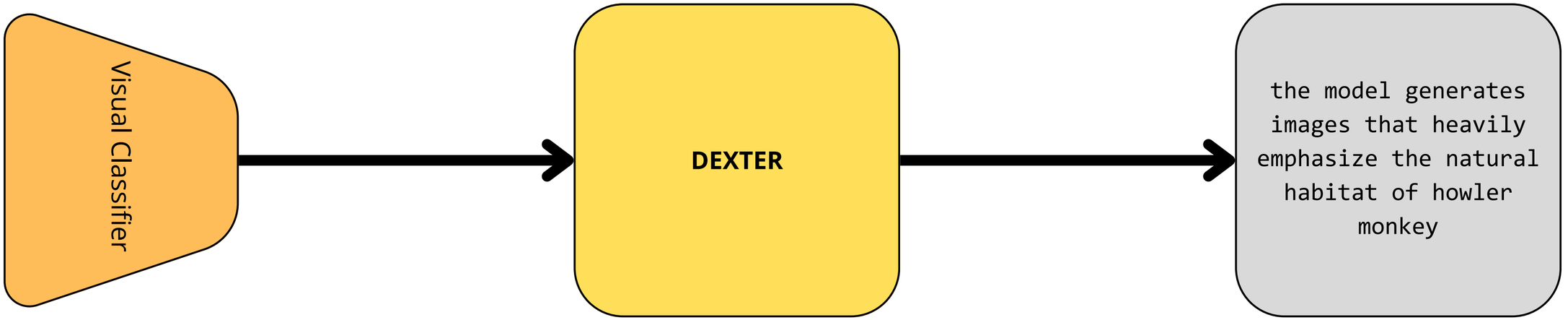
Is possible to explain a visual classifier decision-making process with text?



1. Investigating model reasoning for **“Howler Monkey”** classification
2. DEXTER optimization process

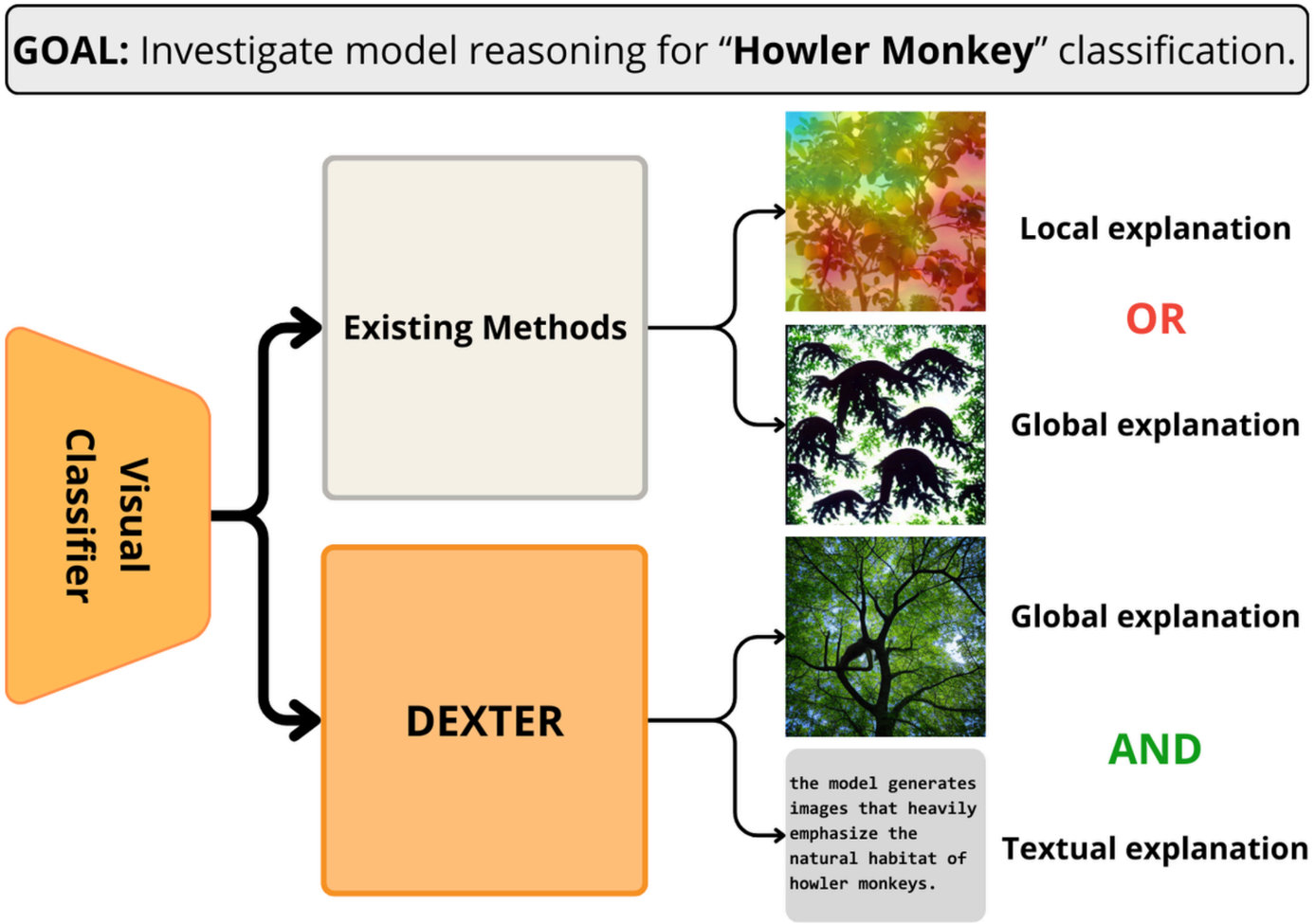
Motivation

Is possible to explain a visual classifier decision-making process with text?

