Explaining in Diffusion: Explaining a Classifier Through Hierarchical Semantics with Text-to-Image Diffusion Models

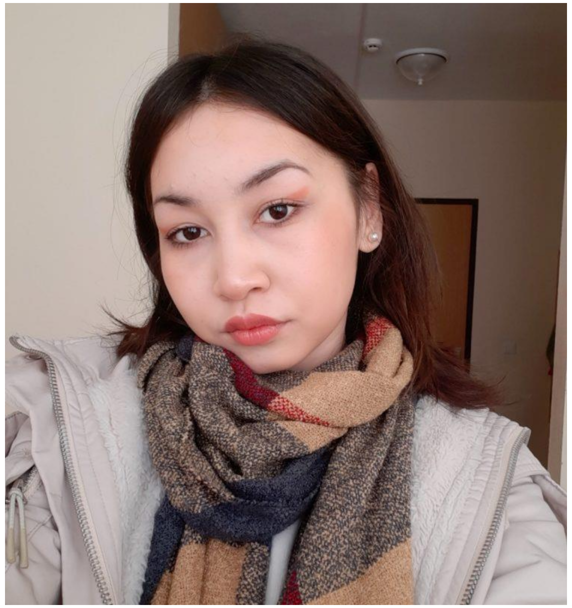
by Tahira Kazimi et al.

CVPR 2025 (poster)

Jakub Świstak

MI2.AI Seminar, Warsaw, May 5th, 2025

Authors



Tahira Kazimi

Ritika Allada

Pinar Yanardag

Key contributions

- **DiffEx**: a training-free, hierarchical explainer built on vision–language models and text-to-image diffusion.
- **Semantic corpus**: VLM-driven, multi-domain hierarchy of concepts for broad reuse.
- **Unified scope**: handles both single-concept (e.g. age) and complex-scene (e.g. architecture) classifiers without retraining.
- **Empirical gains**: delivers richer, more interpretable explanations than prior methods across binary and multiclass tasks (e.g. facial recognition, retinal health).

Motivation

-**Drawbacks of GAN-based counterfactuals** - Prior methods like StylEx require training a new GAN per classifier, depend on manual attribute labeling, and are confined to single-concept domains, making them resource-intensive and less scalable .

-Under-utilization of diffusion models: While recent work explores diffusion-based counterfactuals, existing approaches either produce semantically shallow edits, rely on domain-specific DDPMs, or incur high computational costs, failing to leverage large-scale latent diffusion models for complex scenes

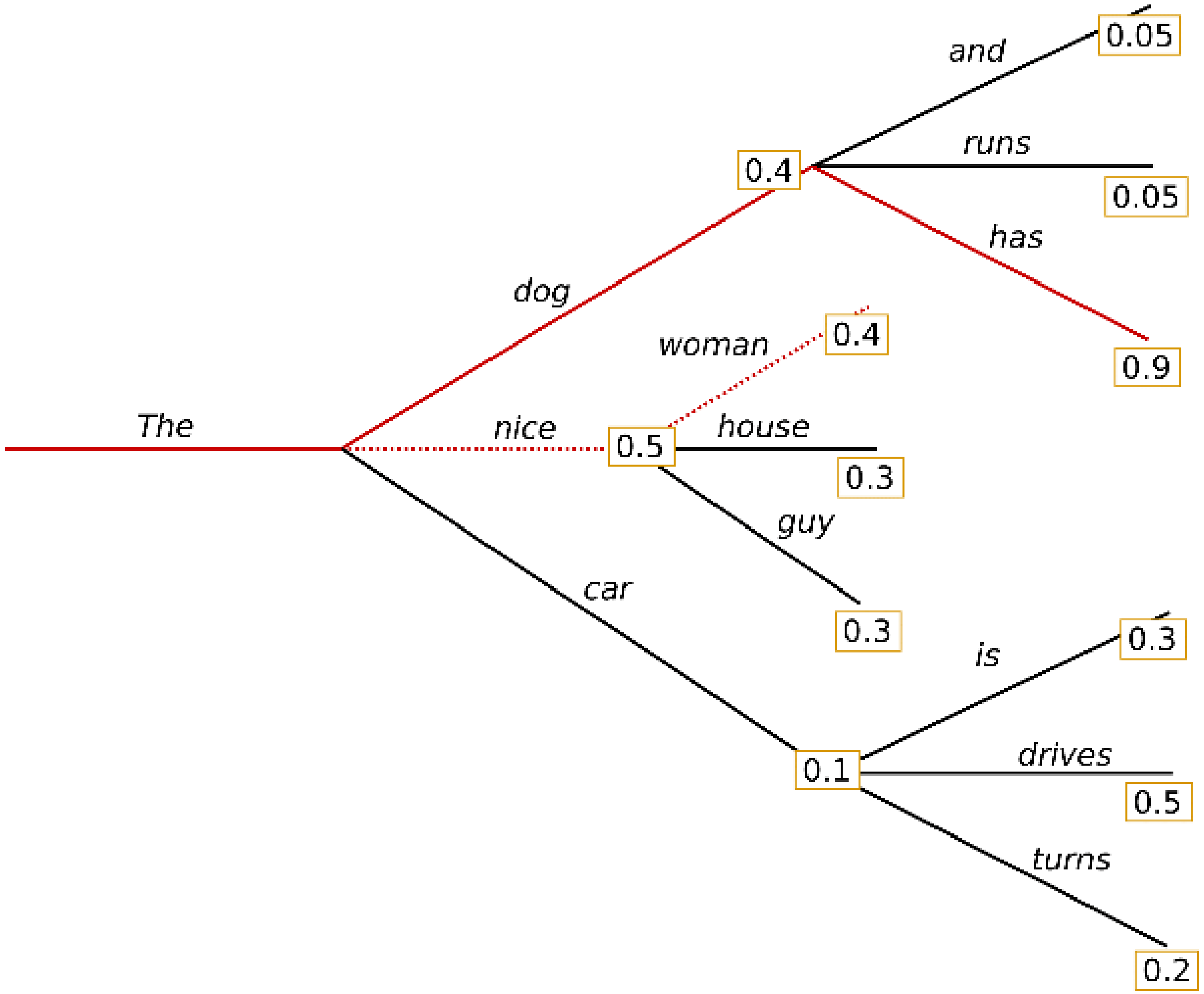
-Absence of hierarchical explanations: No prior research systematically unpacks how different semantic levels—from coarse attributes to fine subtypes—jointly influence classifier logits, leaving a gap in comprehensive, multi-level interpretability

