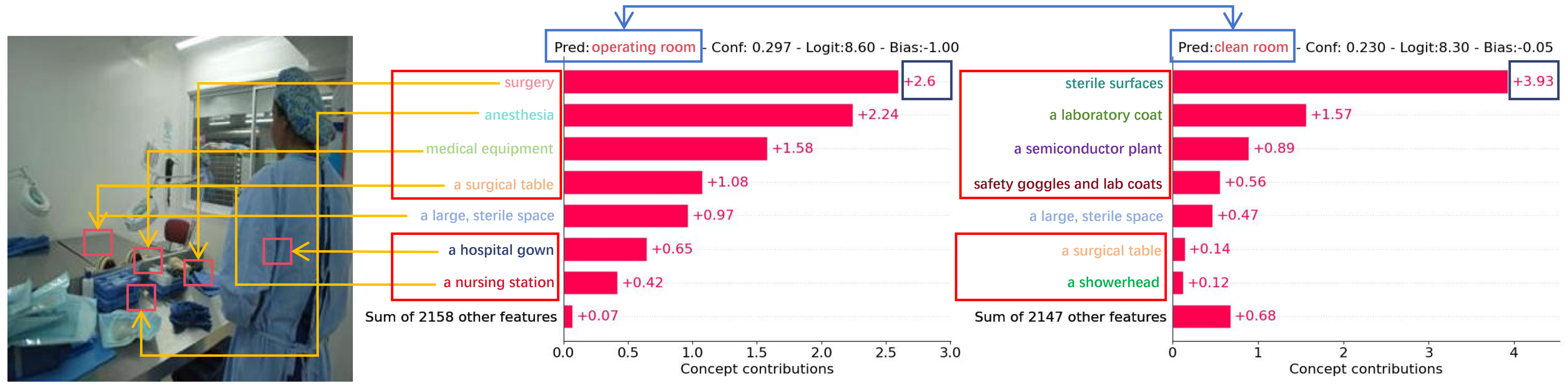
Faithful Vision-Language Interpretation via Concept Bottleneck Models

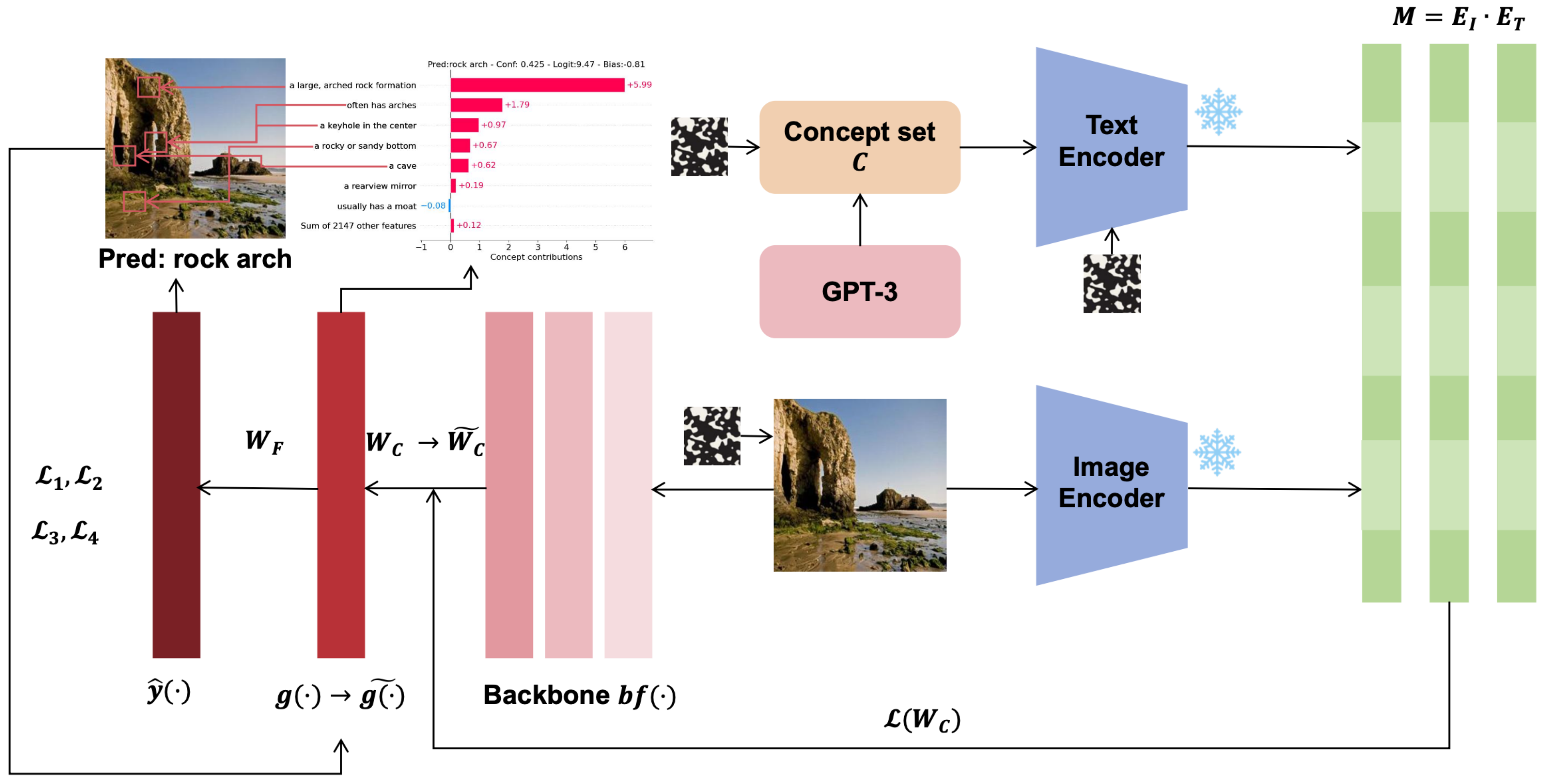
Songning Lai*, Lijie Hu*, Junxiao Wang, Laure Berti-Equille, and Di Wang

