# Thematic session

Paper presentation: Group2

MultiX

# Zeno Geradts

# Experience with retraction of papers based on suspected AI generation as Chief Editor FSI Digital Investigation

# case

One of the registered reviewers in this journal, states that it observes an extremely high rate and the possibility of Generative AI text (please refer to "image.png" attached). Additionally, if one reads through the writing, it bears the distinctive style of Generative AI writing.

https://doi.org/10.1016/j.fsidi.2024.301749

https://doi.org/10.1016/j.fsidi.2024.301800

https://doi.org/10.1016/j.fsidi.2024.301745

# Some proof

Upon receiving your email, I conducted an analysis using ZeroGPT. To test the accuracy of this tool, I input the concluding sentences of your last email:

**"Please note that if we do not have an adequate and timely response, we may be forced to conclude that the allegations are truthful. I look forward to hearing from you soon. Yours sincerely,"**

The analysis indicated that approximately **24%** of this content was **flagged as AI-generated (Screenshot attached below).** This suggests that AI detection tools, which operate based on preset data and machine learning algorithms, may yield false positives, especially when language is good and well structured.

# Author confesses

I assure you of my unwavering commitment to upholding the highest ethical standards in my research and writing. I have taken this experience as a valuable lesson and will ensure complete transparency in all future endeavors.

View PDF Download full issue

Add to Mendeley Share Cite

https://doi.org/10.1016/j.fsidi.2024.301860

Full text access

Refers to

RETRACTED: Leveraging metadata in social media forensic investigations: Unravelling digital clues- A survey study

Forensic Science International: Digital Investigation, Volume 50, September 2024, Pages 301798

Akarshan Suryal

View PDF

This article has been retracted: please see Elsevier Policy on Article Withdrawal (https://www.elsevier.com/about/policies/article-withdrawal).

This article has been retracted following an allegation that raises concerns this article may have been generated by Generative AI.

# Case 2

I was recently looking through FSI:DI for updated materials on Windows forensics and noticed that a retraction was made due to the alleged use of Generative AI.

See:  Suryal, Retraction notice to "Leveraging metadata in social media forensic investigations: Unravelling digital clues- A survey study" [Forensic Sci. Int.: Digit. Invest. 50 (2024) 301798], Forensic Science International: Digital Investigation, Volume 52, 2025, 301860, ISSN 2666-2817, https://doi.org/10.1016/j.fsidi.2024.301860. (https://www.sciencedirect.com/science/article/pii/S2666281724001872

# Case 2

I was recently looking through FSI:DI for updated materials on Windows forensics and noticed that a retraction was made due to the alleged use of Generative AI.

See:  Suryal, Retraction notice to "Leveraging metadata in social media forensic investigations: Unravelling digital clues- A survey study" [Forensic Sci. Int.: Digit. Invest. 50 (2024) 301798], Forensic Science International: Digital Investigation, Volume 52, 2025, 301860, ISSN 2666-2817, https://doi.org/10.1016/j.fsidi.2024.301860. (https://www.sciencedirect.com/science/article/pii/S2666281724001872

Coincidentally, I was examining another paper which on the surface appeared to be relevant to an ongoing matter I have currently awaiting trial.  I am afraid I must report that there appears to be several linguistic markers which may place the paper into the same category.

See:  Soni, Manpreet Kaur, Khalid Aziz, Decoding digital interactions: An extensive study of TeamViewer's Forensic Artifacts across Windows and android platforms, Forensic Science International: Digital Investigation, Volume 51, 2024, 301838, ISSN 2666-2817, https://doi.org/10.1016/j.fsidi.2024.301838. (https://www.sciencedirect.com/science/article/pii/S2666281724001653)

I base my assessment on linguistic patterns that have been noted as appearing more frequently in Generative AI works, most notably those produced with GPT-3.5 like models, which appear within a sentence in an irregular manner. These phrases include:

"Delve" (first word, first point in highlights)

"The pervasive influence" (first line of abstract)

"In the rapidly evolving digital landscape" (first line of introduction)

"<noun> stands as <verb> in the realm of <noun>" (first line second paragraph of introdcution)

"In the realm of <noun>" (first line first paragraph of Related research)

"Before delving" (first line second paragraph of Methodology)

claiming "significant milestone" in conclusion

"endeavours" (first line in future work)

I note that the authors are listed as from the same institution being the "Lovely Professional University" located in Phagwara, Punjab India.

While these linguistic patterns may be explained through poor langauge understanding, ie. English as a second language, and perhaps even a cultural shift of Generative AI itself now creating a bias in how non-english speakers view the langauge, given the above retraction combined with the circumstantial issue presented that the authors share the same institution, I would recommend further investigation.