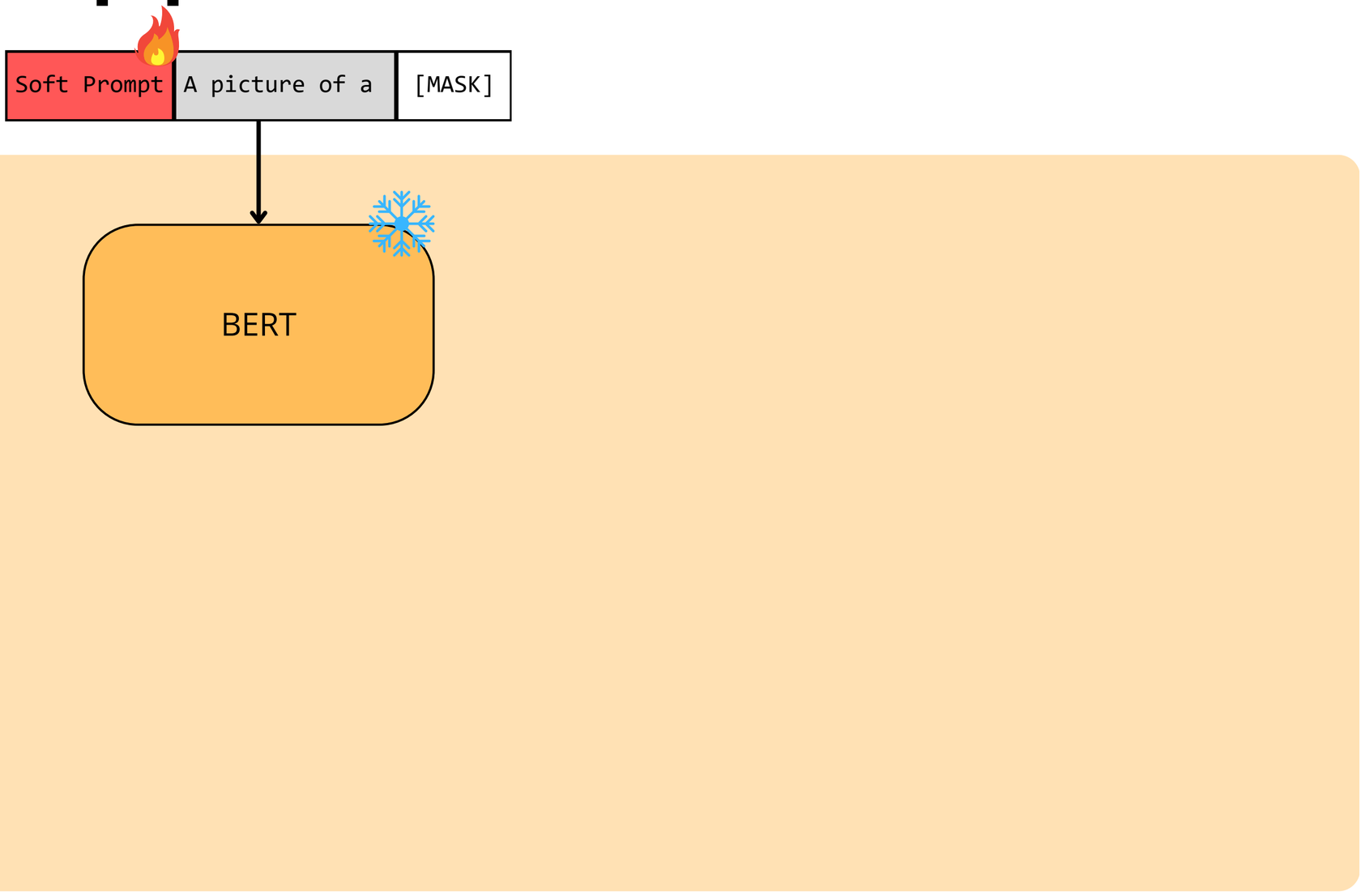


1. Investigating model reasoning for **“Howler Monkey”** classification
2. DEXTER optimization process
3. Textual explanation of the visual classifier behaviour