Introduction

Traditional Concept Bottleneck Models (CBMs) require substantial manual annotation, which label-free CBM effectively addresses by leveraging factual information from pre-trained models. However, this convenience comes with inherent instability in pre-trained models. We addressed it in this paper.



Definition and Framework


$M = E_I \cdot E_T$
Pred:rock arch - Conf: 0.425 - Logit:9.47 - Bias:-0.81
a large, arched rock formation +5.99
often has arches +1.79
a keyhole in the center +0.97
a rocky or sandy bottom +0.67
a cave +0.62
a rearview mirror +0.19
usually has a moat −0.08
Sum of 2147 other features +0.12
Concept contributions
Pred: rock arch
Concept set C
GPT-3
Text Encoder
Image Encoder
W_F
$W_C \to \widetilde{W_C}$
$\mathcal{L}_1, \mathcal{L}_2$
$\mathcal{L}_3, \mathcal{L}_4$
$\hat{y}(\cdot)$
$g(\cdot) \to \widetilde{g(\cdot)}$
Backbone $bf(\cdot)$
$\mathcal{L}(W_C)$

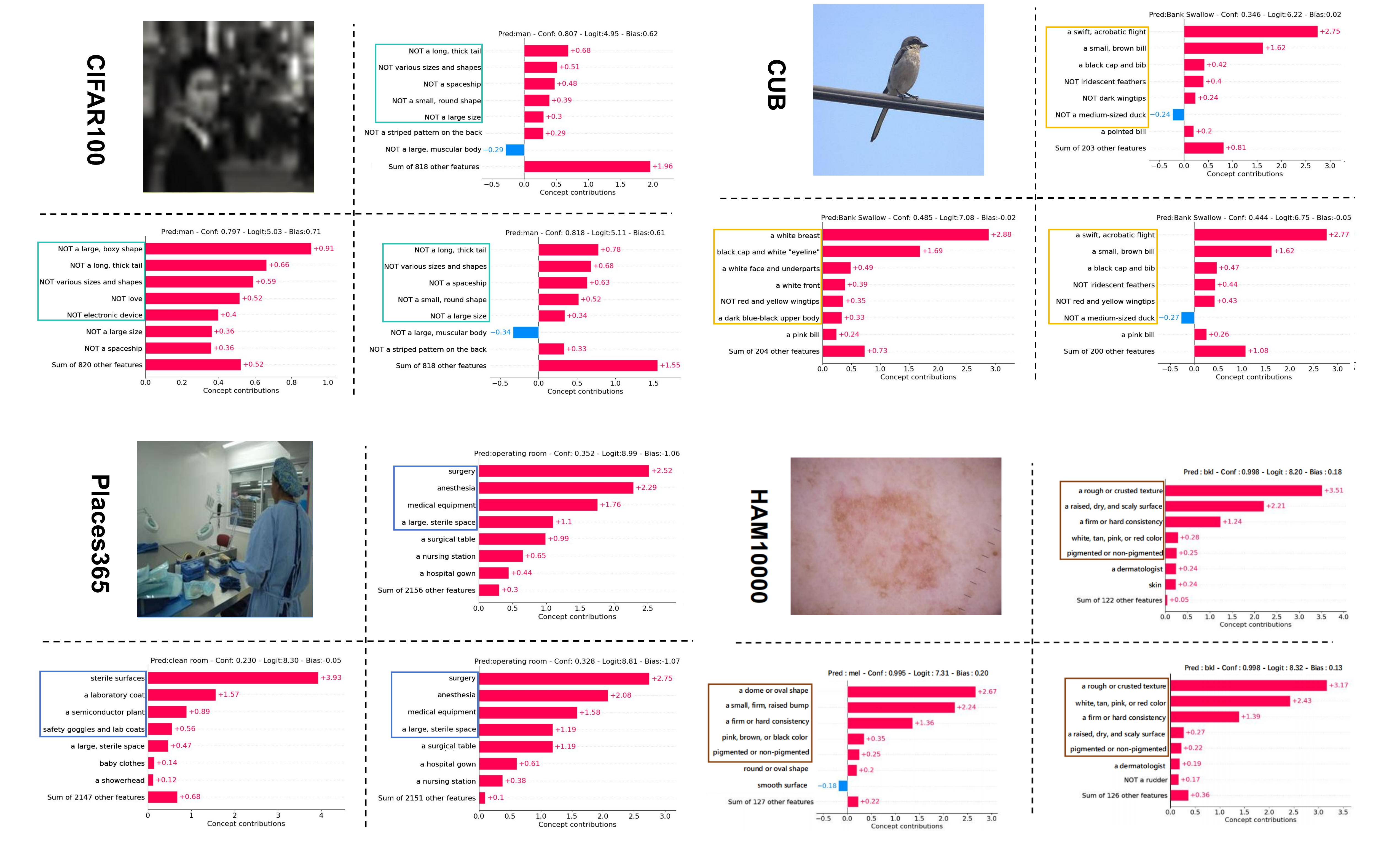
$$\min_{\tilde{W}_c} \mathbb{E}_x [\lambda_1 \underbrace{D(y(x,\tilde{\boldsymbol{c}}), y(x,\boldsymbol{c}))}_{\mathcal{L}_1} + \lambda_2 \underbrace{\mathcal{L}_{k_1}(\tilde{\boldsymbol{g}}(x), \boldsymbol{g}(x))}_{\mathcal{L}_2} + \lambda_3 \underbrace{\max_{||\delta|| \leq R_2} D(y(x,\tilde{\boldsymbol{c}}), y(x,\tilde{\boldsymbol{c}}+\boldsymbol{\delta}))}_{\mathcal{L}_3} + \lambda_4 \underbrace{\max_{||\rho|| \leq R_1} \mathcal{L}_{k_2}(\tilde{\boldsymbol{g}}(x), \tilde{\boldsymbol{g}}(x)+\boldsymbol{\rho})}_{\mathcal{L}_4}] .$$

Definition 1 (Faithful Vision-Language Concept). *Under the same concept space, i.e., under the set of concepts generated by GPT3 at one time, we call a matrix $\tilde{W}_c$ is a $(D, R, \alpha, \beta, k_1, k_2)$-Faithful Vision-Language Concept (FVLC) model for the vanilla concept if it satisfies for any input x:*