-Need for automated, domain-agnostic semantics extraction: Reducing reliance on manual prompt engineering, the authors employ vision-language models to auto-construct a large-scale, hierarchical semantic corpus spanning diverse domains

-Desire for a training-free, unified approach: Motivated to explain both single-concept (e.g., facial age) and complex scene classifiers (e.g., urban vs. rural scenes) without retraining or domain-specific customizations

-Scalability and usability: By combining off-the-shelf diffusion editing tools (e.g., Ledits++) with a beam-search-inspired algorithm, the authors aim for an efficient, broadly applicable framework (DiffEx) that ranks and expands only the most impactful semantic features

Beam search


The
dog
0.4
and
0.05
runs
0.05
has
0.9
nice
0.5
woman
0.4
house
0.3
guy
0.3
car
0.1
is
0.3
drives
0.5
turns
0.2

DiffEx

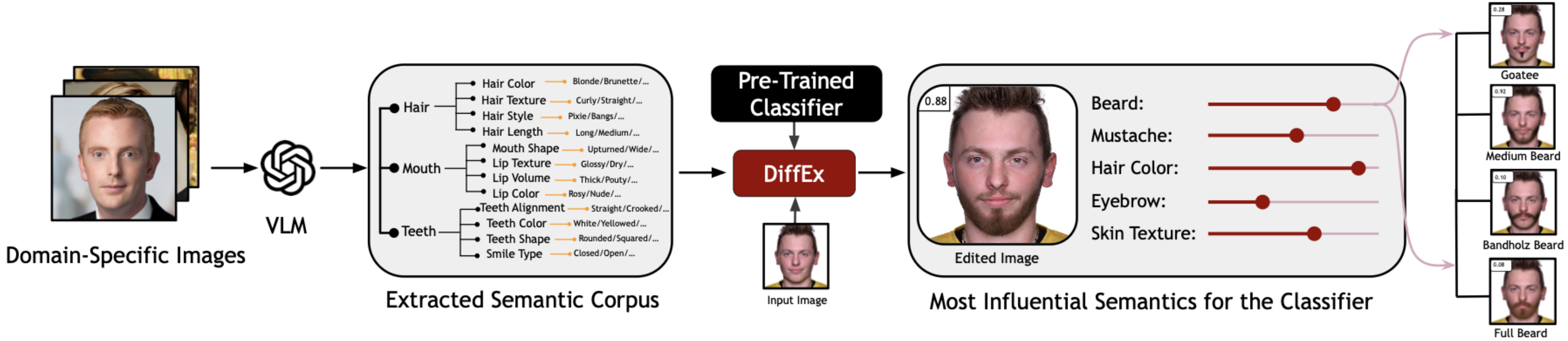


Figure 3. **An Overview of DiffEx.** Our pipeline processes a set of sample domain-specific images and a text prompt using a VLM to generate a hierarchical semantic corpus of attributes relevant to a specific domain. Based on this corpus, DiffEx identifies and ranks the most influential features affecting the classifier's decisions, sorting them from most to least impactful (rightmost image). The hierarchical explanation of semantics (such as beard and its subtypes) provides a fine-grained understanding of which features drive classifier outputs.