I would note that the "Lovely Professional University" has been embroiled in several integrity scandals recently regarding improper peer review process and mishandling of conflicts of interest (via Retraction Watch) and note that since approximately 2021 there has been 53 retractions noted in the public database. Please note that I am on a plane right now and unable to sort by date but have simply searched for appearances of "Lovely Professional University" in the csv as text and assumed the last entry is roughly the last date. However, if this is correct that means they are being retracted at a rate of ~13 per annum!

ISSN 2692-465X

The Retraction Watch Database

PLEASE NOTE: The best way to make use of this database is to download it from here, thanks to the acquisition of the database by Crossref. We have also made changes to the search functionality to maintain reliability. For more information, read our User Guide.

Please consider a U.S. tax-deductible donation to help us to continue to update and improve this important resource.

Author(s): Type to search
Country(s):
Title: Type to search
Article Type(s):
Journal:
Publisher:
Affiliation(s): Lovely Professional University

Original Paper
From Date: To:
PubMedID: mm/dd/yyyy
DOI:

Retraction or Other Notices
From Date: To:
PubMedID: mm/dd/yyyy
DOI:

Clear Search

Search

Your search returned a large number of results. Only 50 are displayed. Narrow your search to view all results

| Retraction or Other Notices<br>Title/Subject(s)/Journal --- Publisher/Affiliation(s)/Retraction Watch Post URL(s)<br>50 Items Displayed Out of 52 Item(s) Found | Reason(s) | Author(s) | Original Paper Date/PubMedID/DOI | Retraction or Other Notic Date/PubMedID/DOI |
|---|---|---|---|---|
| Optimization of lycorine using Response Surface Methodology, extraction methods and in vitro antioxidant and anti-diabetic activities from the roots of Giant Spider Lily: A medicinally important bulbous herb<br>(BLS) Biochemistry; (BLS) Plant Biology/Botany;<br>*South African Journal of Botany --- Elsevier*<br>Plant Physiology Laboratory, Department of Botany, Shivaji University, Kolhapur, Maharashtra 416004, India<br>Department of Biotechnology, Lovely Professional University, Phagwara 144411, | +Concerns/Issues with Peer Review<br>+Conflict of Interest<br>+Investigation by Journal/Publisher<br>+Rogue Editor | S B Patel<br>S S Otari<br>Vijay Kumar<br>Anshu Rastogi<br>M M Lekhak<br>S G Ghane | 09/15/2022<br>00000000<br>10.1016/j.sajb.2022.04.022 | 02/21/2025<br>00000000<br>10.1016/j.sajb.2024.11.035 |

# Elina (Eleni) Sergidou

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

## From data to a validated score-based LR system: A practitioner's guide

Anna Jeannette Leegwater [a], Peter Vergeer [a], Ivo Alberink [a], Leen V. van der Ham [a], Judith van de Wetering [a], Rachid El Harchaoui [a], Wauter Bosma [a], Rolf J.F. Ypma [a], Marjan J. Sjerps [a,b,*]

[a] *Netherlands Forensic Institute, PO Box 24044, The Hague 2490 AA, the Netherlands*
[b] *Korteweg-de Vries Institute for mathematics, University of Amsterdam, PO Box 94248, Amsterdam 1090 GE, the Netherlands*