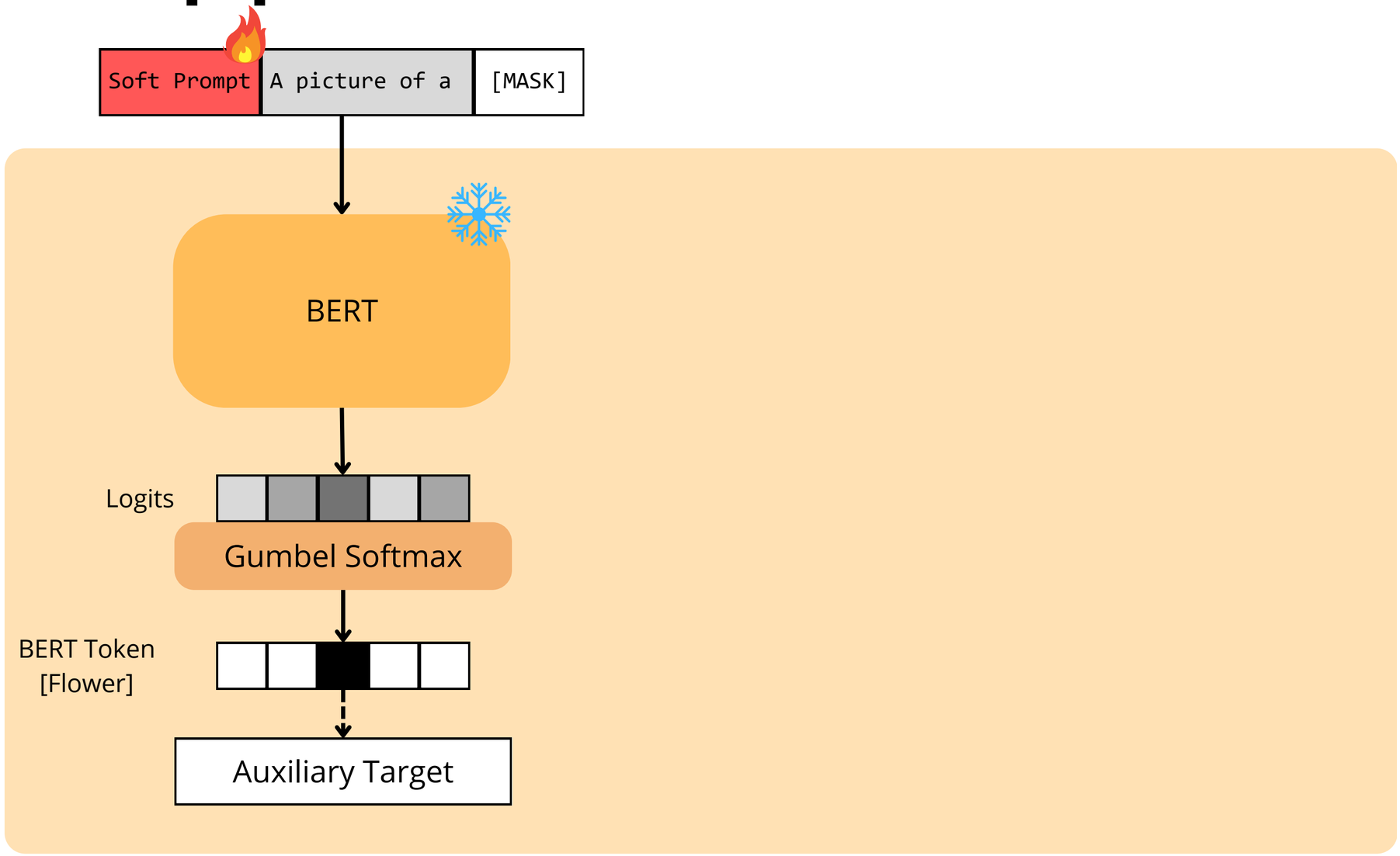
DEXTER and state-of-the-art



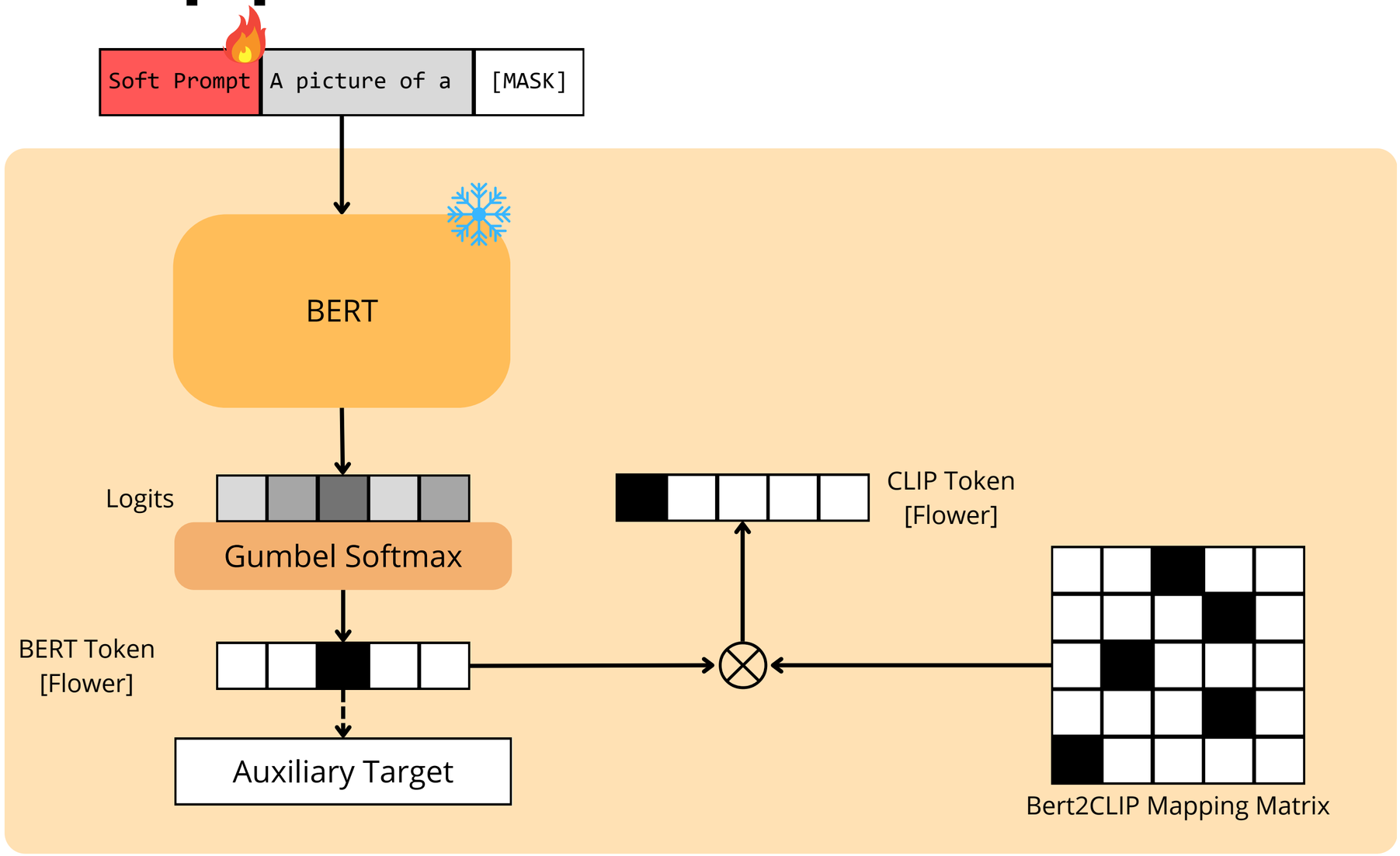
Text pipeline



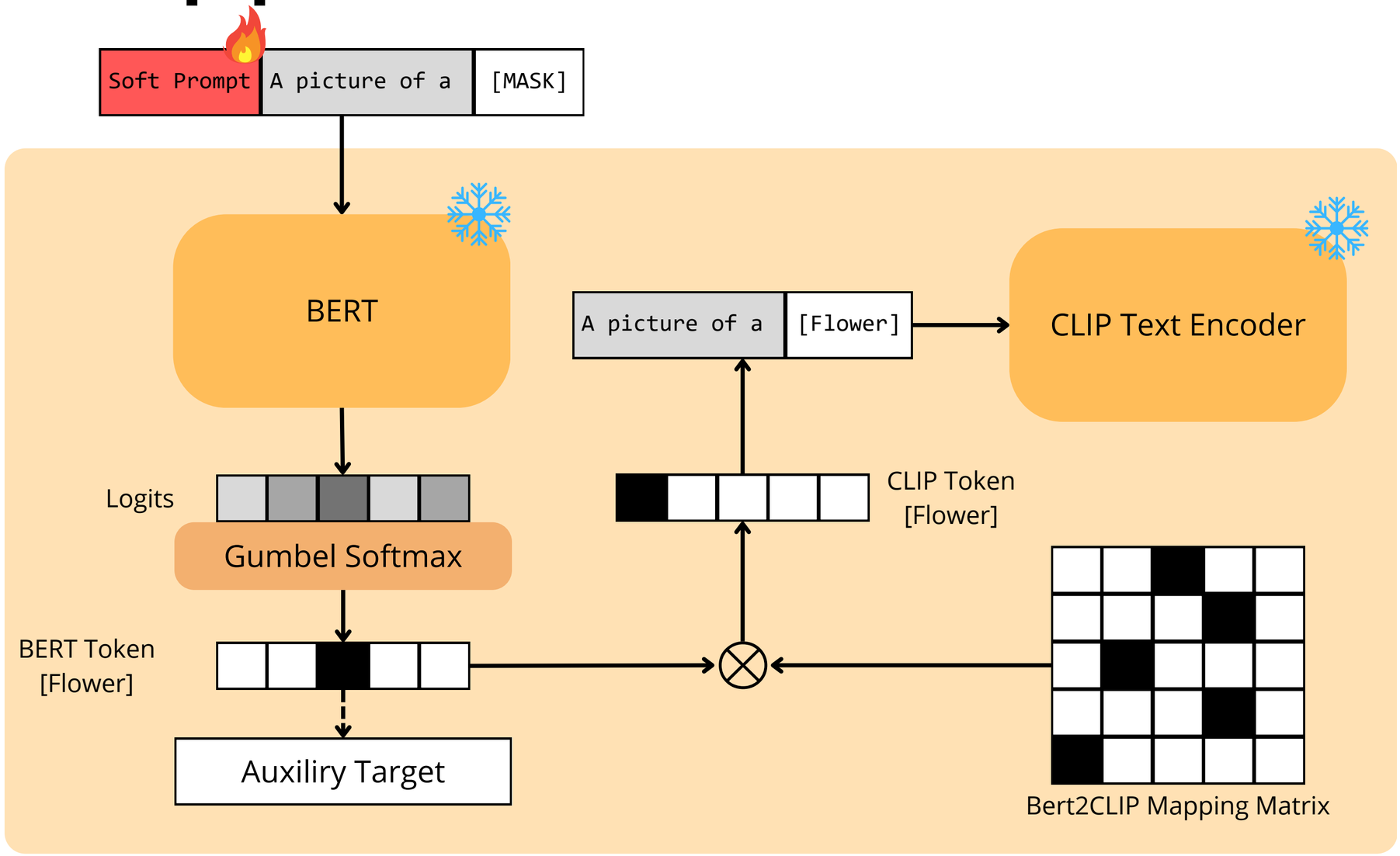
