CLIDE: Zero-Shot Detection of AI-Generated Images via Conditional Likelihood

Roy Betser, Omer Hofman,
Roman Vainstein, Guy Gilboa

WACV 2026

Motivation

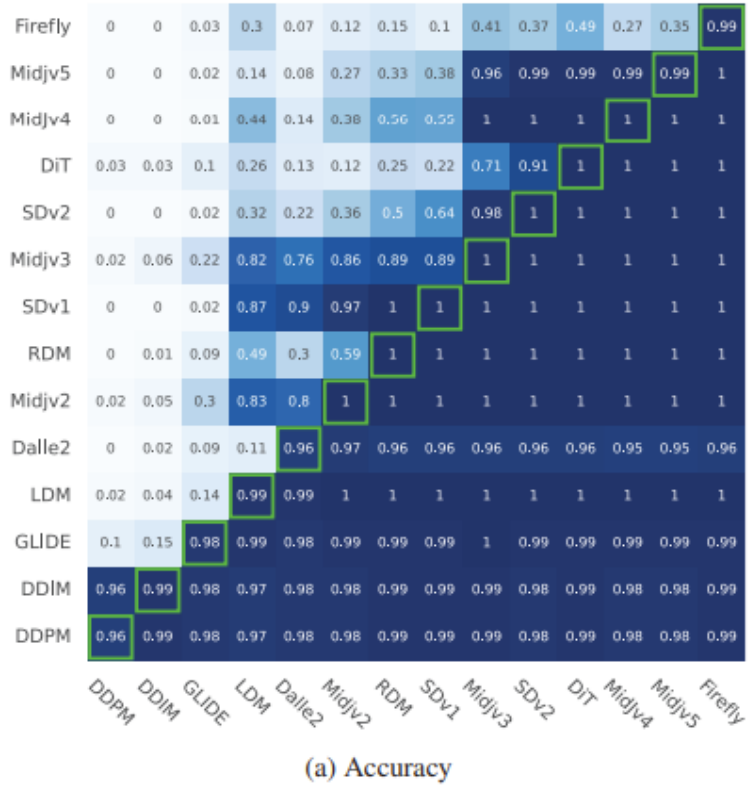
- Generative models produce highly photorealistic images.
- Risk: spreading misinformation.

Real

Generated

The Challenge

- Early detection methods –
  - Fully-supervised
  - Few-shot



(a) Accuracy

- Key issues:
  - Rapidly emerging generative models.
  - Training data maintenance.

Image from Epstein et al. 2023

Zero-shot Detection

- Zero-shot detection:
  - No access to generated content.
  - No task-specific training.

- Existing zero-shot methods
  - **Implicitly** model real-image distributions.
  - Often based on some images **transformation.**

Background - Whitened CLIP

- Given a set I of images:
  - Process all images with and encoder: $\mathrm{X} = \mathrm{CLIP}(\mathrm{I})$.
  - Compute an empirical mean μ and covariance matrix - Σ.
- Whitening matrix:
  - Satisfies $W^T W = \Sigma^{-1}$.
  - Not unique.
  - Whitening transform: $Y = W\hat{X}$.
  - Whitened data Y has zero mean, unit variance and identity covariance matrix.