- *(Similarity of Explanation) $V_{k_1}(\tilde{\boldsymbol{g}}(x), \boldsymbol{g}(x)) \geq \beta_1$ for some $1 \geq \beta_1 \geq 0$;*
- *(Stability of Explanation) $V_{k_2}(\tilde{\boldsymbol{g}}(x), \tilde{\boldsymbol{g}}(x)+\boldsymbol{\rho}) \geq \beta_2$ for some $1 \geq \beta_2 \geq 0$ and all $\|\boldsymbol{\rho}\| \leq R_1$, where $\|\cdot\|$ is a norm and $R_1 \geq 0$;*
- *(Closeness of Prediction) $D(y(x,\tilde{\boldsymbol{c}}), y(x,\boldsymbol{c})) \leq \alpha_1$ for some $\alpha_1 \geq 0$, where D is some probability distance or divergence;*
- *(Stability of Prediction) $D(y(x,\tilde{\boldsymbol{c}}), y(x,\tilde{\boldsymbol{c}}+\boldsymbol{\delta})) \leq \alpha_2$ for all $\|\boldsymbol{\delta}\| \leq R_2$, where D is some probability distance or divergence, $\|\cdot\|$ is a norm and $R_2 \geq 0$,*

where $\tilde{\boldsymbol{g}}(x) = \tilde{W}_c bf(x)$, $y(x,\boldsymbol{c}) = W_F g(x)$, and $y(x,\tilde{\boldsymbol{c}}) = W_F \tilde{\boldsymbol{g}}(x)$, $y(x,\tilde{\boldsymbol{c}}+\boldsymbol{\delta}) = W_F(\tilde{\boldsymbol{g}}(x)+\boldsymbol{\delta})$. For any given x, $\tilde{c} = \tilde{\boldsymbol{g}}(x)$ is a $(D, R, \alpha, \beta, k_1, k_2)$-FVLC. Here, $\alpha = \min\{\alpha_1, \alpha_2\}$, $\beta = \max\{\beta_1, \beta_2\}$, and $R = \min\{R_1, R_2\}$.

Evaluating Interpretability and Utility



Stability Evaluation

Method	CIFAR10	CIFAR100	CUB	Places365
Standard (No interpretability)	88.80%	70.10%	76.70%	48.56%
P-CBM (CLIP)	84.50%	56.00%	N/A	N/A
Label-free CBM	86.32%	65.42%	74.23%	43.63%
WP1(5%) - base	86.47%	65.13%	74.08%	43.57%
WP1(5%) - **FVLC**	86.34%	**65.43%**	73.96%	43.67%
WP1(10%) - base	86.25%	65.09%	73.97%	43.67%
WP1(10%) - **FVLC**	86.39%	64.90%	73.92%	43.62%
WP2 - base	86.41%	65.16%	73.96%	43.54%
WP2 - **FVLC**	86.22%	65.34%	**74.44%**	**44.55%**
IP - base	86.62%	65.36%	74.39%	43.64%
IP - **FVLC**	**86.88%**	65.29%	74.01%	43.71%
WP1(5%)+WP2 - base	86.49%	65.17%	73.90%	43.67%
WP1(5%)+WP2 - **FVLC**	86.43%	65.33%	73.92%	43.49%
WP1(10%)+WP2 - base	86.30%	64.87%	73.82%	43.61%
WP1(10%)+WP2 - **FVLC**	86.38%	65.06%	74.01%	43.44%
WP1(10%)+WP2+IP - base	85.96%	64.41%	73.74%	43.32%
WP1(10%)+WP2+IP - **FVLC**	86.70%	65.14%	74.36%	43.46%

