University of Central Florida

NEURAL INFORMATION PROCESSING SYSTEMS

Paper

https://openreview.net/forum?id=ek5KY0D3CL#discussion

Mitigating Object Hallucination in Large Vision-Language Models via False Discovery Controlled Visual Data Splitting

Chang Liu[1], Yu Tian[2], and Rui Xie[1]*
[1] School of Data, Mathematical, and Statistical Sciences
[2] Institute of Artificial Intelligence
University of Central Florida, Orlando, FL

- LVLMs generate objects not supported by visual evidences, which arise partly from **visual uncertainty** and over-reliance on language priors.
- **Visual uncertainty** provides informative signals about the reliability of visual grounding in LVLMs rather than being merely a source of error.
- Existing approaches largely rely on heuristic decoding signals and lack explicit control over the proportion of hallucinated objects at the image level.



Query: On this table, there are

Q1: Is there fork?

Q2: Is there broccoli?

Q3: Is there carrot?

Q4: Is there Shrimp?

Raw Outputs

Fork ✓
Broccoli ✓
Carrot ✓
Shrimp ✓
(Hallucinated)

Within this image, Mirror Statistics via FDR control at 10%

Rejected Object: Shrimp

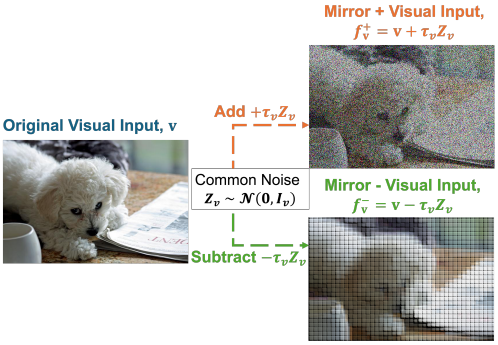
CORAL Selected Object

Fork ✓
Broccoli ✓
Carrot ✓
~~Shrimp~~ ✗

Challenge: Modeling Visual Uncertainty

- We introduce **uncertainty-aware visual data splitting** to construct two paired visual inputs via **symmetric perturbations**, inducing controlled variability in LVLM perception, where $Z_v \sim \mathcal{N}(0, \mathbf{I}_v)$.

$$f_{\mathbf{v}}^{+} = \mathbf{v} + \tau_v Z_v, \qquad f_{\mathbf{v}}^{-} = \mathbf{v} - \tau_v Z_v, \tag{1}$$



Challenge: Statistical Control of Hallucination

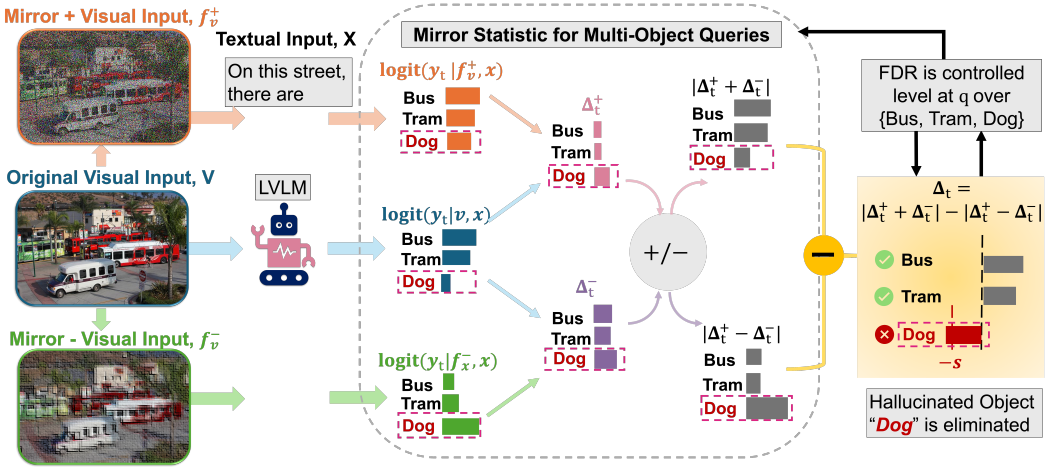
- Construct mirror statistics for estimating and controlling **false discovery rate (FDR)** without access to ground truth.
  - Mirror statistics quantify the consistency of visual evidence across symmetrically perturbed inputs, helping distinguish visually grounded objects from hallucinated ones.