Text pipeline



Text pipeline



Text pipeline



Text pipeline

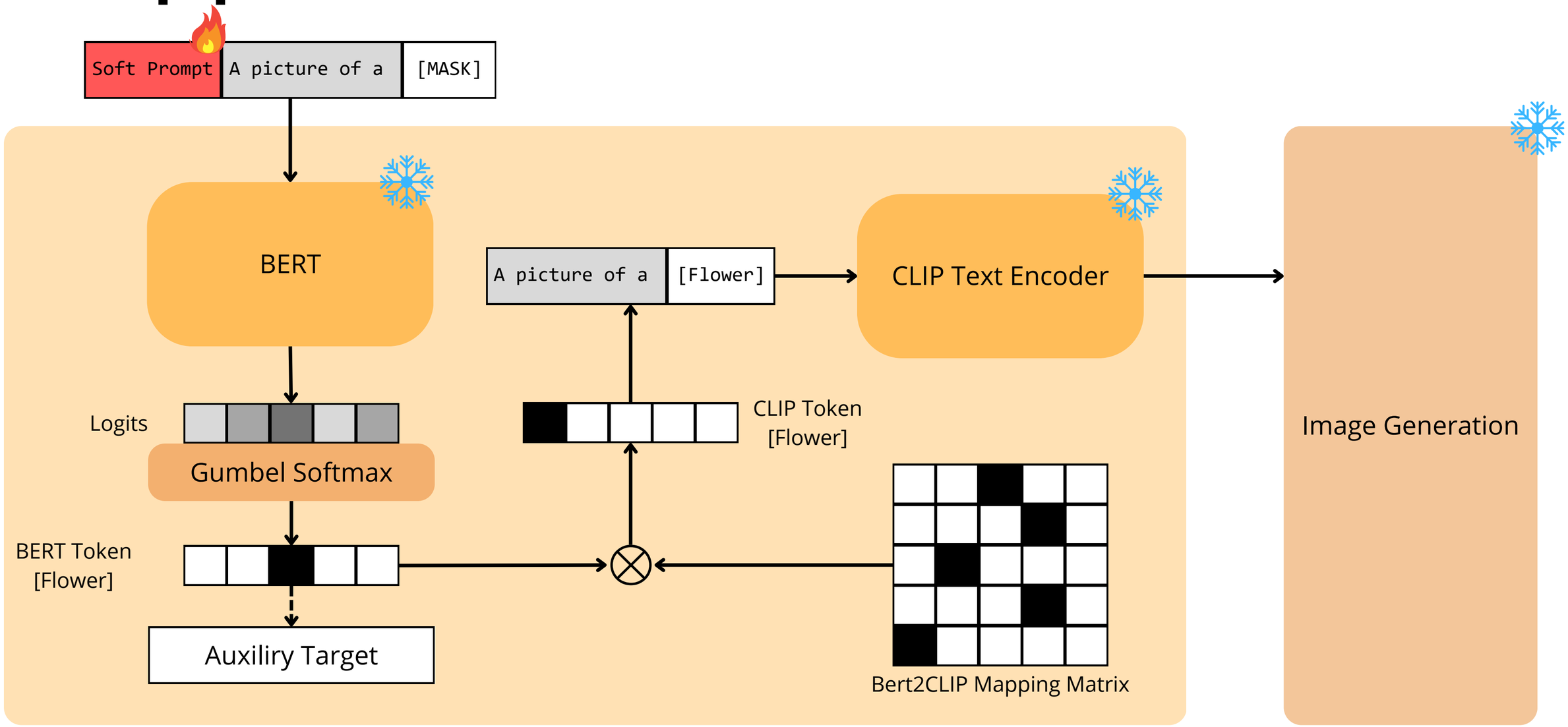


Image Generation

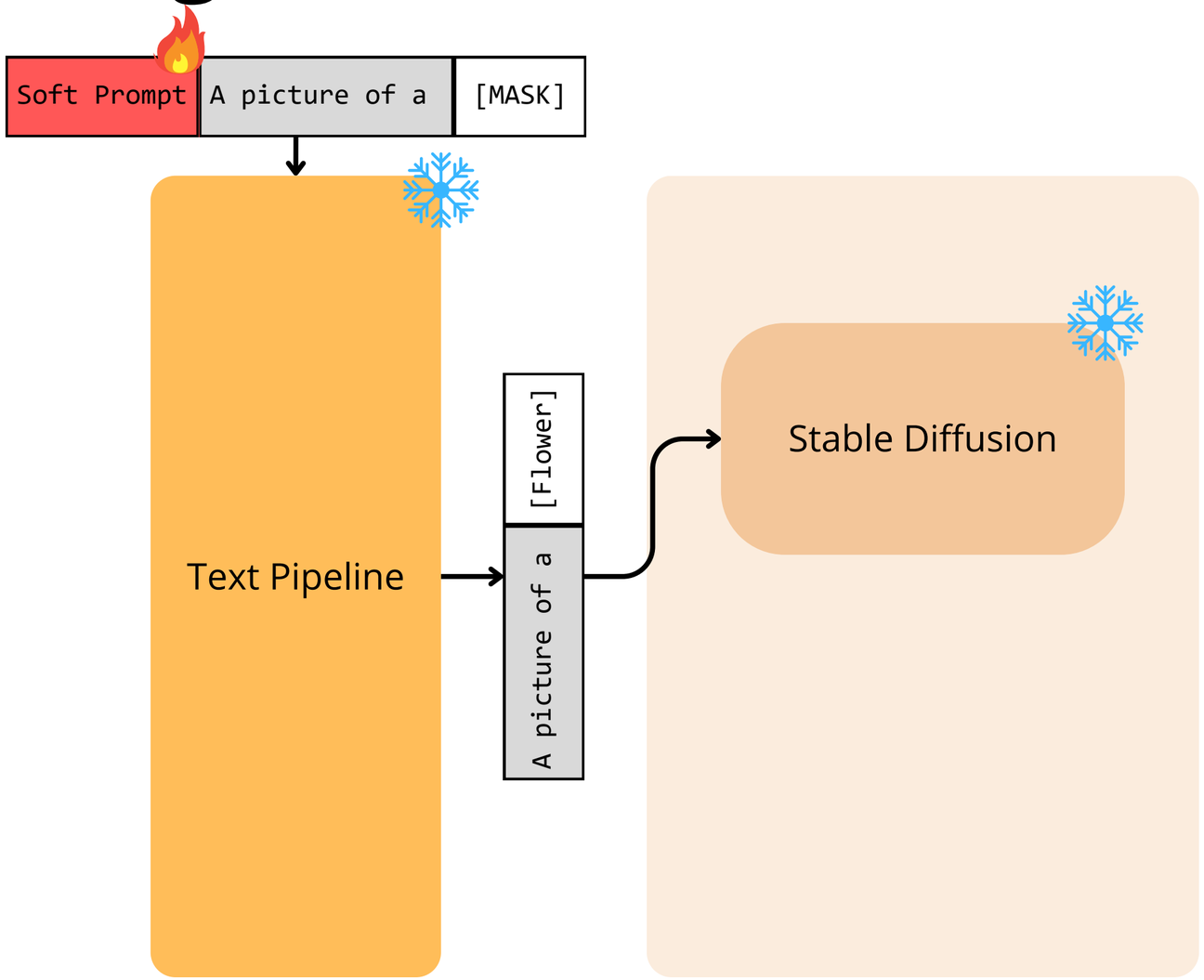