VLM prompt

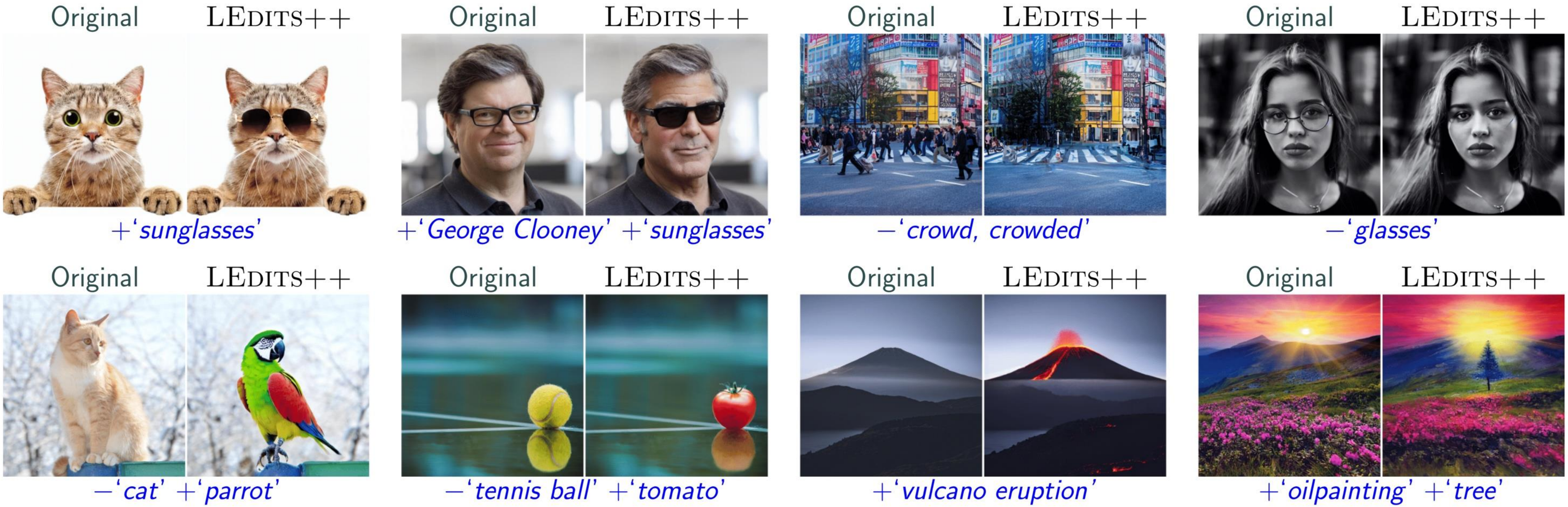
```
[
  {"role": "system",
   "content": ’You are an expert at finding features important for text-based
   image editing using diffusion models, given a set of images. Upon receiving
   a set of images, analyze the given inputs and extract important features and
   keywords that can be used for text-based image editing using diffusion models.
   Analyze the set of images and identify key features that define or are significant
   within the specified domain. These features are encoded to guide generative
   diffusion model for fine-grained image editing of subjects.
   List all different categories related to that specific feature. For example, for
       human features, it
   ranges from skin texture to expression, accessories, eyebrow shape, etc.
   Output must be in the format given, a sample output is given below, give the output
   only without any other descriptive text. Do not restrict your answers to the given
   sample, come up with all features. I want detailed fine-grained features.
[{
      "Face": {"oval face" , "rectangular face", "round face",}
      "Skin Texture": {"smooth skin", "freckled skin", "blemish skin", "scar skin"},
      "Skin Color": {"light colored skin", "dark colored skin"},
      "Eyes Shape": {"round eyes", "almond eyes"},
      "Eyes Color": {"blue colored eyes", "green colored eyes", "hazel colored eyes"},
      "Eyebrows": {"thin eyebrows", "bushy eyebrows"},
      "Hair Color": {"dark colored hair", "light colored hair", "blonde hair",
      "brunette hair",},
      "Hair Texture": {"straight hair", "curly hair", "wavy hair",},
      "Hair Length": {"short hair", "long hair", "medium hair"},
      "Nose Shape": {"button nose", "straight nose", "prominent nose",},
      "Mouth Shape": {"full lip", "thin lip"},
      "Lip Color": {"matte lip", "glossy lip",}
       "earrings", "necklace, glasses, sunglasses",
}]
```