$$\text{FDR} = \mathbb{E}\left[\frac{\#\{\text{hallucinated objects identified as grounded}\}}{\#\{\text{all identified objects}\}}\right] \tag{2}$$

e.g., among four predicted objects, one (shrimp) is hallucinated, yielding an FDR of 25%.

$$\text{Power} = \mathbb{E}\left[\frac{\#\{\text{truly grounded objects correctly identified}\}}{\#\{\text{all truly grounded objects}\}}\right] \tag{3}$$

e.g., retains all visually grounded objects (*fork*, *broccoli*, and *carrot*), achieving 100% power.


Mirror + Visual Input, f_v^+
Textual Input, X
On this street, there are
Original Visual Input, V
LVLM
Mirror - Visual Input, f_v^-
Mirror Statistic for Multi-Object Queries
$\text{logit}(y_t|f_v^+, x)$
Bus
Tram
Dog
$\text{logit}(y_t|v, x)$
$\text{logit}(y_t|f_x^-, x)$
Δ_t^+
Δ_t^-
+/−
$|\Delta_t^+ + \Delta_t^-|$
$|\Delta_t^+ - \Delta_t^-|$
FDR is controlled level at q over {Bus, Tram, Dog}
$\Delta_t =$
$|\Delta_t^+ + \Delta_t^-| - |\Delta_t^+ - \Delta_t^-|$
$-s$
Hallucinated Object "Dog" is eliminated

Mirror Statistic

- For each generated token y_t, we compare the LVLM's logits under the original visual input $\mathbf{v}$ and two mirror views $f_{\mathbf{v}}^{+}$ and $f_{\mathbf{v}}^{-}$.

$$\Delta_t^{+} = \text{logit}_\theta(y_t \mid \mathbf{v}, \mathbf{x}, \mathbf{y}_{<t}) - \text{logit}_\theta\big(y_t \mid f_{\mathbf{v}}^{+}, \mathbf{x}, \mathbf{y}_{<t}\big)\,, \tag{4}$$

$$\Delta_t^{-} = \text{logit}_\theta(y_t \mid \mathbf{v}, \mathbf{x}, \mathbf{y}_{<t}) - \text{logit}_\theta\big(y_t \mid f_{\mathbf{v}}^{-}, \mathbf{x}, \mathbf{y}_{<t}\big)\,. \tag{5}$$

- Following the Gaussian mirrors construction (Xing et al., 2023), we define the **mirror statistic** for visual uncertainty

$$\Delta_t = \big|\Delta_t^{+} + \Delta_t^{-}\big| - \big|\Delta_t^{+} - \Delta_t^{-}\big|. \tag{6}$$

  - $\big|\Delta_t^{+} + \Delta_t^{-}\big|$: captures signal strength through consistent responses across symmetric perturbations,
  - $\big|\Delta_t^{+} - \Delta_t^{-}\big|$: reflects the noise cancellation effect.

Null Symmetry for FDR Control

- **Null Symmetry:** Under approximate conditional sign-flip invariance, the mirror statistic is approximately symmetric around zero for hallucinated tokens.

$$H_{0t} : \text{no systematic visual evidence supports token } y_t,$$

and

$$\mathcal{G}_t = \sigma(\mathbf{v}, \mathbf{x}, \mathbf{y}_{<t})$$

- **Conditional Null Symmetry**: Under conditional sign-flip invariance, the mirror statistic satisfies the null symmetry property. For Δ_t defined in Eq. (6), if

$$(\Delta_t^+, \Delta_t^-) \mid H_{0t}, \mathcal{G}_t \overset{d}{=} (\Delta_t^+, -\Delta_t^-) \mid H_{0t}, \mathcal{G}_t,$$

then, for every $s > 0$,

$$\Pr(\Delta_t \leq -s \mid H_{0t}, \mathcal{G}_t) = \Pr(\Delta_t \geq s \mid H_{0t}, \mathcal{G}_t). \tag{7}$$