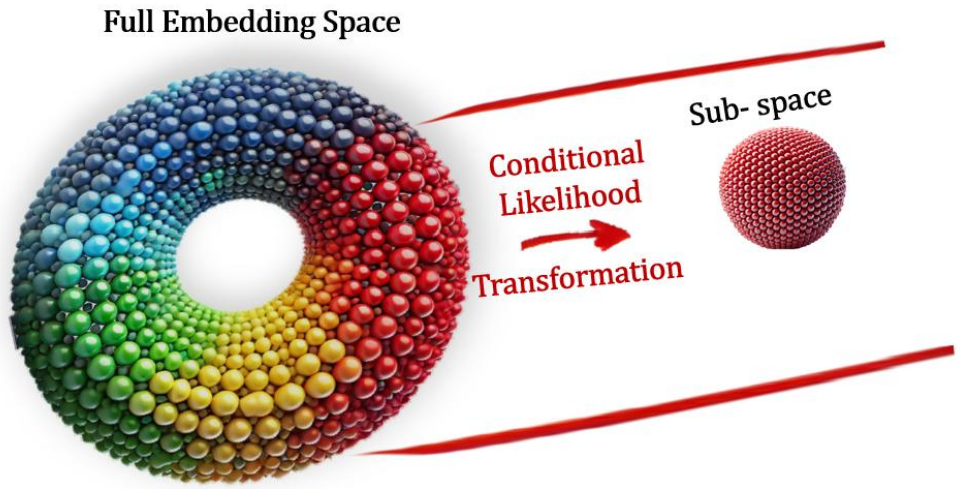
Background - Whitened CLIP

- If the Embeddings are approximately Gaussian:
  - Whitened data will also be approximately Gaussian.
  - Whitened data already has zero mean, unit variance and no correlations.

- Gaussian log-likelihood estimation:
  - $\ell(x) = Log\big(P(x)\big) = -\frac{1}{2}\big(d \cdot \log(2\pi) + \|Wx\|^2\big)$

Conditional Likelihood

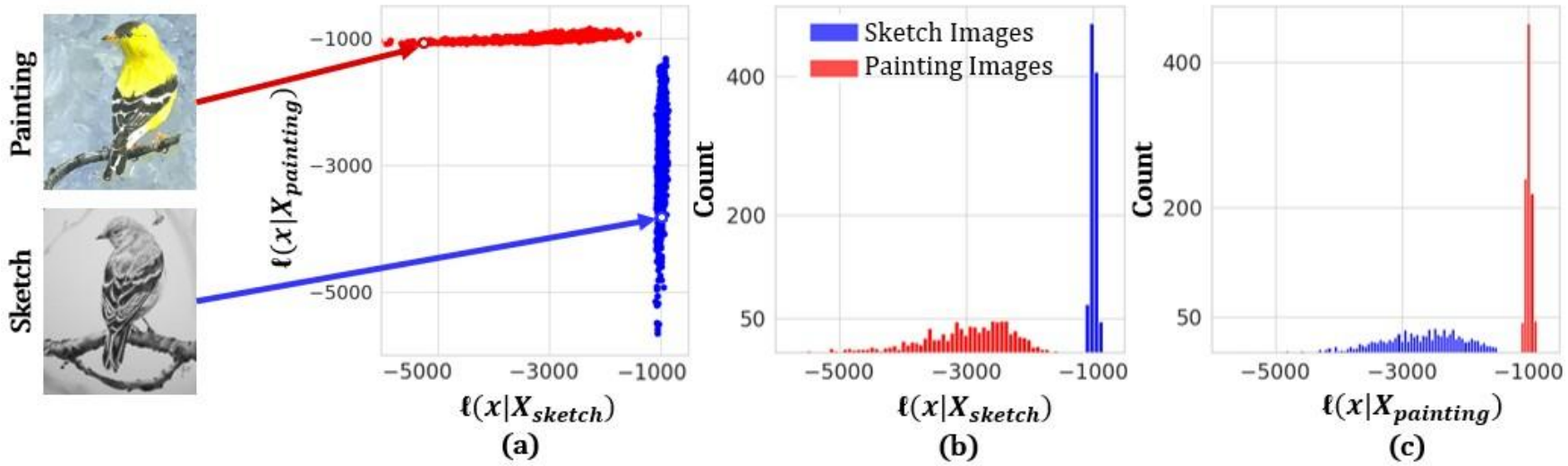
- Motivation – focusing on a sub-space in the full embedding space.
- We find that sub-spaces:
  - Maintain normal distribution.
  - Have effectively lower ranked covariance matrix.