Table H. **Face Domain Keyword-Extraction Prompt Used in GPT-4.** The text above shows the prompt we fed into the VLM in order to find potential attributes in the face domain.

```
[
  {"role": "system",
   "content": ’You are an expert at finding features important for text-based
   image editing using diffusion models, given a set of images. Upon receiving
   a set of images, analyze the given inputs and extract important features and
   keywords that can be used for text-based image editing using diffusion models.
   Analyze the set of images and identify key features that define or are significant
   within the specified domain. These features are encoded to guide generative
   diffusion model for fine-grained image editing of subjects.
   List all different categories related to that specific feature. For example, for
       DOMAIN_NAME features, it
   ranges from ATTRIBUTE_1 to ATTRIBUTE_2, ATTRIBUTE_3, ATTRIBUTE_4, etc.
   Output must be in the format given, a sample output is given below, give the output
   only without any other descriptive text. Do not restrict your answers to the given
   sample, come up with all features. I want detailed fine-grained features.
[{
     "ATTRIBUTE_1": {"sub_attribute_1_1" , "sub_attribute_1_2", "sub_attribute_1_3",}
     "ATTRIBUTE_2": {"sub_attribute_2_1", "sub_attribute_2_2", "sub_attribute_2_3"},
     "ATTRIBUTE_3": {"sub_attribute_3_1", "sub_attribute_3_2"},
     "ATTRIBUTE_4": {"sub_attribute_4_1", "sub_attribute_4_2"},
     "ATTRIBUTE_5": {"sub_attribute_5_1", "sub_attribute_5_2", "sub_attribute_5_3"},
}]
```

Table G. **Prompt Template for Keyword-Extraction.** The text above illustrates the standard format used to input text prompts into GPT-4 for extracting potential attributes across different domains. “DOMAIN_NAME” refers to a specific domain, such as facial features, bird species, etc. “ATTRIBUTE_1, ATTRIBUTE_2, etc.” refer to the Level 1 (broad) categories, while “sub_attribute_1_1, sub_attribute_1_2, etc.” refer to Level 2 (finer-grained) categories.