# Meike Kombrink

*Article*

## Universal Image Vaccine Against Steganography

**Shiyu Wei, Zichi Wang * and Xinpeng Zhang**

# Idea: make it impossible to NOT detect stego

And do so before someone can use the image for steganography

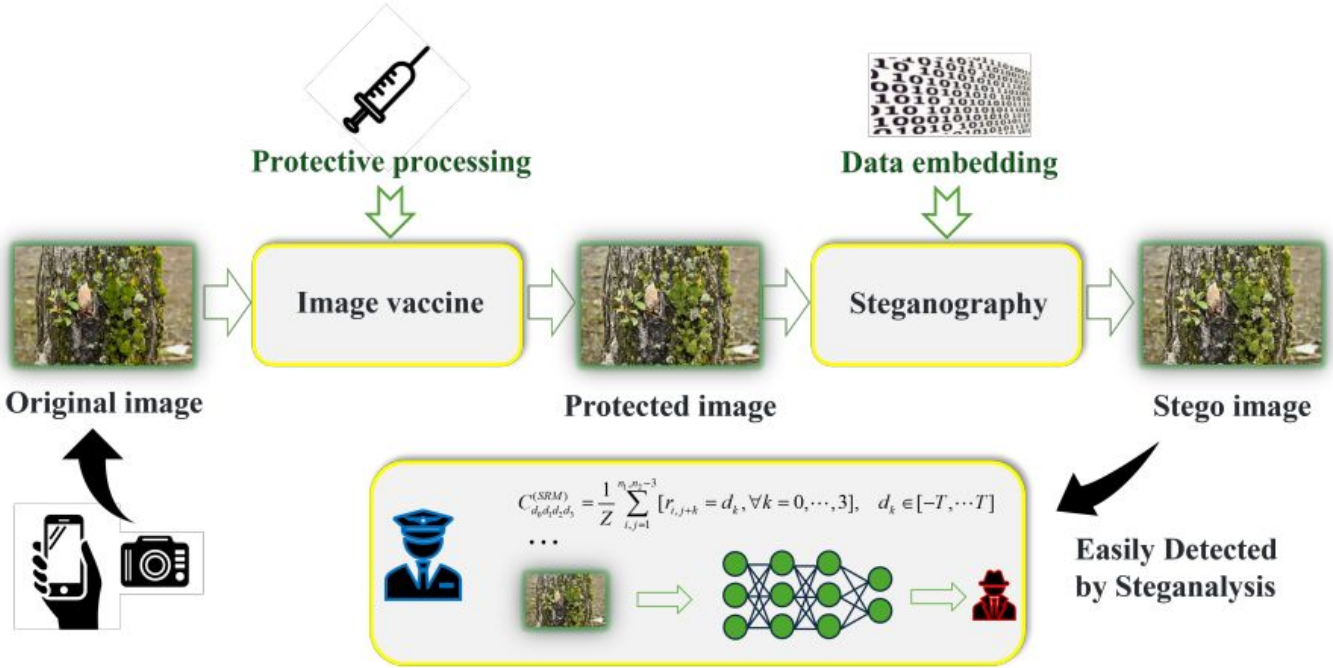
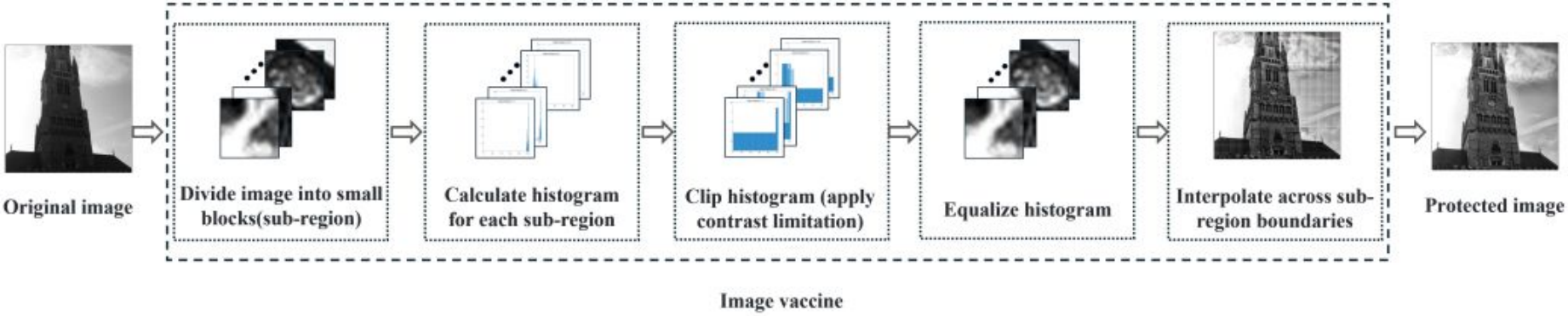