Controlling Hallucination via FDR

- **FDR Estimation:**

$$\widehat{\mathrm{FDR}}(s) = \frac{\#\{y_t \mid \Delta_t \leq -s\}}{\#\{y_t \mid \Delta_t \geq s\} \vee 1}. \tag{8}$$

- **FDR Control:** Given a target FDR level q (e.g., $q = 0.1$)„ CORAL selects a data-driven threshold to retain visually grounded tokens while limiting the estimated false discovery proportion.
- **Data-Driven Thresholding:** For a target FDR level q, select

$$T_q = \min\left\{s > 0 : \widehat{\mathrm{FDR}}(s) \leq q\right\}. \tag{9}$$

Retain $\{y_t \in \mathcal{S} : \Delta_t \geq T_q\}$ and suppress the remaining tokens.

Image-Level FDR Control at 10%: FDR and Power Results

Table 1: Evaluation of overall FDR control and overall power across multiple LVLM architectures on MSCOCO (3000 runs). Overall FDR and power denote averages of per-image FDR and power. **Bold** indicates the best result and underline indicates the second-best result.

Method	LLaVA-OneVision-7B		Qwen2.5-VL-7B		InternVL3-8B		Average	
	Overall FDR↓	Overall Power↑	Overall FDR↓	Overall Power↑	Overall FDR↓	Overall Power↑	Overall FDR↓	Overall Power↑
Random								
Regular	0.0935 (± 0.0133)	75.36 (± 1.88)	0.0944 (± 0.0114)	77.43 (± 1.22)	0.0946 (± 0.0122)	73.52 (± 1.77)	0.0942 (± 0.0111)	75.44 (± 1.44)
VCD (2024)	0.0892 (± 0.0188)	78.50 (± 1.23)	0.0908 (± 0.0132)	79.14 (± 1.53)	0.0899 (± 0.0109)	78.86 (± 1.34)	0.0900 (± 0.0145)	78.83 (± 1.31)
MARINE (2025)	0.0846 (± 0.0122)	81.25 (± 1.17)	0.0866 (± 0.0116)	80.05 (± 1.60)	0.0889 (± 0.0133)	79.84 (± 1.33)	0.0867 (± 0.0124)	80.38 (± 1.21)
AGLA (2025)	0.0799 (± 0.0223)	84.92 (± 1.11)	0.0821 (± 0.0114)	84.67 (± 1.41)	0.0810 (± 0.0144)	80.92 (± 1.33)	0.0810 (± 0.0123)	83.50 (± 1.50)
CORAL(Ours)	**0.0691 (± 0.0113)**	**91.43 (± 1.42)**	**0.0718 (± 0.0332)**	**94.60 (± 1.49)**	**0.0799 (± 0.0122)**	**86.88 (± 1.12)**	**0.0736 (± 0.0149)**	**90.97 (± 1.44)**
Popular								
Regular	0.0971 (± 0.0205)	80.11 (± 1.28)	0.0903 (± 0.0111)	77.01 (± 1.64)	0.0937 (± 0.0106)	50.80 (± 1.69)	0.0937 (± 0.0141)	69.31 (± 1.54)
VCD (2024)	0.0893 (± 0.0187)	83.11 (± 1.32)	0.0911 (± 0.0133)	75.10 (± 1.26)	0.0875 (± 0.0211)	67.22 (± 1.11)	0.0893 (± 0.0177)	75.14 (± 1.23)
MARINE (2025)	0.0773 (± 0.0134)	85.22 (± 1.55)	0.0883 (± 0.0220)	78.28 (± 1.25)	0.0701 (± 0.0221)	87.20 (± 1.11)	0.0786 (± 0.0192)	83.57 (± 1.30)
AGLA (2025)	0.0733 (± 0.0177)	88.13 (± 1.69)	0.0861 (± 0.0233)	82.22 (± 1.51)	0.0741 (± 0.0188)	81.14 (± 1.55)	0.0778 (± 0.0199)	83.83 (± 1.58)
CORAL(Ours)	**0.0552 (± 0.0122)**	**90.42 (± 0.24)**	**0.0802 (± 0.0200)**	**84.10 (± 1.15)**	**0.0675 (± 0.0155)**	**88.93 (± 1.21)**	**0.0676 (± 0.0159)**	**87.82 (± 0.87)**
Adversarial								
Regular	0.0988 (± 0.0114)	51.22 (± 1.55)	0.0973 (± 0.0108)	52.19 (± 1.13)	0.0953 (± 0.0133)	54.88 (± 1.97)	0.0971 (± 0.0118)	52.76 (± 1.55)
VCD (2024)	0.0903 (± 0.0115)	56.32 (± 1.56)	0.0935 (± 0.0109)	53.16 (± 1.77)	0.0901 (± 0.0131)	60.59 (± 1.26)	0.0913 (± 0.0118)	56.69 (± 1.53)
MARINE (2025)	0.0864 (± 0.0105)	61.36 (± 1.99)	0.0903 (± 0.0117)	60.22 (± 1.53)	0.0896 (± 0.0156)	70.46 (± 1.59)	0.0888 (± 0.0126)	64.01 (± 1.70)
AGLA (2025)	0.0775 (± 0.0198)	75.77 (± 1.28)	0.0899 (± 0.0155)	73.99 (± 1.11)	0.0826 (± 0.0211)	81.36 (± 1.78)	0.0833 (± 0.0188)	77.04 (± 1.39)
CORAL(Ours)	**0.0694 (± 0.0155)**	**88.33 (± 1.45)**	**0.0839 (± 0.0111)**	**83.63 (± 1.55)**	**0.0733 (± 0.0122)**	**87.59 (± 1.33)**	**0.0755 (± 0.0129)**	**86.52 (± 1.44)**