Conditional Likelihood

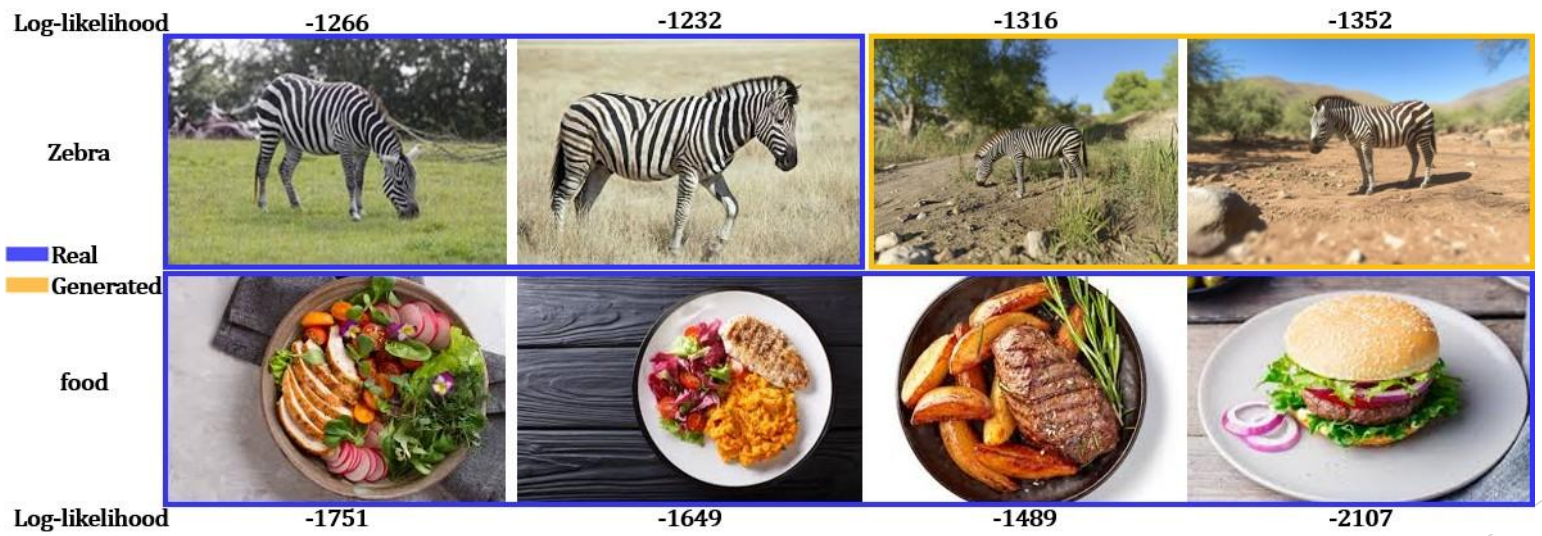
- For a set X of samples from sub-space D ($x_i \in D$):

$$\ell(x|X,m) = -\frac{1}{2}(m \cdot \log(2\pi) + \|W(X,m)\hat{x}\|^2$$



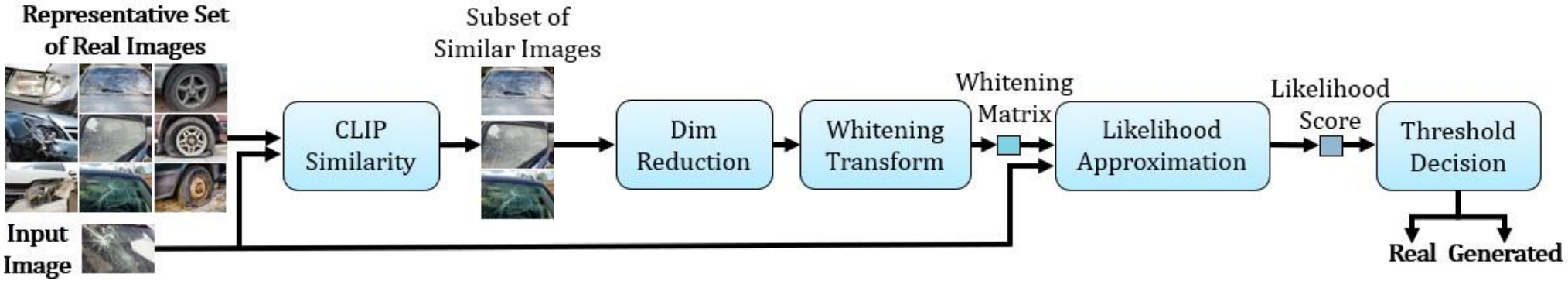
Local Neighborhood

- Goal: Using only **real** images for whitening then **generated** images will become outliers.
- Problem: CLIP embeddings hold additional semantic information.