**Figure 1.** Image vaccine protection scheme against steganography.

# Goal: Make a UNIVERSAL vaccine



# Works better than Histogram Equalization

(a) Original-image1

(b) HE-image1

(c) CLAHE-image1

(d) Original-image2

(e) HE-image2

(f) CLAHE-image2

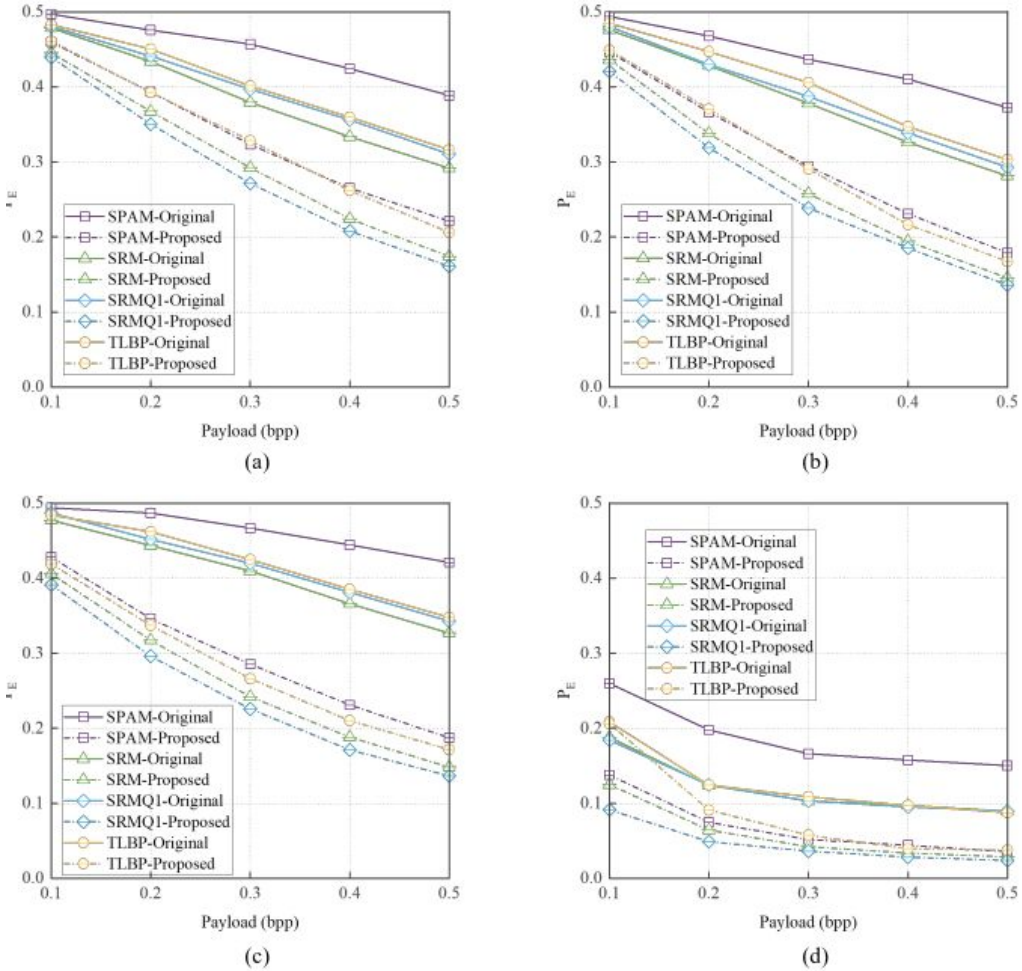



**Figure 4.** Comparison of our method with traditional steganalytic methods against various steganography techniques on BOWS2: (**a**) WOW, (**b**) SUNIWARD, (**c**) MiPOD, and (**d**) UT-GAN.

# My thoughts

Love the idea/concept

But a LARGE different in the image is seen, not sure that is desirable

# Conor McCarthy

## Training Language Models for Social Deduction with Multi-Agent Reinforcement Learning

Bidipta Sarkar
Stanford University
Stanford, United States of America
bidiptas@cs.stanford.edu

Warren Xia
Stanford University
Stanford, United States of America
waxia@cs.stanford.edu

C. Karen Liu
Stanford University
Stanford, United States of America
karenliu@cs.stanford.edu

Dorsa Sadigh
Stanford University
Stanford, United States of America
dorsa@cs.stanford.edu

Sarkar, B., Xia, W., Liu, C. K., & Sadigh, D. (2025). Training Language Models for Social Deduction with Multi-Agent Reinforcement Learning. *arXiv preprint arXiv:2502.06060*.

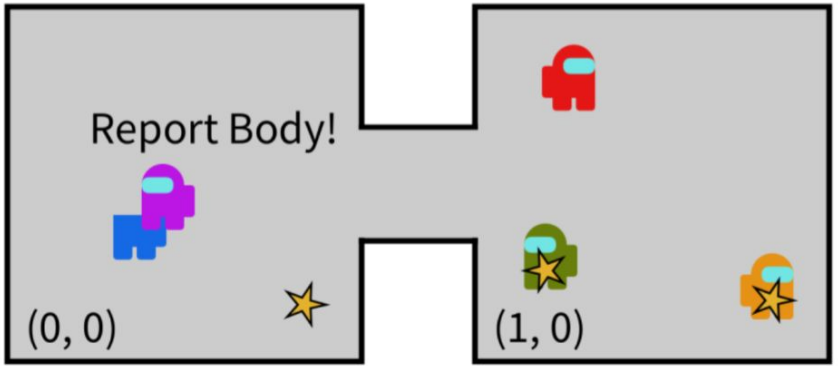
Chicago

- Multi-agent (LLM) settings require agents to communicate in a shared language
- Especially in “partially observable” settings, sharing knowledge is key
- Agents must share and parse messages