Image Generation

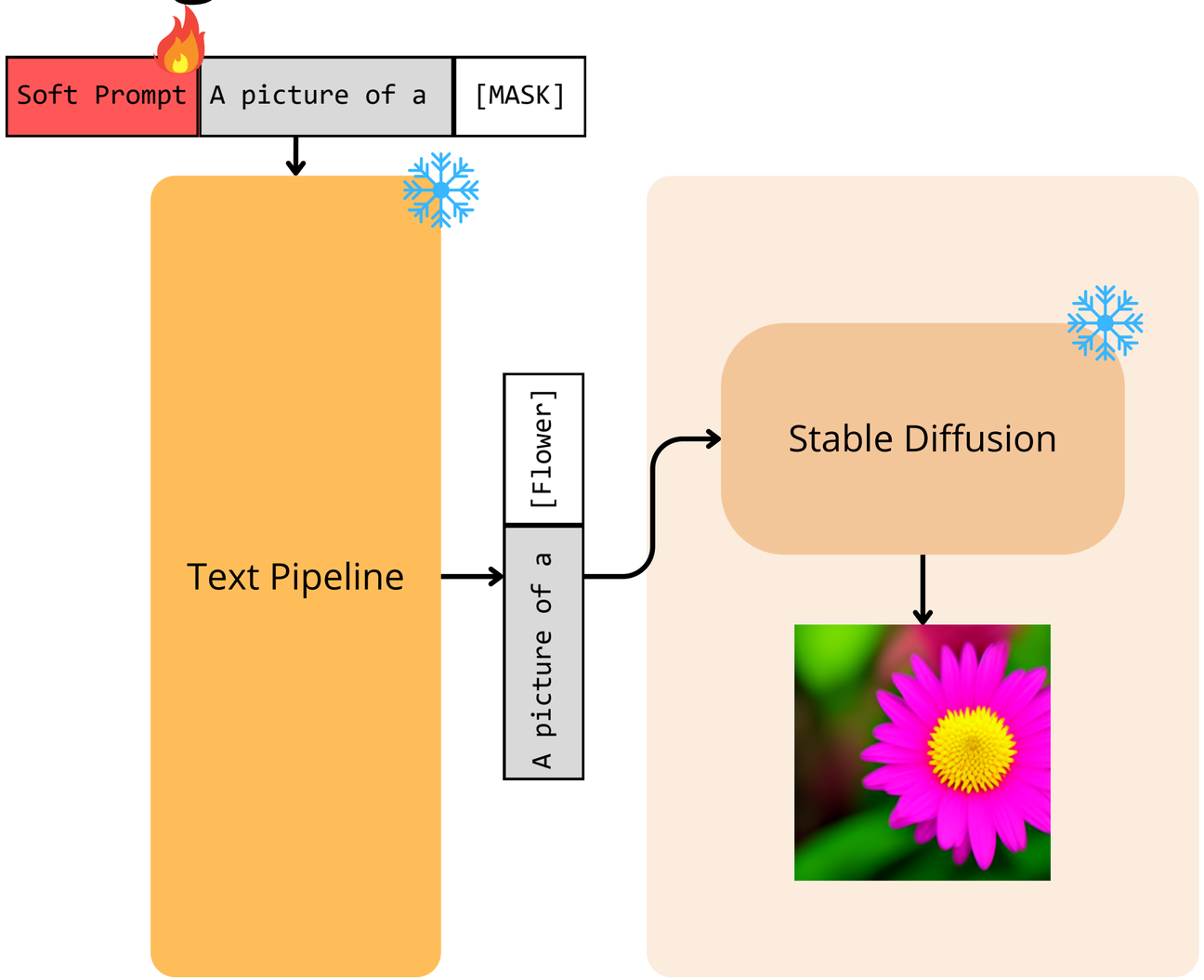


Image Generation

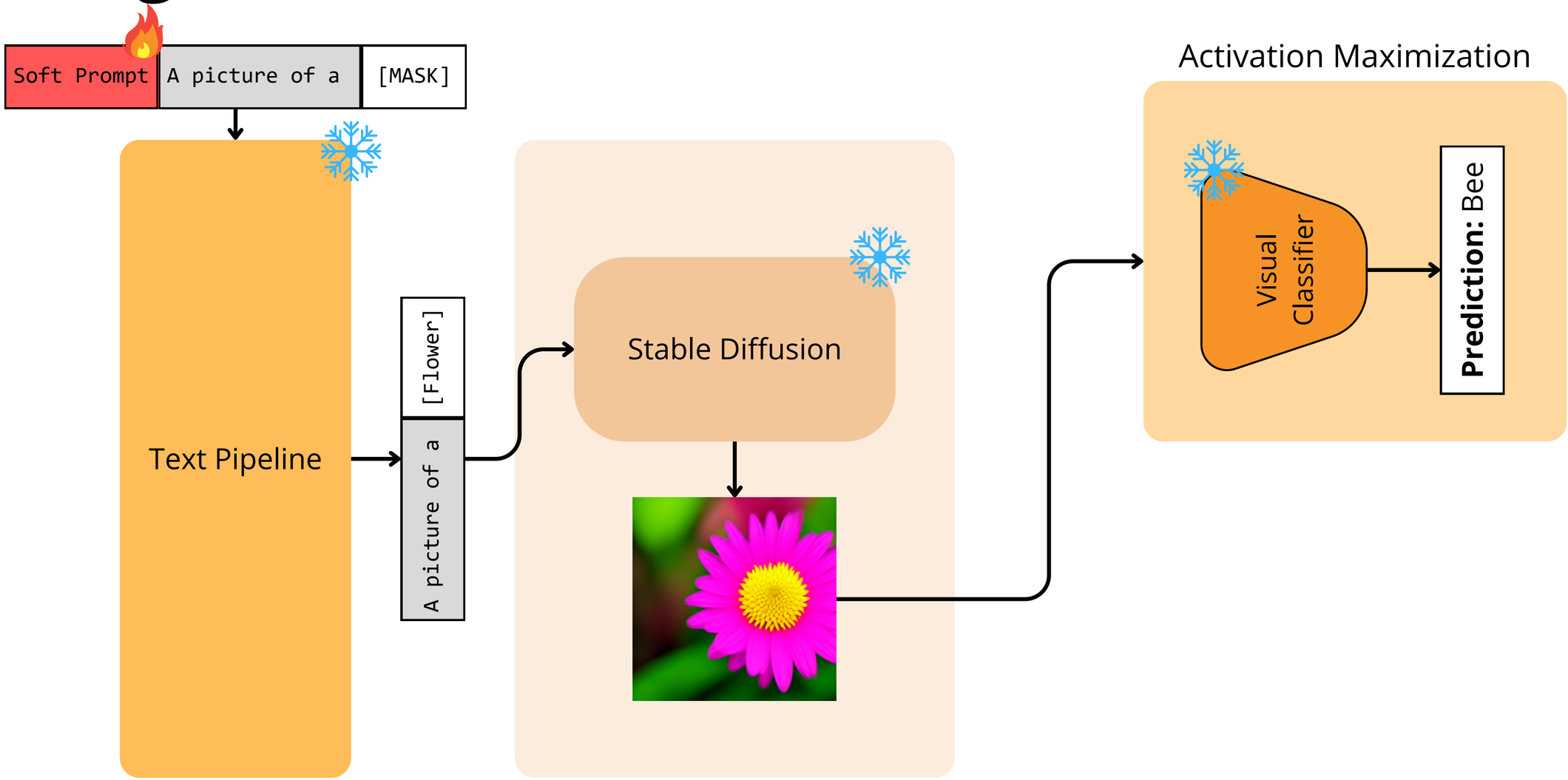
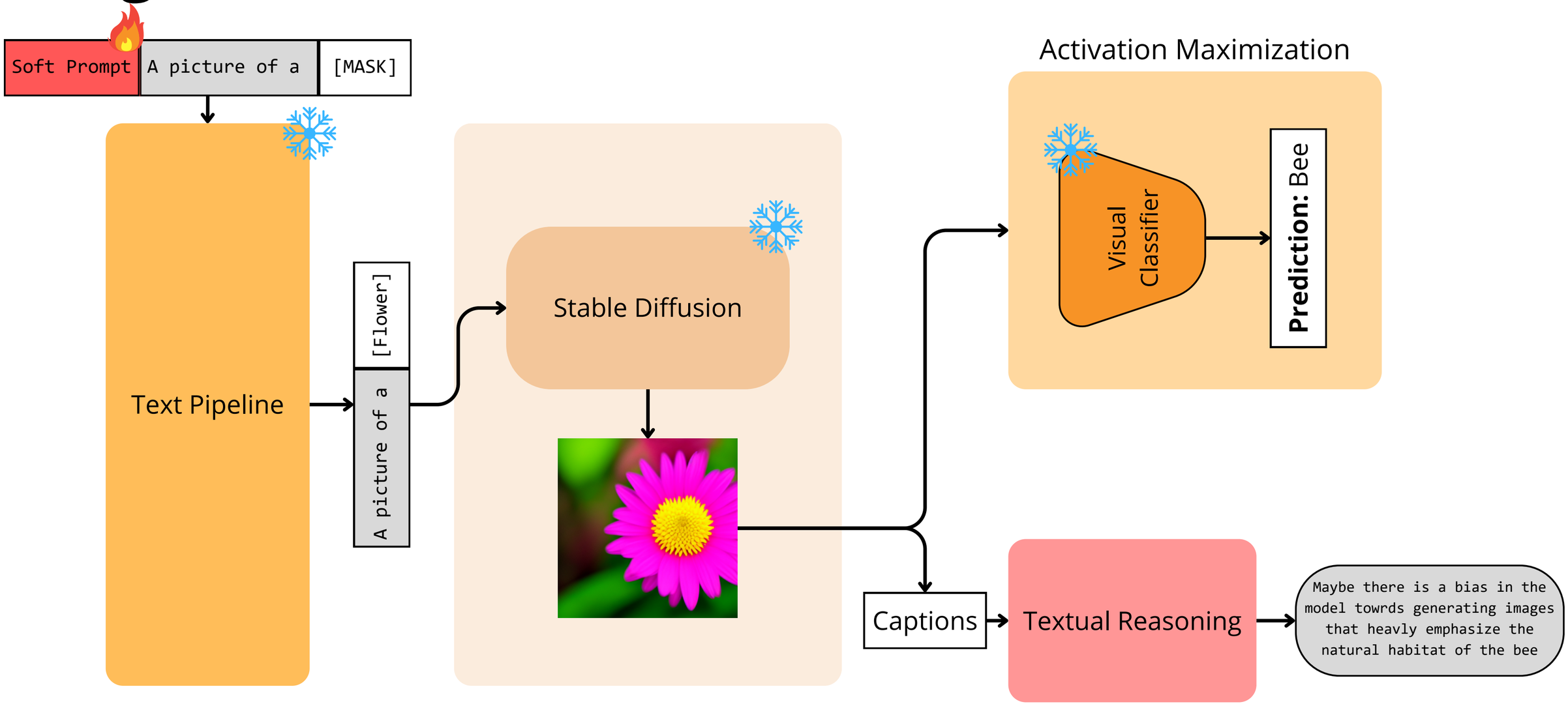