LEDITS++ Limitless Image Editing using Text-to-Image Models



DiffEx algorithm

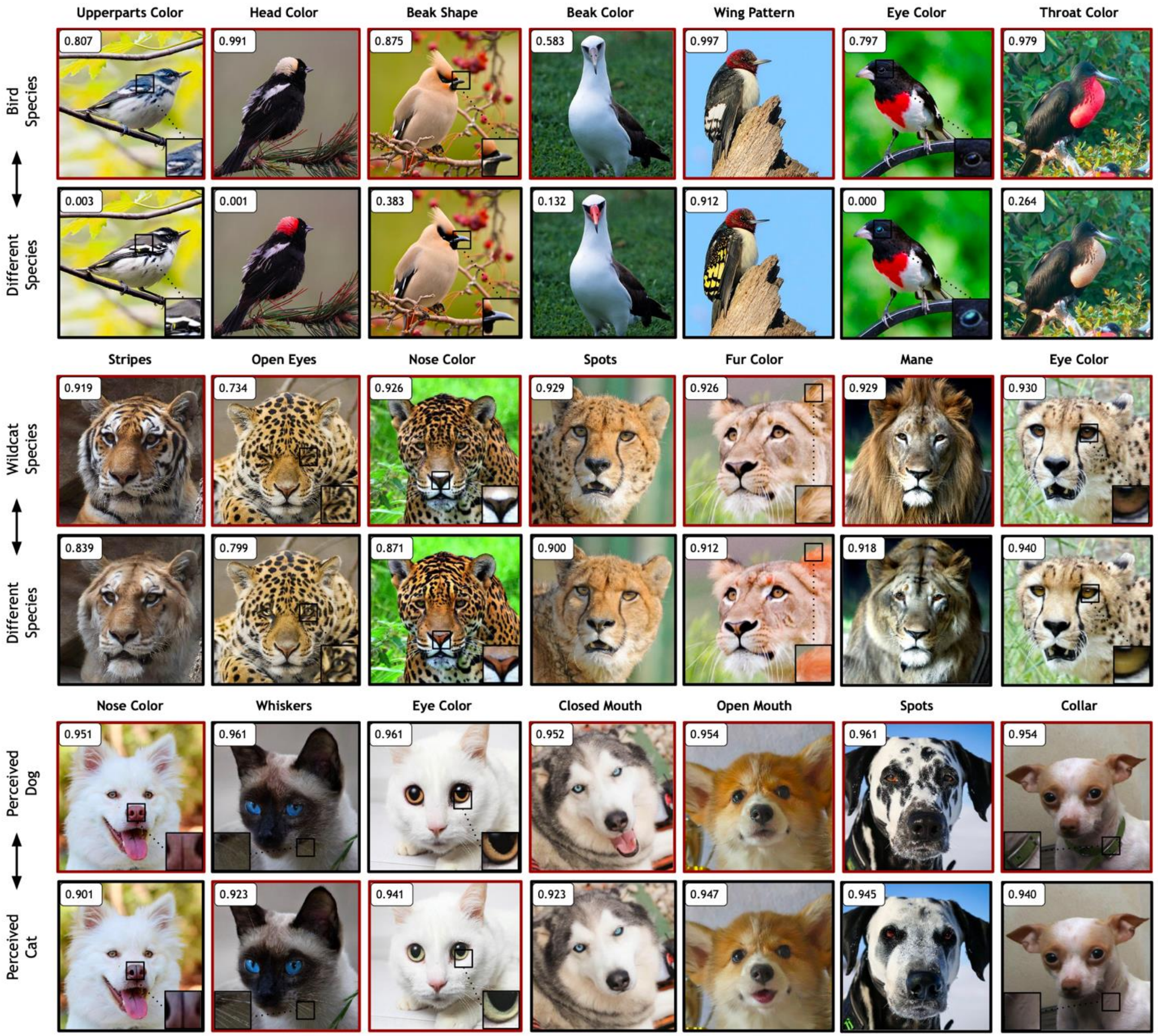
Algorithm 1 DiffEx

Require: Hierarchical structure $\mathcal{H}$ with semantic groups and features, beam width B, classifier or scoring function f, scoring threshold δ

Ensure: Optimal semantics maximizing f

Initialize $S \leftarrow$ root-level groups in $\mathcal{H}$ {Initial candidate set at top-level groups}
Initialize beam $\mathcal{B} \leftarrow \emptyset$
Score each candidate $s \in S$ using the scoring function $f(s)$
Select top B candidates with $f(b) \geq \delta$ and store in beam $\mathcal{B}$ {Apply thresholding to filter relevant candidates}
while $S \neq \emptyset$ **do**
    Initialize $S_{\text{next}} \leftarrow \emptyset$
    for each candidate $b \in \mathcal{B}$ **do**
        Expand b by adding sub-features from its next level in $\mathcal{H}$ to form new candidates
        for each new combination b' generated from b **do**
            if $f(b') > f(b)$ **then**
                Add b' to S_{next}
            end if
        end for
    end for
    Set $S \leftarrow S_{\text{next}}$ {Move to next level in hierarchy}
end while
Return highest-scoring combination from final $\mathcal{B}$ as the optimal joint semantic combination