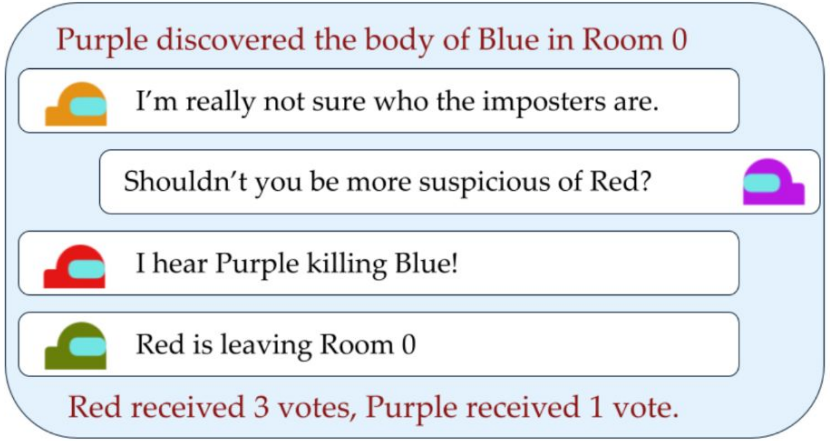
INNERSLOTH PRESENTS

# AMONG US

## Gameplay Phase




## Discussion Phase



- Multi Agent RL (MARL) requires large datasets for settings requiring natural language communication
- Sparse reward signal replaced with dense “Imposter Belief” reward signal
- “Speaking” - rewarded for changing other crewmates’ beliefs about imposter
- “Listening” - rewarded for changing own beliefs about imposter
- RWKV language model [2]

[2] Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Gv, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan Wind, Stanisław Woźniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. 2023. RWKV: Reinventing RNNs for the Transformer Era. In Findings of the Association for Computational Linguistics: EMNLP 2023, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 14048–14077. https://doi.org/10.18653/v1/2023.findings-emnlp.936

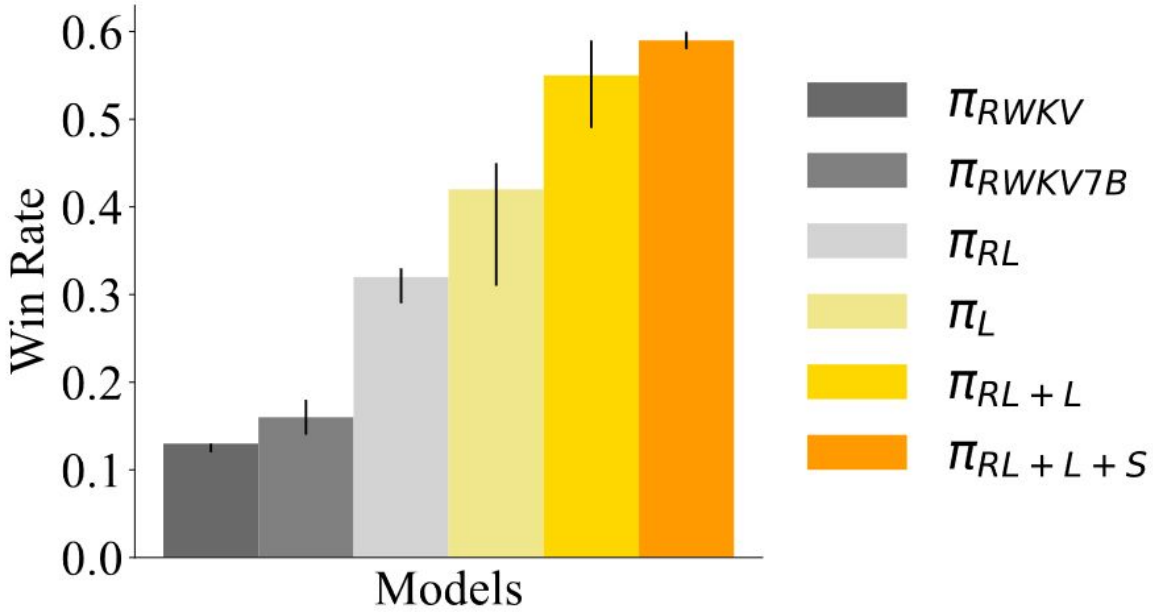
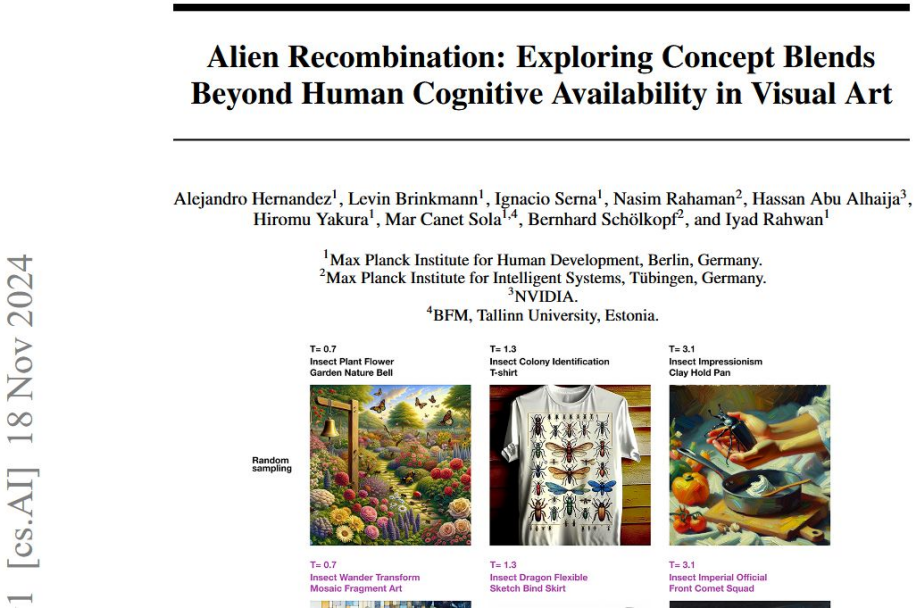