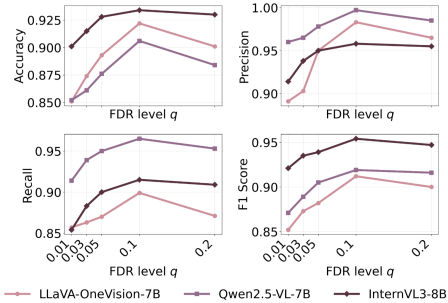


Figure 4: Ablation study on the effect of FDR target level (q) on the performance of LLaVA-OneVision-7B, Qwen2.5-VL-7B, InternVL3-8B using POPE metrics with $q = \{0.01, 0.03, 0.05, 0.1, 0.2\}$.

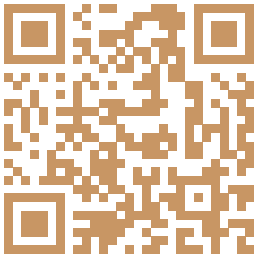
Table 5: Ablation study on POPE metrics using the MSCOCO dataset with LLaVA-OneVision-7B. Results are mean ± std over 10 runs. **Bold** indicates the best result.

Method	Accuracy ↑	Precision ↑	Recall ↑	F1 Score ↑
Regular	85.87 (± 1.77)	83.41 (± 2.13)	88.08 (± 1.47)	85.72 (± 1.66)
w/o Visual Uncertainty Splitting	88.12 (± 0.62)	86.78 (± 0.74)	89.07 (± 0.71)	87.90 (± 0.58)
w/o Mirror Statistic	88.76 (± 0.55)	87.07 (± 0.49)	90.33 (± 0.61)	88.67 (± 0.53)
w/o Overall FDR Control	89.36 (± 0.48)	86.42 (± 0.92)	92.37 (± 0.56)	89.27 (± 0.41)
CORAL (Ours)	**92.17 (± 0.89)**	**88.25 (± 1.46)**	**94.87 (± 1.61)**	**91.64 (± 1.83)**

Thank you!