Image Generation



Auxiliary Target

1st step: Initialize a Pseudo-target and a best loss variable:

Auxiliary Target

1st step: Initialize a Pseudo-target and a best loss variable:

- random word from BERT Vocabulary `[Human]`

Auxiliary Target

1st step: Initialize a Pseudo-target and a best loss variable:

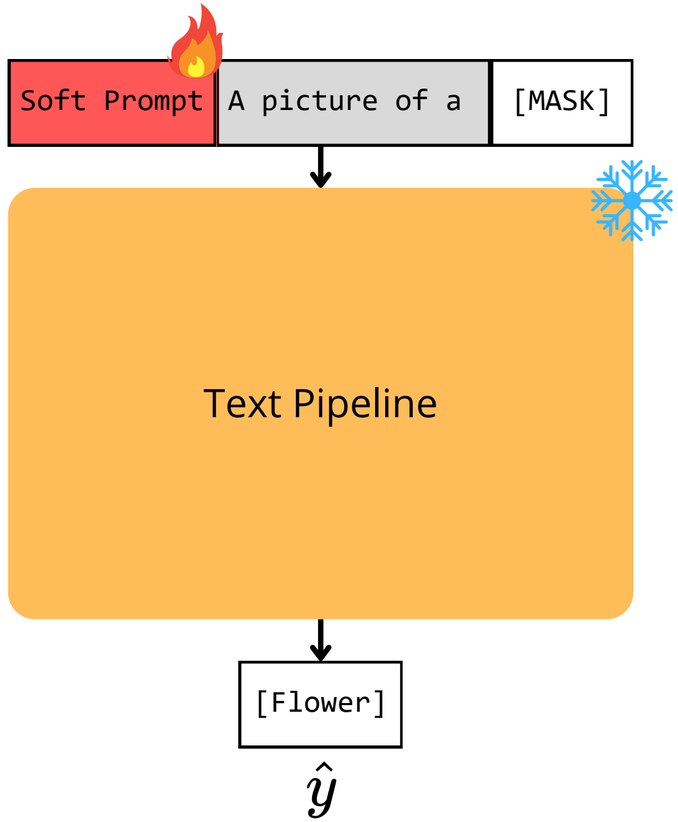
- random word from BERT Vocabulary `[Human]`
- $\text{best loss} = +\infty$