**Figure 3: Win rates for crewmates trained with different algorithms over the "base" environment: $2 \times 2$ grid of rooms, 4 tasks per crewmate, and 5 players. Error bars represent the maximum and minimum expected win rates across the three independently trained runs with different seeds.**

# Thanos Efthymiou



**Alien Recombination: Exploring Concept Blends Beyond Human Cognitive Availability in Visual Art**

Alejandro Hernandez[1], Levin Brinkmann[1], Ignacio Serna[1], Nasim Rahaman[2], Hassan Abu Alhaija[3], Hiromu Yakura[1], Mar Canet Sola[1,4], Bernhard Schölkopf[2], and Iyad Rahwan[1]

[1]Max Planck Institute for Human Development, Berlin, Germany.
[2]Max Planck Institute for Intelligent Systems, Tübingen, Germany.
[3]NVIDIA.
[4]BFM, Tallinn University, Estonia.





Hernandez, A., Brinkmann, L., Serna, I., Rahaman, N., Abu Alhaija, H., Yakura, H., Canet Sola, M., Schölkopf, B., & Rahwan, I. (2024). *Alien Recombination: Exploring Concept Blends Beyond Human Cognitive Availability in Visual Art*. To appear in *NeurIPS 2024 Workshop on Creativity & Generative AI*.

# Alien Recombination - Motivation

- Human creativity is constrained by cultural, social, and cognitive limitations.
- Many possible concept combinations remain unexplored due to these biases.
- AI can transcend human limits to generate entirely new conceptual blends.
- Offers the potential for disruptive innovation in art and other creative fields.

# Alien Recombination - Method

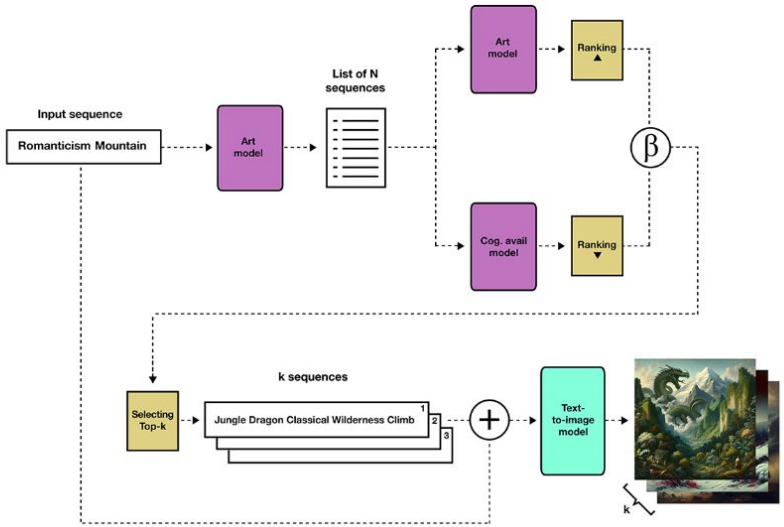



Figure 4: Schematic representation of the Alien Recombination method. The Art model generates N sequences from an input sequence. These sequences are ranked by both the Art model (ascending perplexity) and the Cognitive Availability model (descending perplexity). The ranking and selection process, termed "Alien sampling", employs a weighted rank aggregation method parameterized by $\beta$. Increasing $\beta$ prioritizes sequences that are more distant from what is cognitively available, thus enhancing *alieness*. The top-$k$ sequences (user-defined $k$) resulting from this fused ranking are then processed by a text-to-image model (DALL-E [10] in this study) to generate images, using the prompt: `A painting that contains the concepts: <input sequence + generated sequence>`.