Method	CIFAR10		CIFAR100		CUB		Places365	
	TCPC	TOPC	TCPC	TOPC	TCPC	TOPC	TCPC	TOPC
WP1(5%) - base	1.55E-01	6.32E-02	1.01E-01	7.17E-02	1.26E-01	1.85E-01	1.59E-01	6.40E-02
WP1(5%) - **FVLC**	**1.12E-03**	**8.55E-03**	**2.81E-03**	**4.51E-03**	**1.05E-02**	**1.50E-03**	**1.38E-03**	**1.30E-03**
WP1(10%) - base	1.99E-01	8.36E-02	1.94E-01	1.31E-01	2.32E-01	3.41E-01	2.26E-01	1.14E-01
WP1(10%) - **FVLC**	**1.19E-03**	**7.40E-03**	**3.67E-03**	**4.55E-03**	**1.19E-02**	**1.53E-03**	**1.39E-03**	**1.25E-03**
WP2 - base	1.53E-01	4.99E-02	1.36E-01	6.67E-02	1.43E-01	1.73E-01	1.40E-01	6.37E-02
WP2 - **FVLC**	**1.10E-02**	**8.72E-03**	**3.35E-03**	**4.55E-03**	**1.05E-02**	**1.53E-03**	**1.55E-03**	**1.29E-03**
IP - base	1.68E-01	6.28E-02	1.38E-01	8.81E-02	1.71E-01	2.23E-01	1.73E-01	8.09E-02
IP - **FVLC**	**8.02E-03**	**8.29E-03**	**3.24E-03**	**4.56E-03**	**1.04E-02**	**1.59E-03**	**1.50E-03**	**1.25E-03**
WP1(5%)+WP2 - base	1.85E-01	3.50E-02	1.28E-01	6.65E-02	1.44E-01	1.79E-01	1.60E-01	6.32E-02
WP1(5%)+WP2 - **FVLC**	**1.20E-02**	**7.46E-03**	**3.67E-03**	**4.55E-03**	**9.81E-02**	**1.51E-03**	**1.54E-03**	**1.28E-03**
WP1(10%)+WP2 - base	1.17E-01	8.62E-02	1.93E-01	1.32E-01	1.76E-01	3.45E-01	2.52E-01	1.17E-01
WP1(10%)+WP2 - **FVLC**	**1.18E-02**	**9.41E-03**	**2.06E-02**	**1.44E-02**	**1.87E-02**	**3.79E-02**	**2.74E-02**	**1.18E-02**
WP1(10%)+WP2+IP - base	1.36E-01	1.05E-01	2.22E-01	1.55E-01	1.95E-01	3.54E-01	2.62E-01	1.44E-01
WP1(10%)+WP2+IP - **FVLC**	**1.43E-02**	**1.11E-02**	**2.39E-02**	**1.77E-02**	**2.21E-02**	**4.54E-02**	**3.35E-02**	**1.34E-02**