Paper



Code

Appendix

POPE Evaluation on MSCOCO

Table 2: Evaluation with POPE score across multiple LVLM architectures on the MSCOCO dataset. We report individualized Accuracy and F1 score (mean ± std over 3000 runs). **Bold** indicates the best result and underline indicates the second-best result.

Method	LLaVA-OneVision-7B		Qwen2.5-VL-7B		InternVL3-8B		Average	
	Accuracy↑	F1 Score↑	Accuracy↑	F1 Score↑	Accuracy↑	F1 Score↑	Accuracy↑	F1 Score↑
Random								
Regular	85.87 (± 1.77)	85.72 (± 1.66)	86.20 (± 1.30)	86.98 (± 1.28)	87.15 (± 1.88)	86.26 (± 1.80)	86.41 (± 1.65)	86.32 (± 1.58)
VCD (2024)	86.33 (± 1.23)	88.86 (± 1.43)	87.35 (± 1.83)	90.06 (± 2.03)	88.15 (± 1.53)	88.05 (± 1.45)	87.28 (± 1.53)	88.99 (± 1.64)
MARINE (2025)	87.21 (± 1.35)	89.09 (± 1.09)	89.72 (± 1.11)	90.33 (± 2.09)	89.23 (± 1.35)	89.15 (± 1.19)	88.72 (± 1.27)	89.52 (± 1.46)
AGLA (2025)	88.15 (± 1.65)	89.97 (± 1.21)	90.01 (± 1.23)	91.17 (± 1.51)	92.53 (± 1.42)	93.54 (± 1.77)	90.23 (± 1.43)	91.56 (± 1.50)
CORAL (Ours)	**92.17 (± 0.89)**	**91.64 (± 1.83)**	**90.63 (± 1.54)**	**91.89 (± 1.83)**	**93.55 (± 1.52)**	**95.35 (± 1.87)**	**92.12 (± 1.32)**	**92.96 (± 1.84)**
Popular								
Regular	82.72 (± 1.99)	83.16 (± 1.85)	83.88 (± 1.48)	84.79 (± 1.52)	84.63 (± 1.48)	85.26 (± 1.15)	83.74 (± 1.65)	84.40 (± 1.51)
VCD (2024)	84.24 (± 1.46)	86.16 (± 1.30)	86.32 (± 1.83)	89.34 (± 2.04)	85.29 (± 1.14)	86.25 (± 2.00)	85.28 (± 1.48)	87.25 (± 1.78)
MARINE (2025)	86.81 (± 1.21)	88.19 (± 2.09)	88.72 (± 1.38)	90.33 (± 1.18)	87.67 (± 1.53)	88.37 (± 1.81)	87.73 (± 1.37)	88.96 (± 1.69)
AGLA (2025)	89.28 (± 1.11)	89.34 (± 1.70)	88.14 (± 1.23)	90.79 (± 1.15)	89.42 (± 1.32)	90.17 (± 1.22)	88.95 (± 1.22)	90.10 (± 1.36)
CORAL (Ours)	**90.37 (± 1.89)**	**89.91 (± 1.33)**	**90.00 (± 1.45)**	**91.28 (± 1.83)**	**91.33 (± 1.35)**	**93.57 (± 1.44)**	**90.57 (± 1.56)**	**91.59 (± 1.53)**
Adversarial								
Regular	80.28 (± 2.21)	81.02 (± 2.07)	83.17 (± 1.37)	83.46 (± 2.08)	84.59 (± 1.33)	84.96 (± 1.80)	82.68 (± 1.64)	83.15 (± 1.98)
VCD (2024)	81.21 (± 1.89)	82.96 (± 1.82)	85.43 (± 1.88)	85.24 (± 2.04)	85.88 (± 1.65)	86.99 (± 1.77)	84.17 (± 1.81)	85.06 (± 1.88)
MARINE (2025)	85.26 (± 2.11)	85.29 (± 2.09)	87.42 (± 1.11)	88.73 (± 1.91)	87.25 (± 1.75)	88.11 (± 1.16)	86.64 (± 1.66)	87.38 (± 1.72)
AGLA (2025)	86.44 (± 1.61)	**88.14 (± 2.01)**	91.39 (± 1.23)	89.73 (± 2.15)	90.11 (± 1.32)	89.83 (± 1.10)	89.31 (± 1.39)	89.23 (± 1.75)
CORAL (Ours)	**88.27 (± 1.81)**	87.99 (± 2.03)	**93.75 (± 1.54)**	**90.87 (± 1.83)**	**91.03 (± 1.00)**	**90.48 (± 1.83)**	**91.02 (± 1.45)**	**89.78 (± 1.90)**