# A

T=1.9

Classical Landscape
Giant Forest Sunrise

T=2.2

Tower Cottage Restoration
Wilderness Island

T=2.5

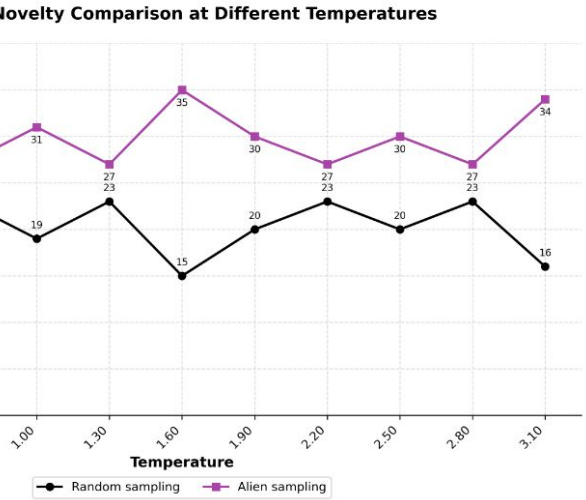
Quiet Mill Breeze
Folk Moment

T=2.8

Medieval River Lost
Glimpse Hunter

T=3.1

Ceremony Arrange
Civilization Indoor
Voyage




T=2.5

Comet Orbit Investigation
Scientist Exploration

T=2.8

Atmosphere Visitor Clock
Folklore Folk

T=3.1

Fire Burst Pour
Charge Sword

Alien sampling

# Alien Recombination - Takeaways

- AI can explore unexplored conceptual spaces, leading to novel creative outputs.
- Potential of AI to push artistic boundaries beyond human cognitive limits.
- Combines generative models with structured concept spaces.
- Encourages rethinking creativity as a combinatorial problem solvable by AI.

# Stijn van Lierop

## CLIPping the Deception: Adapting Vision-Language Models for Universal Deepfake Detection

Sohail Ahmed Khan
University of Bergen
Bergen, Norway
sohail.khan@uib.no

Duc-Tien Dang-Nguyen
University of Bergen
Bergen, Norway
ductien.dangnguyen@uib.no

Khan, Sohail Ahmed, and Duc-Tien Dang-Nguyen. “CLIPping the Deception: Adapting Vision-Language Models for Universal Deepfake Detection.” In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, 1006–15. Phuket Thailand: ACM, 2024. https://doi.org/10.1145/3652583.3658035.

# Problem & Idea

Background

- Problem: many detectors overfit on generator-specific artifacts
- Vision language models may capture subtle features of real images better
- But what is the optimal way to use them?
- Would incorporating text be of added value?

Key contributions

- What is the most effective transfer learning strategy?
  - Incorporate textual information as well
  - Robustness analysis
- Less training data needed compared to some SOTA methods

Khan, Sohail Ahmed, and Duc-Tien Dang-Nguyen. “CLIPping the Deception: Adapting Vision-Language Models for Universal Deepfake Detection.” In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, 1006–15. Phuket Thailand: ACM, 2024. https://doi.org/10.1145/3652583.3658035.

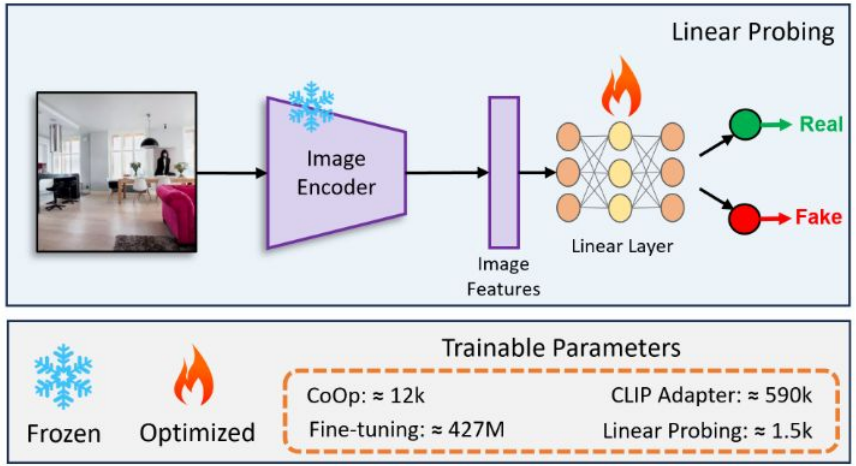



**Figure 2: In this figure, we present four distinct transfer learning strategies that are explored for $real/fake$ image classification. At bottom right we list the number of trainable parameters for each approach. Images from [13, 50].**