CLIDE

- Detection pipeline:



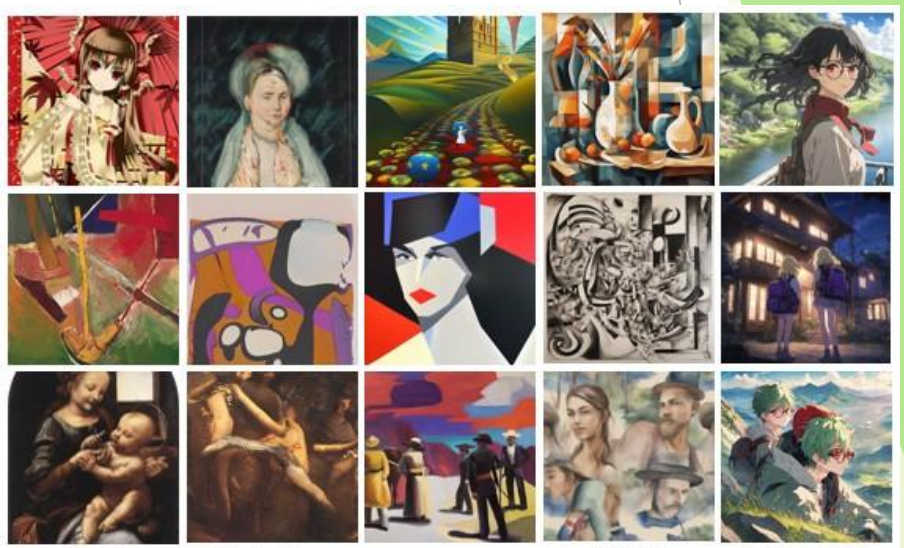
Results - General Images

Generative Model	AEROBLADE [59]				RIGID [30]				ZED [18]				Manifold Bias [11]				Ours			
	AUC ↑	AP ↑	F1 ↑	Acc ↑	AUC ↑	AP ↑	F1 ↑	Acc ↑	AUC ↑	AP ↑	F1 ↑	Acc ↑	AUC ↑	AP ↑	F1 ↑	Acc ↑	AUC ↑	AP ↑	F1 ↑	Acc ↑
ProGan [34]	0.54	0.49	0.65	0.54	0.45	0.46	0.16	0.48	0.74	0.75	0.7	0.64	**0.96**	**0.97**	**0.88**	**0.88**	0.95	**0.96**	0.87	**0.88**
StyleGan [35]	0.58	0.59	0.63	0.51	0.71	0.72	0.56	0.65	0.75	**0.77**	0.71	0.65	0.68	0.74	0.61	0.68	**0.76**	0.71	**0.75**	**0.73**
StyleGan2 [36]	0.74	0.76	0.68	0.61	0.52	0.54	0.31	0.53	0.75	0.8	0.68	0.61	0.62	0.66	0.48	0.6	**0.8**	**0.82**	**0.73**	**0.68**
BigGAN [9]	0.65	0.61	0.68	0.6	0.62	0.63	0.42	0.57	0.45	0.47	0.63	0.5	0.9	0.9	0.8	0.81	**0.96**	**0.96**	**0.87**	**0.88**
GauGan [51]	0.7	0.71	0.68	0.6	0.32	0.39	0.1	0.45	0.33	0.41	0.6	0.44	**0.98**	**0.98**	**0.91**	**0.9**	0.83	0.85	0.76	0.73
CycleGan [75]	0.76	0.73	0.73	0.69	0.34	0.39	0.05	0.43	0.22	0.38	0.6	0.43	0.97	0.97	0.89	0.9	**0.98**	**0.98**	**0.9**	**0.91**
CRN [15]	0.34	0.51	0.48	0.59	0.25	0.36	0.01	0.42	**0.99**	**0.99**	**0.92**	**0.92**	0.93	0.9	0.88	0.88	0.98	**0.99**	0.91	**0.92**
SD V1.4 [60]	0.47	0.47	0.63	0.51	0.79	0.77	0.63	0.69	0.62	0.65	0.65	0.54	0.71	0.62	0.4	0.57	**0.9**	**0.91**	**0.82**	**0.82**
SD V1.5 [60]	0.49	0.49	0.64	0.52	0.77	0.76	0.63	0.68	0.61	0.63	0.65	0.54	0.72	0.64	0.44	0.58	**0.9**	**0.91**	**0.83**	**0.82**
Guided DM [20]	0.55	0.49	0.66	0.57	0.51	0.51	0.25	0.5	0.58	0.55	0.68	0.59	0.8	0.8	0.66	0.7	**0.87**	**0.87**	**0.79**	**0.78**