MME Evaluation on MSCOCO

Table 3: Evaluation with MME score (mean ± std over 3000 runs) across multiple LVLM architectures on MSCOCO. **Bold** indicates the best result and underline indicates the second-best result.

Model	Method	Object		Attribute	Relation		Total ↑
		Existence ↑	Count ↑	Color ↑	Position ↑	Commonsense ↑	
LLaVA-OneVision-7B	Regular	190.33 (± 6.50)	145.53 (± 15.20)	170.66 (± 9.10)	160.25 (± 8.30)	70.16 (± 6.50)	736.93 (± 24.80)
	VCD (2024)	186.25 (± 7.22)	147.25 (± 11.44)	175.35 (± 15.58)	165.36 (± 2.55)	79.42 (± 5.74)	753.63 (± 18.76)
	MARINE (2025)	191.26 (± 4.55)	150.44 (± 10.15)	177.45 (± 10.96)	165.36 (± 7.19)	82.33 (± 5.29)	766.84 (± 17.47)
	AGLA (2025)	190.54 (± 6.22)	153.32 (± 15.37)	178.42 (± 3.22)	170.35 (± 4.22)	84.36 (± 5.21)	776.99 (± 15.96)
	CORAL (Ours)	**193.67 (± 3.21)**	**160.43 (± 10.11)**	**180.35 (± 8.46)**	**175.24 (± 10.58)**	**85.52 (± 4.22)**	**795.21 (± 18.12)**
Qwen2.5-VL-7B	Regular	185.52 (± 5.80)	140.46 (± 12.20)	165.16 (± 8.60)	155.22 (± 7.50)	80.53 (± 6.20)	726.89 (± 21.30)
	VCD (2024)	187.35 (± 7.36)	145.63 (± 10.45)	173.35 (± 8.23)	160.22 (± 11.34)	82.55 (± 6.29)	749.10 (± 11.61)
	MARINE (2025)	189.83 (± 4.25)	153.35 (± 10.24)	176.25 (± 7.21)	165.14 (± 10.44)	84.21 (± 4.87)	768.78 (± 11.88)
	AGLA (2025)	190.15 (± 3.28)	158.33 (± 11.38)	179.35 (± 4.15)	168.13 (± 9.83)	87.28 (± 2.28)	783.24 (± 13.42)
	CORAL (Ours)	**194.24 (± 3.18)**	**161.35 (± 10.01)**	**182.53 (± 6.98)**	**170.86 (± 8.14)**	**88.91 (± 3.71)**	**797.89 (± 12.67)**
InternVL3-8B	Regular	192.25 (± 6.20)	150.22 (± 10.80)	175.48 (± 7.40)	165.16 (± 6.50)	95.38 (± 5.20)	778.49 (± 19.80)
	VCD (2024)	193.26 (± 3.28)	159.25 (± 10.97)	178.25 (± 9.25)	175.35 (± 7.73)	95.15 (± 1.54)	801.26 (± 13.36)
	MARINE (2025)	**193.87 (± 2.82)**	163.28 (± 3.29)	180.93 (± 5.18)	177.35 (± 4.26)	96.18 (± 1.58)	811.61 (± 13.42)
	AGLA (2025)	193.17 (± 3.28)	165.26 (± 7.77)	183.85 (± 5.28)	177.96 (± 4.26)	95.15 (± 1.77)	815.39 (± 12.81)
	CORAL (Ours)	193.46 (± 4.26)	**168.36 (± 10.53)**	**185.29 (± 13.19)**	**178.54 (± 1.35)**	**96.28 (± 1.98)**	**821.93 (± 12.30)**