# Method

Training set: ProGAN / LSUN

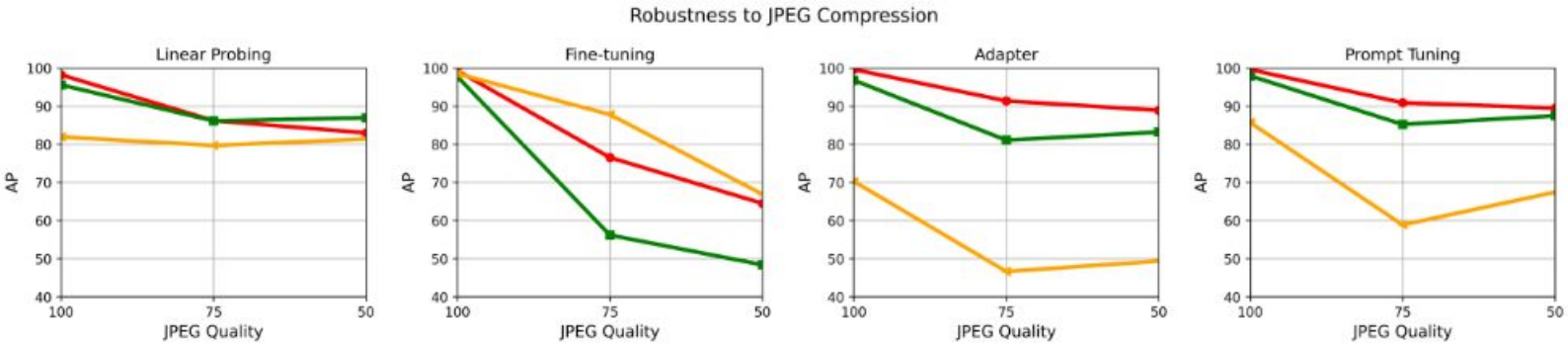
**Table 1: This table showcases the statistics of the test datasets. Certain datasets include their own collection of *real* images. However, for datasets that lack their own *real* images, we utilize LAION's [44] images instead.**

| Generator | Num. *real*/*fake* | Real Data Source | Image Resolution | Family |
|---|---|---|---|---|
| ProGAN [21] | 4k / 4k | LSUN | 256 x 256 | GAN |
| BigGAN [3] | 2k / 2k | ImageNet | 256 x 256 | GAN |
| CycleGAN [52] | 1k / 1k | Various | 256 x 256 | GAN |
| EG3D [4] | 1k / 1k | LAION | 512 x 512 | GAN |
| GauGAN [36] | 5k / 5k | COCO | 256 x 256 | GAN |
| StarGAN [6] | 2k / 2k | CelebA | 256 x 256 | GAN |
| StyleGAN [24] | 1k / 1k | LSUN | 256 x 256 | GAN |
| StyleGAN2 [25] | 1k / 1k | Various | ≈ 256 x 256 | GAN |
| StyleGAN3 [23] | ≈ 1k / 1k | Various | 512 x 512 | GAN |
| Taming-T [12] | 1k / 1k | LAION | 256 x 256 | GAN |
| DALL-E (mini) [10] | 1k / 1k | LAION | 256 x 256 | - |
| Glide [31] | 1k / 1k | LAION | 256 x 256 | Diff. |
| Guided [34] | 1k / 1k | LAION | 256 x 256 | Diff. |
| LDM [40] | 1k / 1k | LAION | 256 x 256 | Diff. |
| Stable Diff. [40] | 1k / 1k | LAION | 512 x 512 | Diff. |
| SDXL [37] | 1k / 1k | LAION | 1024 x 1024 | Diff. |
| Deepfakes [41] | ≈ 2.7k / 2.7k | YouTube | ≈ 256 x 256 | - |
| FaceSwap [41] | 2.8k / 2.8k | YouTube | ≈ 256 x 256 | - |
| Midjourney-V5 | 1k / 1k | LAION | Various | Comm. |
| Adobe Firefly | 1k / 1k | LAION | Various | Comm. |
| DALL-E 3 | 1k / 1k | LAION | Various | Comm. |

Khan, Sohail Ahmed, and Duc-Tien Dang-Nguyen. "CLIPping the Deception: Adapting Vision-Language Models for Universal Deepfake Detection." In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, 1006–15. Phuket Thailand: ACM, 2024. https://doi.org/10.1145/3652583.3658035.

# Results

- In general: prompt-tuning works best overall
  - **+ 5.01% mAP** & **+6.61% average accuracy** compared to SOTA
- Training on more data does increase performance, but no significant differences
- However, linear probing approach seems to be more robust against perturbations




Khan, Sohail Ahmed, and Duc-Tien Dang-Nguyen. “CLIPping the Deception: Adapting Vision-Language Models for Universal Deepfake Detection.” In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, 1006–15. Phuket Thailand: ACM, 2024. https://doi.org/10.1145/3652583.3658035.