Ablation Study

Method	Setting			CIFAR10		CIFAR100		CUB		Places365	
	$\mathcal{L}_2$	$\mathcal{L}_3$	$\mathcal{L}_4$	TCPC	TOPC	TCPC	TOPC	TCPC	TOPC	TCPC	TOPC
WP1(10%) - FVLC				1.99E-01	8.36E-02	1.94E-01	1.31E-01	2.32E-01	3.41E-01	2.26E-01	1.14E-01
	✓			2.09E-02	3.14E-02	2.81E-02	4.88E-02	4.08E-01	7.56E-02	4.69E-02	6.14E-02
		✓		1.80E-02	1.79E-02	2.01E-02	2.85E-02	3.77E-01	4.50E-02	4.48E-02	3.68E-02
			✓	4.78E-03	3.11E-02	1.67E-02	2.19E-02	5.69E-02	5.52E-03	6.04E-03	4.98E-03
	✓	✓		1.67E-02	1.11E-02	5.39E-02	6.85E-02	1.69E-01	2.12E-02	1.98E-02	1.81E-02
	✓		✓	1.65E-03	1.01E-02	5.08E-03	6.43E-03	1.61E-02	2.14E-03	1.95E-03	1.72E-03
		✓	✓	1.63E-03	1.02E-02	5.04E-03	6.27E-03	1.63E-02	2.10E-03	1.94E-03	1.71E-03
	✓	✓	✓	**1.19E-03**	**7.40E-03**	**3.67E-03**	**4.55E-03**	**1.19E-02**	**1.53E-03**	**1.39E-03**	**1.25E-03**
WP2 - FVLC				1.53E-01	4.99E-02	1.36E-01	6.67E-02	1.43E-01	1.73E-01	1.40E-01	6.37E-02
	✓			7.62E-02	2.29E-02	2.02E-02	2.08E-02	6.09E-02	5.22E-03	1.04E-02	5.31E-03
		✓		5.37E-02	1.84E-02	1.46E-02	1.16E-02	6.80E-02	2.95E-03	1.06E-02	2.36E-03
			✓	5.19E-02	1.46E-02	1.31E-02	9.34E-03	5.30E-02	3.24E-03	6.83E-03	2.12E-03
	✓	✓		4.57E-02	1.82E-02	1.45E-02	1.01E-02	5.01E-02	3.46E-03	6.30E-03	2.94E-03
	✓		✓	2.73E-02	6.96E-03	7.10E-03	4.60E-03	2.42E-02	1.91E-03	3.97E-03	2.36E-03
		✓	✓	2.47E-02	8.00E-03	5.29E-03	5.25E-03	2.40E-02	1.70E-03	3.82E-03	1.96E-03
	✓	✓	✓	**1.10E-02**	**8.72E-03**	**3.35E-03**	**4.55E-03**	**1.05E-02**	**1.53E-03**	**1.55E-03**	**1.29E-03**
IP - FVLC				1.68E-01	6.28E-02	1.38E-01	8.81E-02	1.71E-01	2.23E-01	1.73E-01	8.09E-02
	✓			6.39E-02	2.55E-02	2.56E-02	3.57E-02	8.04E-02	1.27E-02	1.16E-02	9.80E-03
		✓		3.82E-02	3.91E-02	1.53E-02	2.16E-02	4.77E-02	7.66E-03	6.99E-03	5.88E-03
			✓	1.63E-02	1.73E-02	6.35E-03	9.35E-03	2.14E-02	3.30E-03	3.01E-03	2.59E-03