Auxiliary Target

2nd step: Compute the masking loss

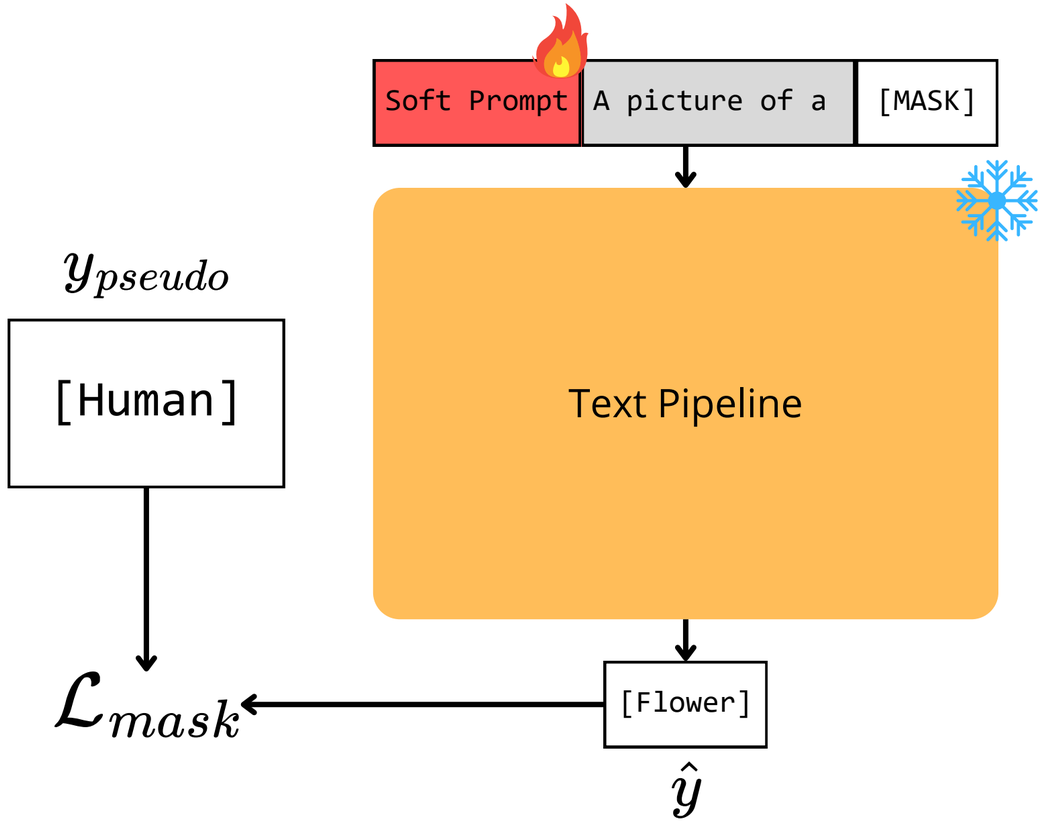
Auxiliary Target

2nd step: Compute the masking loss



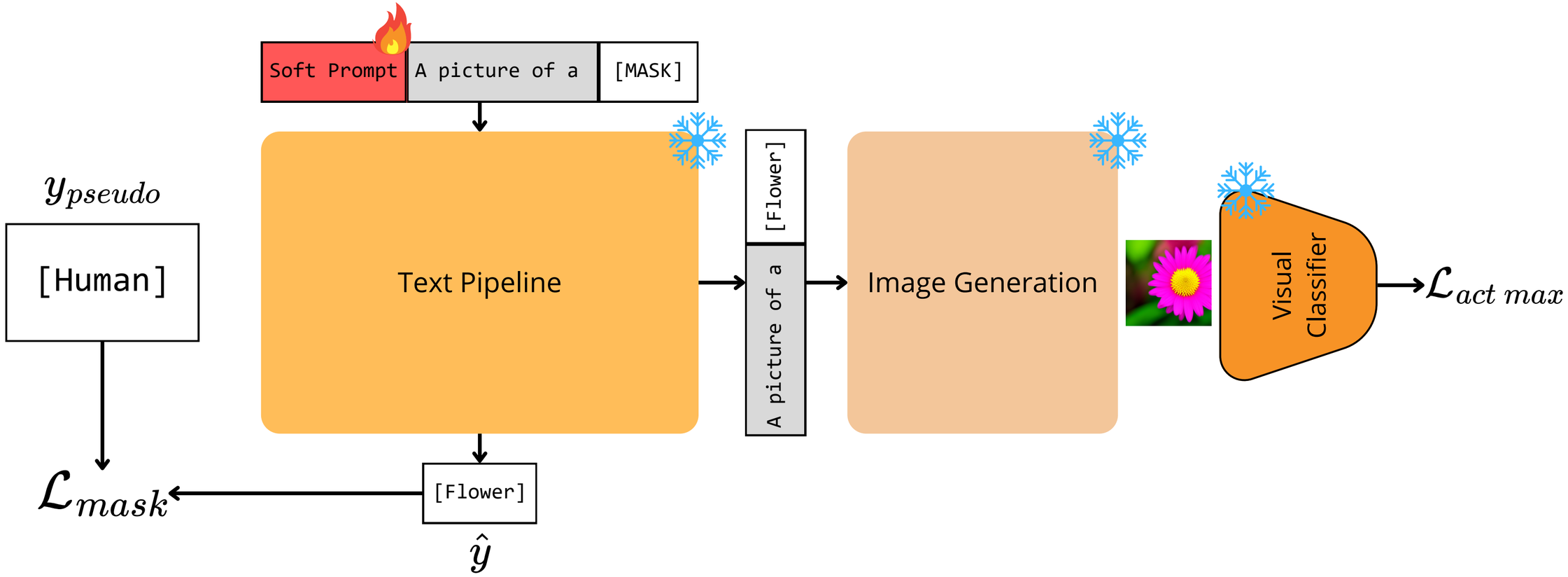
Auxiliary Target

2nd step: Compute the masking loss



Auxiliary Target

3rd step: Complete the forward



Auxiliary Target

4th step: Update the best loss variable and the pseudo target

Auxiliary Target

4th step: Update the best loss variable and the pseudo target

$if\ \mathcal{L}_{act\ max} < \text{best loss}:$

Auxiliary Target

4th step: Update the best loss variable and the pseudo target

$if\ \mathcal{L}_{act\ max} < \text{best loss}:$

- $\implies \text{best loss} = \mathcal{L}_{act\ max}$

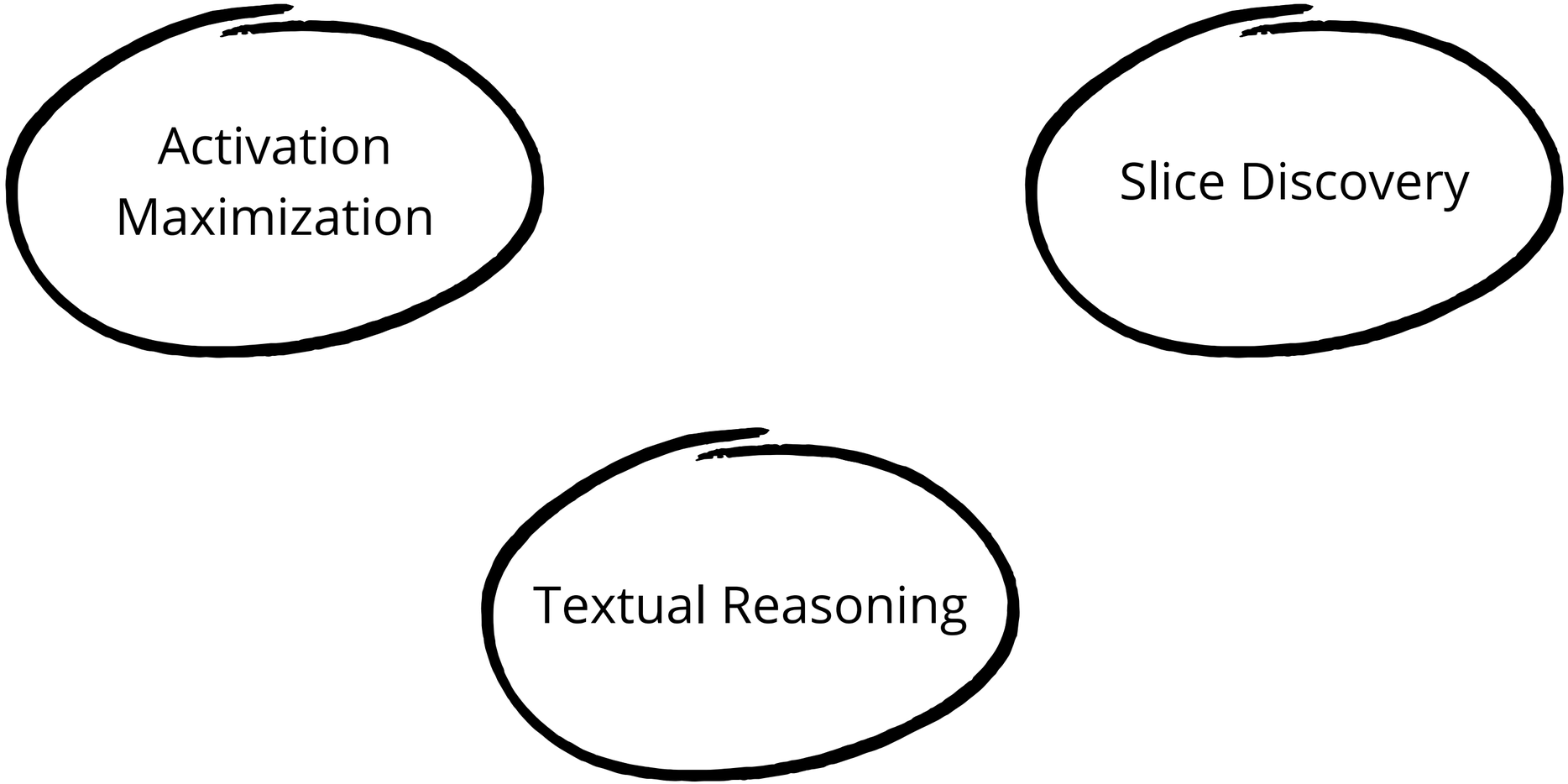
Auxiliary Target

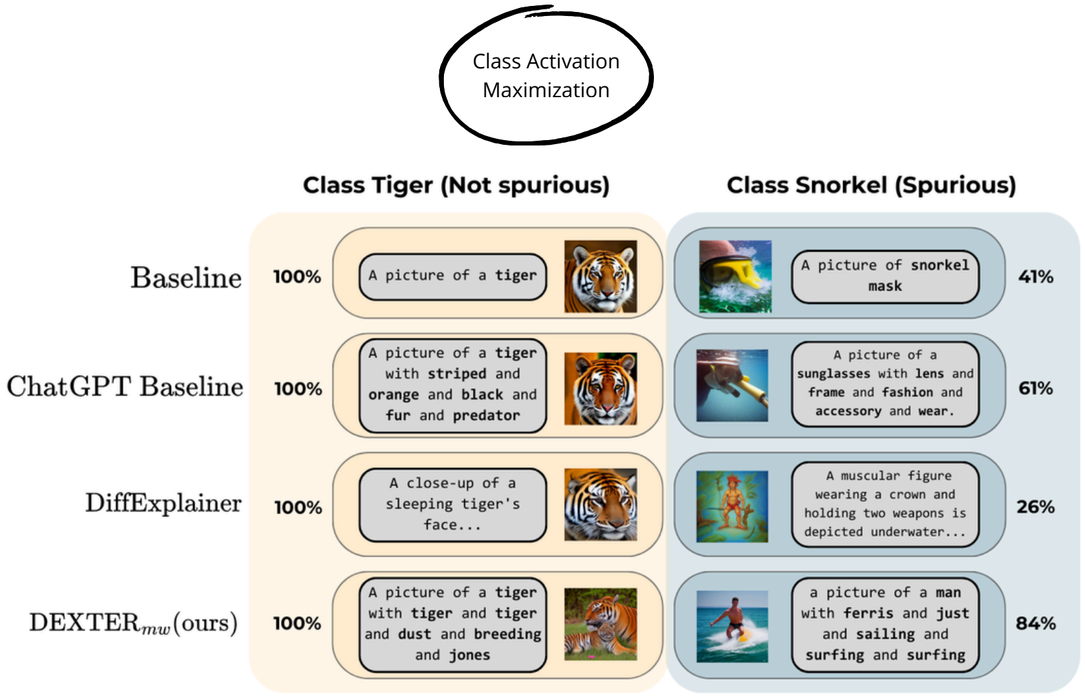
4th step: Update the best loss variable and the pseudo target