LDM 100 [60]	0.48	0.48	0.62	0.48	0.51	0.51	0.25	0.5	0.52	0.56	0.63	0.49	0.84	0.88	0.78	0.79	**0.92**	**0.93**	**0.84**	**0.85**
LDM 200 [60]	0.47	0.48	0.62	0.48	0.51	0.51	0.25	0.5	0.53	0.56	0.63	0.5	0.85	0.88	0.79	0.8	**0.92**	**0.93**	**0.85**	**0.85**
Glide 50 27 [48]	0.06	0.32	0.59	0.42	0.51	0.51	0.25	0.5	**0.98**	**0.98**	**0.91**	**0.92**	0.93	0.94	0.86	0.86	0.91	0.92	0.84	0.84
Glide 100 27 [48]	0.08	0.32	0.6	0.43	0.51	0.51	0.25	0.5	**0.98**	**0.98**	**0.91**	**0.92**	0.92	0.93	0.84	0.84	0.91	0.92	0.83	0.83
Glide 100 10 [48]	0.04	0.31	0.59	0.42	0.51	0.51	0.25	0.5	**0.98**	**0.98**	**0.92**	**0.92**	0.93	0.93	0.85	0.85	0.91	0.92	0.83	0.83
ADM [20]	0.45	0.43	0.65	0.56	0.56	0.57	0.39	0.56	0.47	0.46	0.65	0.55	0.62	0.57	0.32	0.53	**0.89**	**0.88**	**0.84**	**0.84**
DALL·E 3 [6]	0.35	0.41	0.61	0.45	0.51	0.51	0.25	0.5	0.45	0.5	0.62	0.47	0.82	0.84	0.71	0.74	**0.96**	**0.96**	**0.88**	**0.89**
Midjourney [45]	0.29	0.37	0.6	0.43	0.65	0.66	0.48	0.6	**0.92**	**0.93**	0.84	0.84	0.56	0.53	0.27	0.51	0.9	0.84	**0.85**	**0.85**
VDQM [28]	0.51	0.47	0.66	0.57	0.53	0.54	0.32	0.53	0.46	0.46	0.63	0.49	0.83	0.83	0.7	0.73	**0.92**	**0.93**	**0.84**	**0.84**
Wukong [46]	0.48	0.47	0.62	0.49	0.63	0.6	0.37	0.55	0.39	0.45	0.6	0.44	0.73	0.69	0.5	0.62	**0.95**	**0.96**	**0.88**	**0.88**
All	0.52	0.48	0.64	0.53	0.51	0.53	0.28	0.52	0.69	0.66	0.69	0.62	0.85	0.88	0.76	0.78	**0.91**	**0.91**	**0.84**	**0.84**

- Our method **outperforms** all other methods on most models and all models together.