Top 7 discovered attributes across different animal domains



StylEx vs DiffEx comparison

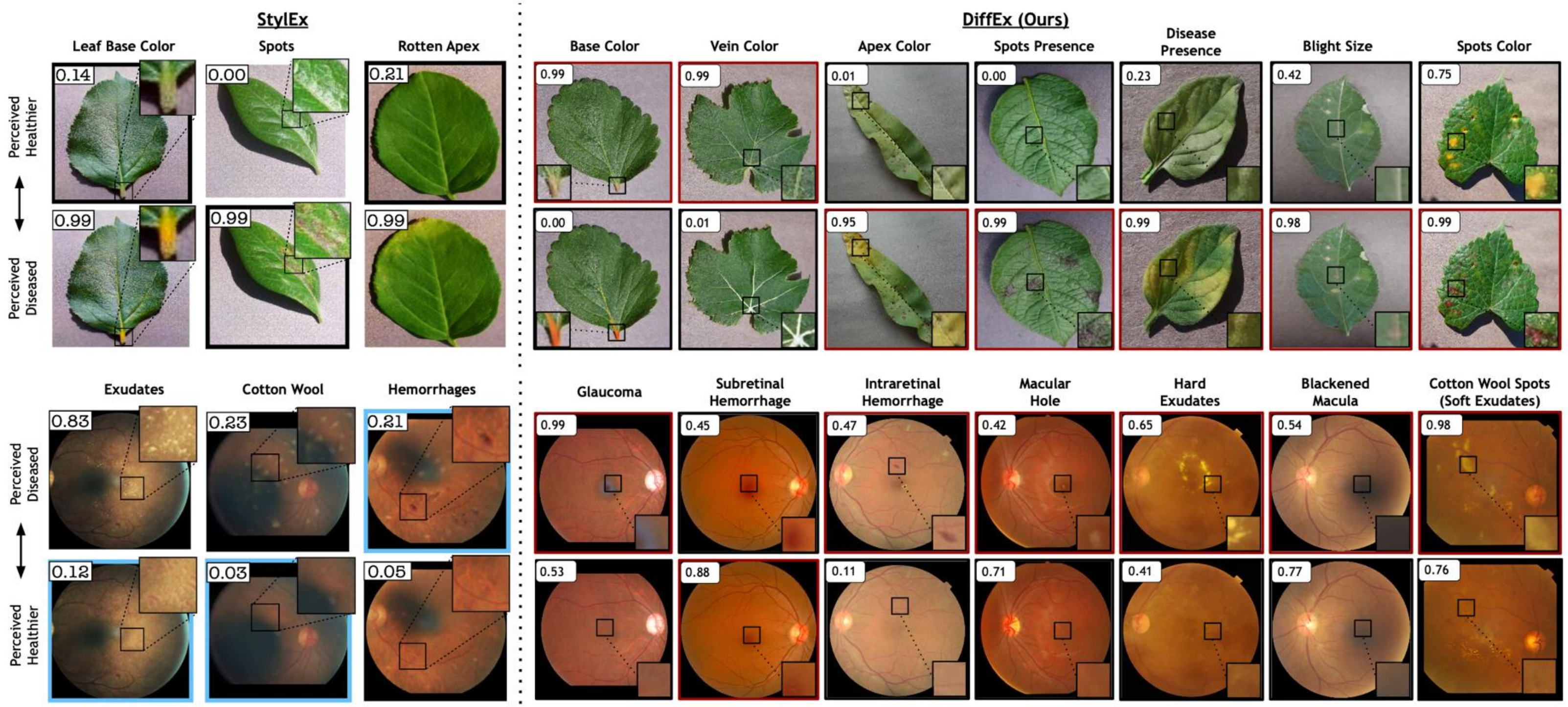
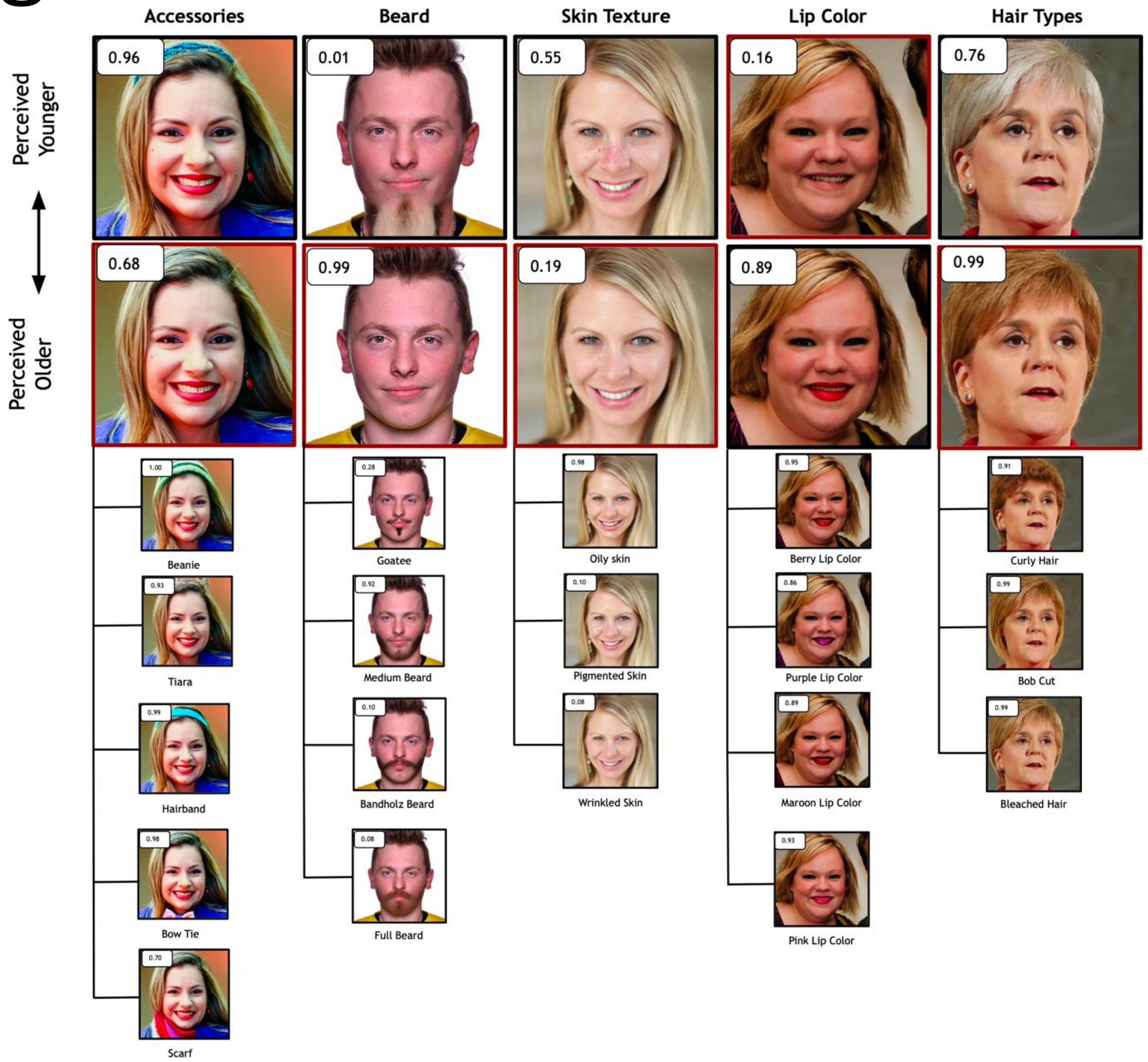
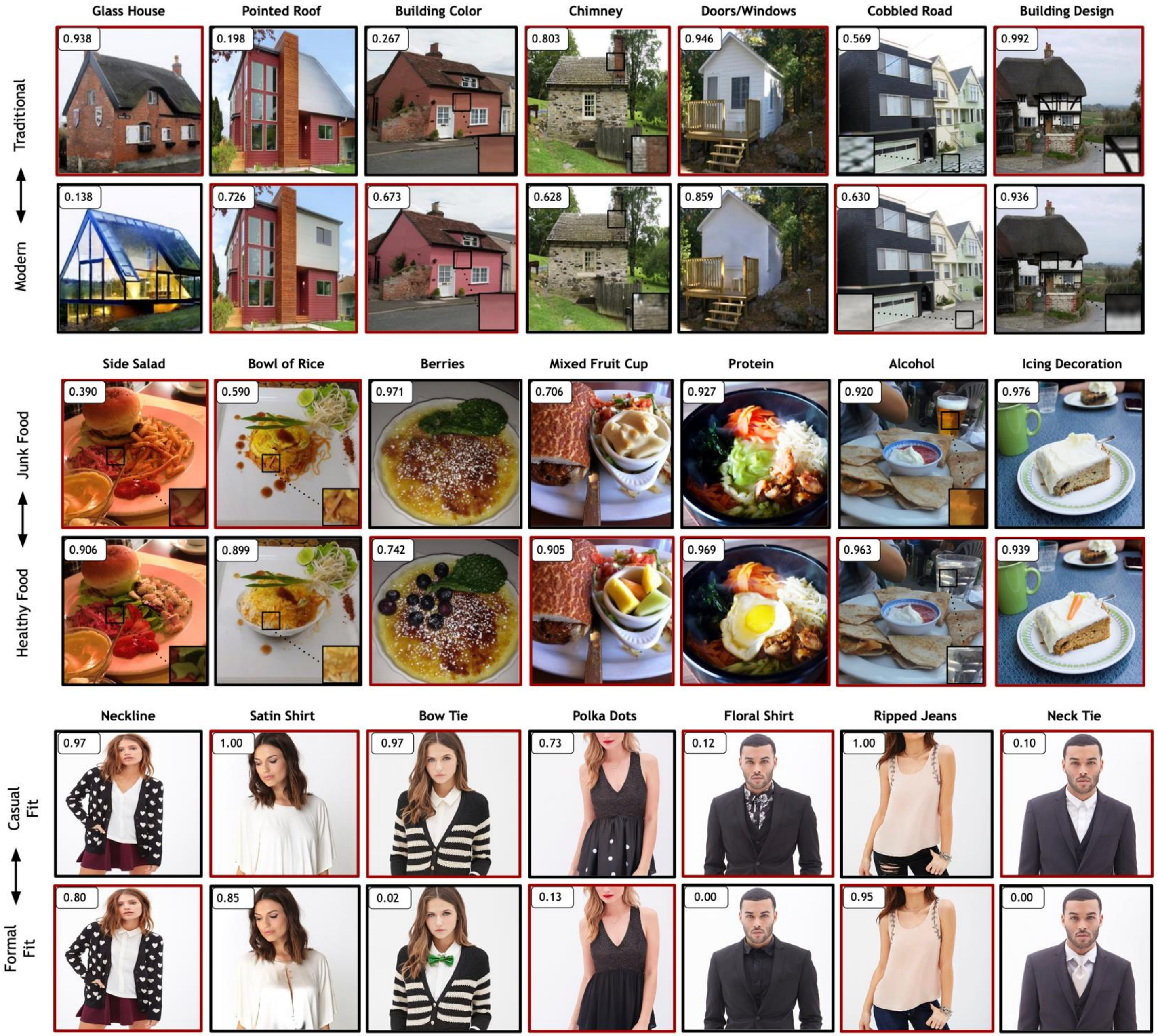