	✓	✓		2.04E-02	2.10E-02	8.35E-03	1.15E-02	2.67E-02	4.04E-03	3.77E-03	3.18E-03
	✓		✓	1.23E-02	1.27E-02	5.09E-03	6.94E-03	1.65E-02	2.45E-03	2.35E-03	1.94E-03
		✓	✓	1.26E-02	1.09E-02	5.07E-03	7.00E-03	1.59E-02	2.53E-03	2.29E-03	1.95E-03
	✓	✓	✓	**8.02E-03**	**8.29E-03**	**3.24E-03**	**4.56E-03**	**1.04E-02**	**1.59E-03**	**1.50E-03**	**1.25E-03**
WP1(10%)+WP2 - FVLC				1.17E-01	8.62E-02	1.93E-01	1.32E-01	1.76E-01	3.45E-01	2.52E-01	1.17E-01
	✓			5.97E-02	6.28E-02	1.43E-01	9.94E-02	1.24E-01	2.58E-01	1.86E-01	8.24E-02
		✓		3.76E-02	6.34E-02	1.48E-01	1.03E-01	1.21E-01	2.63E-01	1.90E-01	8.13E-02
			✓	3.52E-02	2.54E-02	6.03E-02	4.13E-02	5.08E-02	1.07E-01	7.63E-02	3.31E-02
	✓	✓		9.21E-02	5.45E-02	1.28E-01	8.56E-02	1.04E-01	2.24E-01	1.60E-01	6.82E-02
	✓		✓	3.15E-02	2.90E-02	6.88E-02	4.71E-02	5.76E-02	1.21E-01	8.86E-02	3.72E-02
		✓	✓	1.53E-02	1.18E-02	2.78E-02	1.89E-02	2.37E-02	4.88E-02	3.62E-02	1.55E-02
	✓	✓	✓	**1.18E-02**	**9.41E-03**	**2.06E-02**	**1.44E-02**	**1.87E-02**	**3.79E-02**	**2.74E-02**	**1.18E-02**
WP1(10%)+WP2+IP - FVLC				1.36E-01	1.05E-01	2.22E-01	1.55E-01	1.95E-01	3.54E-01	2.62E-01	1.44E-01
	✓			7.37E-02	5.75E-02	1.29E-01	9.24E-02	1.21E-01	2.34E-01	1.74E-01	7.30E-02
		✓		7.85E-02	5.85E-02	1.23E-01	9.17E-02	1.14E-01	2.43E-01	1.70E-01	6.80E-02
			✓	4.59E-02	3.67E-02	7.62E-02	5.57E-02	6.74E-02	1.41E-01	1.08E-01	4.23E-02
	✓	✓		5.81E-02	4.52E-02	9.69E-02	7.11E-02	8.99E-02	1.83E-01	1.34E-01	5.49E-02
	✓		✓	5.84E-02	4.53E-02	9.75E-02	7.17E-02	8.94E-02	1.85E-01	1.36E-01	5.41E-02
		✓	✓	2.91E-02	2.31E-02	4.85E-02	3.56E-02	4.46E-02	9.33E-02	6.78E-02	2.81E-02
	✓	✓	✓	**1.43E-02**	**1.11E-02**	**2.39E-02**	**1.77E-02**	**2.21E-02**	**4.54E-02**	**3.35E-02**	**1.34E-02**

Contact

Corresponding to lijie.hu@kaust.edu.sa