Artistic Images

- Other method's performance **decreases.**
- Our method **maintains high performance**



Generative Model	AEROBLADE [59]				RIGID [30]				ZED [18]				Manifold Bias [11]				Ours			
	AUC ↑	AP ↑	F1 ↑	Acc ↑	AUC ↑	AP ↑	F1 ↑	Acc ↑	AUC ↑	AP ↑	F1 ↑	Acc ↑	AUC ↑	AP ↑	F1 ↑	Acc ↑	AUC ↑	AP ↑	F1 ↑	Acc ↑
StyleGan3 [37]	0.81	0.84	0.72	0.69	0.32	0.39	0.06	0.46	0.38	0.41	0.63	0.5	0.96	0.96	0.9	0.9	**0.98**	**0.99**	**0.93**	**0.93**
SD2.1 [60]	0.57	0.6	0.61	0.48	0.77	0.75	0.52	0.64	0.17	0.34	0.61	0.44	0.91	0.9	0.83	0.84	**0.98**	**0.98**	**0.92**	**0.93**
SDXL [53]	0.62	0.69	0.6	0.46	0.64	0.6	0.3	0.54	0.33	0.43	0.61	0.44	0.83	0.81	0.71	0.75	**0.97**	**0.98**	**0.91**	**0.91**
AnimagineXL [43]	0.66	0.75	0.6	0.47	0.39	0.42	0.08	0.46	0.45	0.59	0.61	0.44	0.28	0.37	0.02	0.44	**0.76**	**0.83**	**0.66**	**0.55**
All	0.65	0.7	0.62	0.52	0.53	0.55	0.26	0.53	0.33	0.41	0.61	0.45	0.74	0.8	0.69	0.73	**0.92**	**0.92**	**0.83**	**0.82**

Damaged Car Images

- Other method's performance **decreases.**
- Our method **maintains high performance**

Generative Model	AEROBLADE [59]				RIGID [30]				ZED [18]				Manifold Bias [11]				Ours			
	AUC ↑	AP ↑	F1 ↑	Acc ↑	AUC ↑	AP ↑	F1 ↑	Acc ↑	AUC ↑	AP ↑	F1 ↑	Acc ↑	AUC ↑	AP ↑	F1 ↑	Acc ↑	AUC ↑	AP ↑	F1 ↑	Acc ↑
Kandinsky2.1 [58]	0.29	0.38	0.6	0.43	0.3	0.38	0.06	0.44	0.68	0.66	0.68	0.61	0.6	0.64	0.48	0.61	**0.97**	**0.98**	**0.93**	**0.93**
SDXL [53]	0.34	0.4	0.61	0.45	0.3	0.38	0.06	0.44	0.83	0.85	0.76	0.73	0.32	0.42	0.15	0.49	**0.97**	**0.97**	**0.91**	**0.91**
DALL·E 3 [6]	0.39	0.45	0.6	0.44	0.4	0.42	0.11	0.42	0.15	0.33	0.59	0.42	0.14	0.33	0.03	0.44	**0.99**	**0.99**	**0.93**	**0.93**
Midjourney [45]	0.46	0.47	0.62	0.47	0.51	0.5	0.23	0.5	0.78	0.75	0.74	0.71	0.4	0.44	0.13	0.48	**0.93**	**0.94**	**0.87**	**0.86**
UnCLIP [57]	0.22	0.35	0.6	0.43	0.69	0.67	0.46	0.6	0.31	0.38	0.61	0.46	0.69	0.68	0.48	0.61	**0.8**	**0.8**	**0.72**	**0.67**
All	0.33	0.4	0.61	0.44	0.45	0.48	0.21	0.49	0.61	0.54	0.68	0.61	0.47	0.53	0.3	0.53	**0.92**	**0.92**	**0.85**	**0.84**

Invoice Images

Real Document

Different price

Request – total amount – 1000$

Different date

Request – Jan 2/4, 2012

Different name

Request – name – John Doe

Generative Model	AEROBLADE [59]				RIGID [30]				ZED [18]				Manifold Bias [11]				Ours			
	AUC ↑	AP ↑	F1 ↑	Acc ↑	AUC ↑	AP ↑	F1 ↑	Acc ↑	AUC ↑	AP ↑	F1 ↑	Acc ↑	AUC ↑	AP ↑	F1 ↑	Acc ↑	AUC ↑	AP ↑	F1 ↑	Acc ↑
GPT-Image-1 [55]	0.53	0.62	0.56	0.47	0.57	0.67	0.12	0.53	0.6	0.56	0.65	0.54	0.61	0.58	0.55	0.62	**0.91**	**0.92**	**0.82**	**0.82**