Figure 6. **Visual Comparison of Key Attributes Identified by StylEx and DiffEx in the Plant Health and Retinal Disease Domains.** This figure illustrates the enhanced capability of our method (DiffEx) in identifying a broader set of significant attributes compared to StylEx within the plant health and retinal disease domains. DiffEx successfully uncovers more detailed and diagnostically relevant features, such as "leaf vein color" and "macular hole," which provide deeper insights into leaf and retina health. In contrast, StylEx primarily identifies general attributes like "leaf base color" and "exudates." For DiffEx, images with black borders represent the counterfactual images, while those with red borders represent the original images. For StylEx, images with blue or black borders are counterfactuals. For a comprehensive comparison of the top attributes discovered by StylEx and DiffEx across various domains, please refer to Table 1.

Hierarchical structure of facial attributes and their impact on age classifier score



Handling Complex Scenes with DiffEx



Top attributes across different domains

Face (Age)		Bird (Species)		Leaves (Health)		Retina Scans (Disease)		Wildcat (Species)		Pet (Cat/Dog)	
StylEx	**Ours**	**StylEx**	**Ours**	**StylEx**	**Ours**	**StylEx**	**Ours**	**StylEx**	**Ours**	**StylEx**	**Ours**
Skin Pigmentation	Eyebrow	Belly Color	Upperparts Color	Base Leaf Color	Base Color	Exudates	Glaucoma	Spots	Stripes	Open Mouth	Nose Color
Eyebrow Thickness	Makeup	Upperparts Color	Head Color	Apex Color	Vein Color	Cotton Wool Spots	Subretinal Hemorrhage	Black Tear Mark	Open Eyes	Closed Mouth	Whiskers
Eyeglasses	Mustache Type	Wing Pattern	Beak Shape	Spots	Apex Color	Hemorrhages	Intraretinal Hemorrhage	Eye Shape + Size	Nose Color	Eye Shape	Eye Color
Hair Color	Teeth	Beak Color	Beak Color	Blight	Spots Presence	Clustered Exudates	Macular Hole	✗	Spots	Dropped Ears	Closed Mouth
Lip Thickness + Position	Lip Volume	Head Color	Wing Pattern	Halos	Disease Presence	✗	Hard Exudates	✗	Fur Color	Pointed Ears	Open Mouth
Bangs	Lip Color	Breast Color	Eye Color	✗	Blight Size	✗	Blackened Macula	✗	Mane	Eye Circumference	Spots
Eye Makeup	Eyelash	✗	Throat Color	✗	Leaf Texture	✗	Soft Exudates	✗	Eye Color	✗	Collar
Facial Hair Color	Beard Type	✗	Wing Color	✗	Spots Color	✗	Retinal Drusen	✗	Tongue	✗	Pointed Ears
✗	Facewear	✗	Crest Presence	✗	Discoloration	✗	Optic Disc Hemorrhage	✗	Pupil Size	✗	Mouth Color
✗	Headwear	✗	Feather Texture	✗	Leaf Orientation	✗	Cataract	✗	Whiskers	✗	Fur Pattern

Table 1. **Comparison of Top Attributes Across Different Domains and Classifiers.** The table above contains a list of the top attributes discovered by DiffEx (Ours) vs. StylEx. The ✗ in the table indicates attributes that were not mentioned in StylEx. It is also important to note that "cotton wool spots" and "soft exudates" refer to the same condition within the retinal disease domain.

Experiments

Rating	Bird Domain	Face Domain
Edit Quality	3.386 ± 0.223	3.659 ± 0.248
Disentanglement	3.163 ± 0.197	3.204 ± 0.213

Table 3. **Edit Quality and Disentanglement Ratings.** The table above provides the average edit quality and faithfulness ratings across different domains from User Study 1. The scoring is done on a scale from 1 to 5.

Method	Crest Presence	Beak Shape	Throat Color	Feather Texture	Eye Color	Beak Color	Head Color	Upperparts Color	Avg. Correct Response
Grad-CAM	36%	50%	56%	35%	47%	65%	59%	76%	53%
StylEx	68%	85%	79%	82%	74%	68%	91%	65%	76.5%
DiffEx (Ours)	**88%**	**91%**	**88%**	**91%**	**82%**	**82%**	**97%**	**88%**	**88.4%**

Table 2. **Comparison with Other Explainability Methods.** The table above displays the percentage of correct attribute selections for the bird class, as chosen by users when viewing outputs from different explainability methods. It also includes the average percentage of correct responses across all attributes for each method. As shown, for each attribute presented, the majority of users identified the correct attribute when viewing the output generated by DiffEx.

Limitations

-Dependence on VLM-curated semantic corpus, meaning quality and scope of the initial semantic data directly limit coverage
-Potential under-representation of specialized or context-specific features, which may be critical for accurate interpretation in niche domains
- Use of off-the-shelf image editing models (e.g. Ledits++) can produce entangled edits, introducing confounding factors that skew classifier scores
-Even minor unintended changes may undermine interpretability in high-stakes settings (e.g. medical imaging)
-Framework improvements hinge on integrating domain-specific semantic adjustments (e.g. task-specific RAGs) or alternative editing methods for robust performance

Thank You for your attention!

MI

Jakub Swistak

MI2.AI Seminar, Warsaw, May 5th, 2025