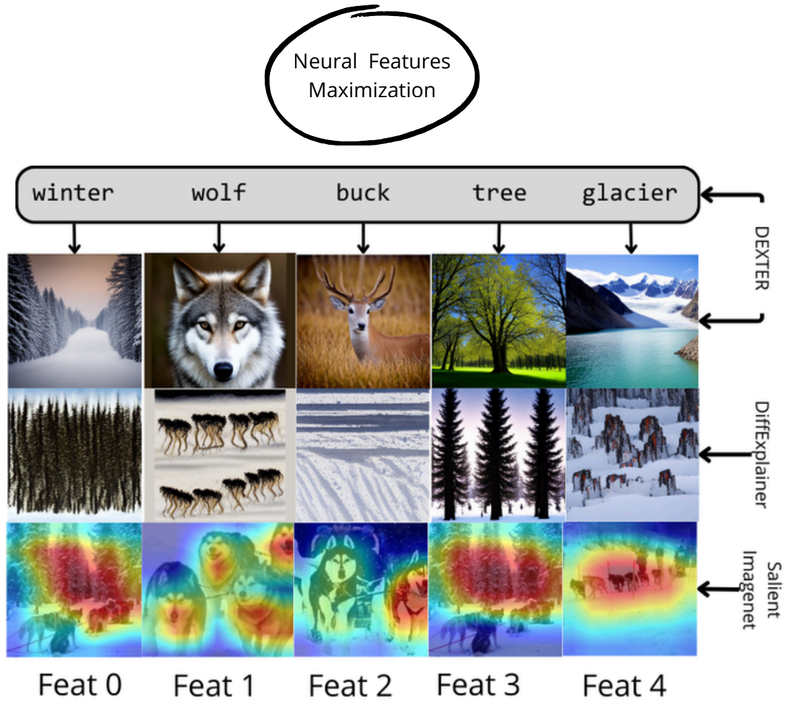
$if\ \mathcal{L}_{act\ max} < \text{best loss}:$

- $\implies \text{best loss} = \mathcal{L}_{act\ max}$
- $\implies y_{pseudo} = \hat{y}$

Evaluation

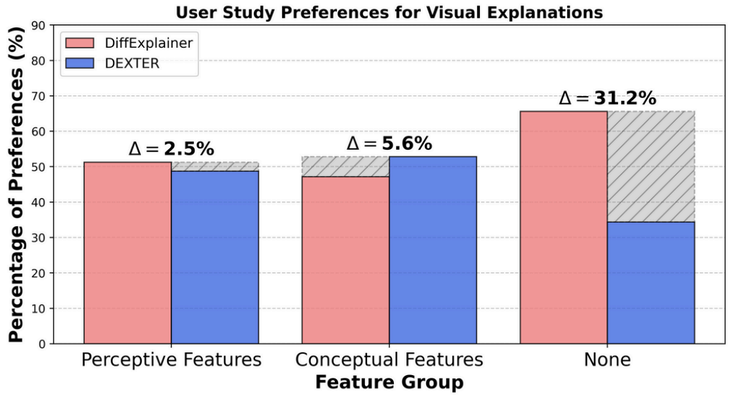






Method	Spurious	Core	Mean
Baseline	35.33	86.40	60.87
ChatGPT description	41.20	78.53	59.87
DiffExplainer	33.20	47.66	39.83
DEXTER (ours)	**63.00**	**87.86**	**75.43**

Comparison between text-prompting strategies



User study results on SalientImagenet

Slice Discovery

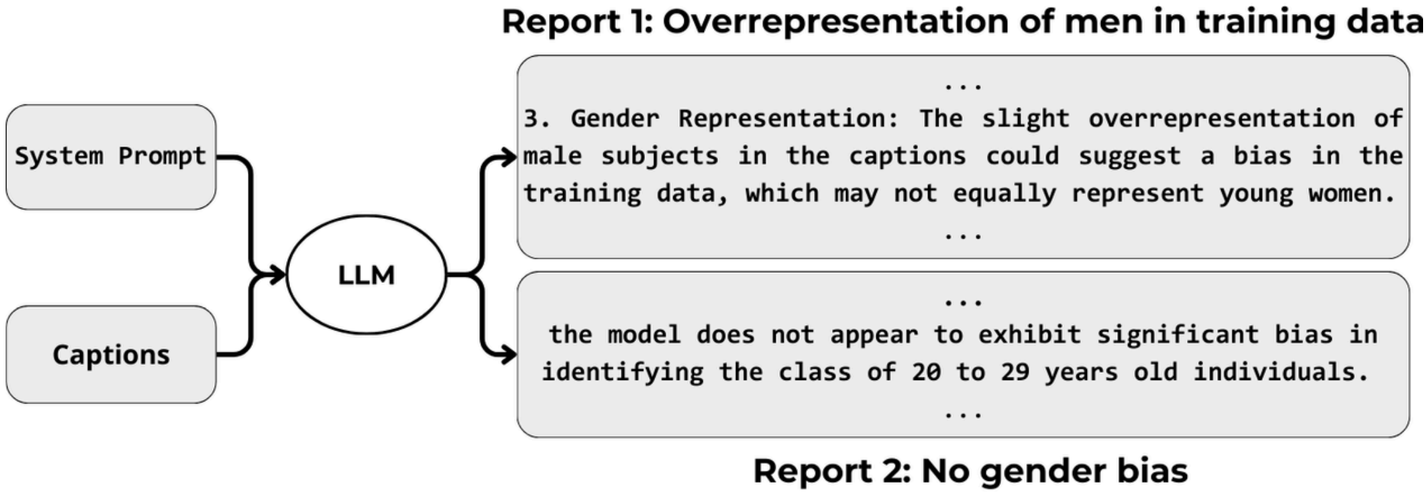
Dataset	Class 0	Class 1
Waterbirds	fence, woods, jungle, backyard	seas, sea, lake, harbor
CelebA	—	woman, person, head, girl

Class-wise words discovered by DEXTER for Waterbirds and CelebA datasets

Method	Data	GT	CelebA		Waterbirds	
			Worst	Avg.	Worst	Avg.
ERM	✓	-	47.7 ± 2.1	94.9	62.6 ± 0.3	97.3
LfF	✓	-	77.2	85.1	78.0	91.2
GEORGE	✓	-	54.9 ± 1.9	94.6	76.2 ± 2.0	95.7
JTT	✓	-	81.5 ± 1.7	88.1	83.8 ± 1.2	89.3
CNC	✓	-	88.8 ± 0.9	89.9	88.5 ± 0.3	90.9
DRO	✓	✓	90.0 ± 1.5	93.3	89.9 ± 1.3	91.5
LADDER	✓	-	89.2 ± 0.4	89.8	$\mathbf{92.4 \pm 0.8}$	93.1
DRO-B2T	✓	-	$\underline{90.4 \pm 0.9}$	93.2	$\underline{90.7 \pm 0.3}$	92.1
Ours	-	-	$\mathbf{91.3 \pm 0.01}$	91.7	90.5 ± 0.1	92.0

DEXTER performance on slice discovery and debiasing

Textual Reasoning



DEXTER explanation reports.

	Class 0 (20-29)		Class 1 (50-59)		Mean
Metric	w Bias	w/o Bias	w Bias	w/o Bias	
STS	0.92	0.85	0.91	0.91	0.90
G-eval$_{con}$	4.58 ± 1.00	4.80 ± 0.47	3.66 ± 0.84	3.68 ± 0.85	4.19 ± 0.52
MOS$_{LLM}$	4.29 ± 0.63	4.80 ± 0.37	4.63 ± 0.44	4.19 ± 0.74	4.48 ± 0.25
MOS$_{humans}$	4.20 ± 0.63	3.89 ± 0.64	4.10 ± 0.67	3.88 ± 0.79	4.01 ± 0.69

Evaluation of DEXTER bias reports

Thank you
for the attention!