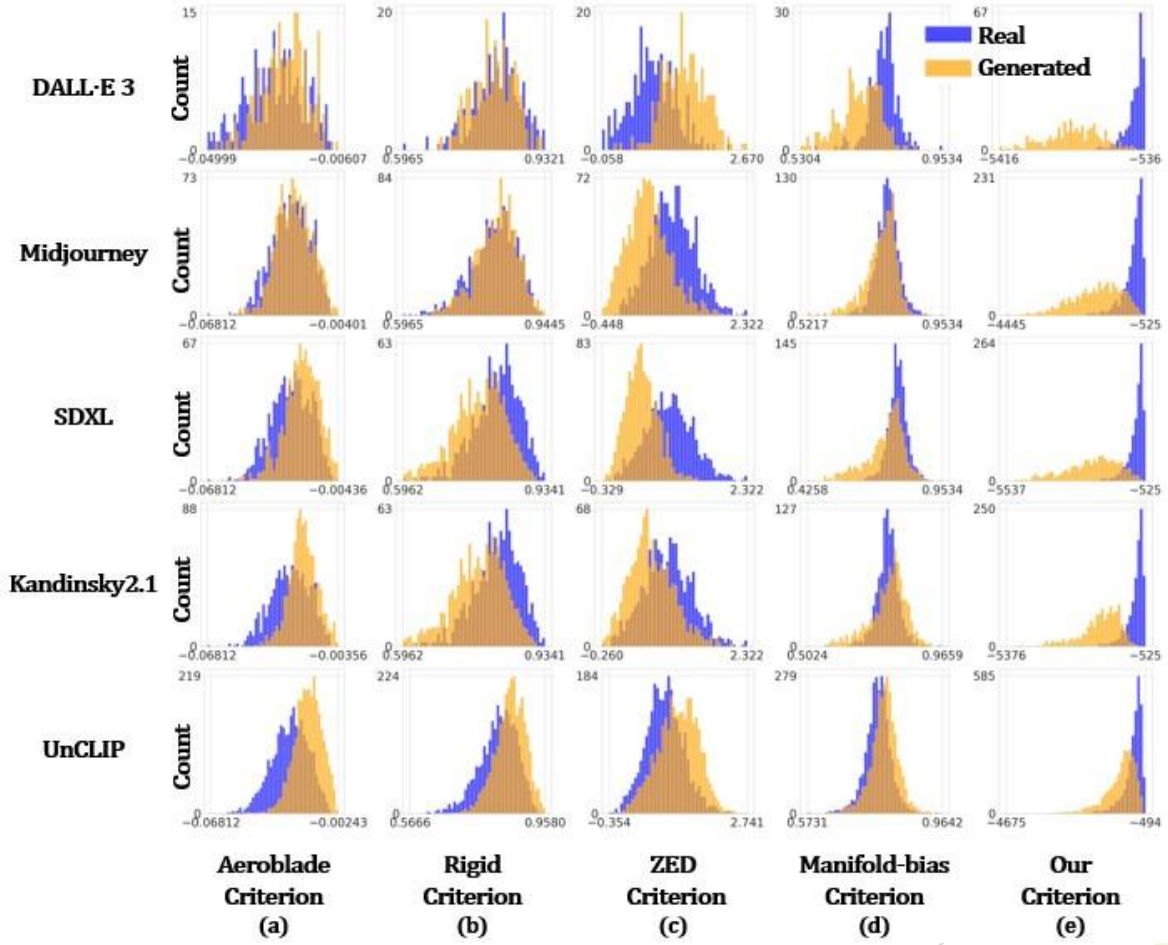
Separation & “Flipping” Phenomena

Other methods:

- Poor separation.
- Distribution “flips” sides for different models.

Our Method:

- Strong separation.
- Consistent sides relations between real and generated images.



Conclusions

- Conditional likelihood can be used to detect generated images in a zero-shot setting:
  - A domain-adaptive detection method.
  - CLIDE outperforms all SOTA methods.
  - Consistent relations between real and generated distributions.

Thank you

Roy Betser – roybe@campus.tehnion.ac.il
roy.betser@fujitsu.com