TIDE: Training Locally Interpretable Domain Generalization Models Enables Test-time Correction

Supplementary Material

Appendix A.

In Section A.1, we show examples where the proposed signature verification step does not detect incorrectly predicted classes. In Section A.2, we qualitatively compare against an additional baseline to highlight the limitations of only learning global class level features, while showing that our method effectively captures local concepts. In Section A.3, we show additional results for concept localization to show further evidence for how our model effectively captures key features across diverse contexts and styles. In Section A.4, we show additional samples of images annotated via our proposed pipeline across all domains from all datasets namely PACS [27], VLCS [13], OfficeHome [50], and DomainNet [38]. Finally, we provide details of prompting GPT3.5 [2] in Section A.5.

A.1. Signature Verification Errors

In this section, we provide examples where the model's misclassifications are not detected during the test-time signature matching step. We demonstrate that this is prevalent particularly when regions in the input image resemble concepts from other classes. For example, as shown in Figure 8, the first row/first column illustrates an image of an elephant from the sketch domain of the PACS dataset. The model incorrectly classifies the image as a guitar due to certain visual similarities: the elephant's `trunk`, with its line-like patterns, is interpreted as the `strings` of a guitar, and the unusual, dot-like appearance of the eyes is misidentified as guitar `knobs`. Similarly, additional examples illustrate this phenomenon: a dog, horse, and house are all misclassified as a person across various domains due to the presence of person-like concepts such as `eyes` and `lips`. Notably, the house from the cartoon domain features drawn elements resembling `eyes` and `lips`, which the model identifies as human facial features. Since these highlighted regions correspond to semantically plausible features for the predicted classes (guitar/person), the detection step does not flag this as an error, mistakenly believing the prediction to be correct.

A.2. Additional Baseline

We show comparisons with an additional baseline NJPP [59] in Figure 9. One can note that similar to the baseline ABA [5] demonstrated in Figure 1 in the main paper, NJPP too struggles to maintain consistent attention under domain shifts, often focusing on irrelevant regions (see first row). On the other hand, our method in the second and third rows accurately highlights stable local features like `beak` and `feathers`.

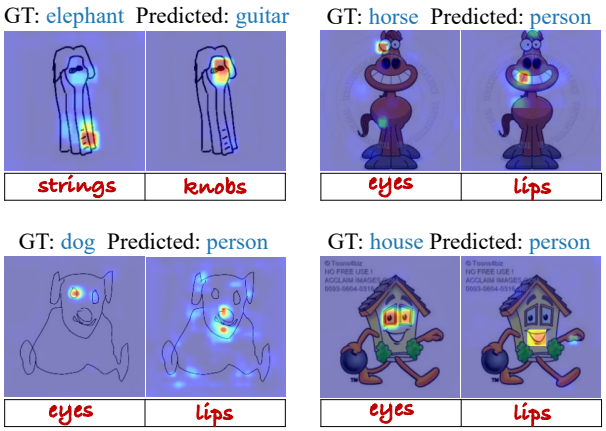

Figure 8. Signature Verification Errors.

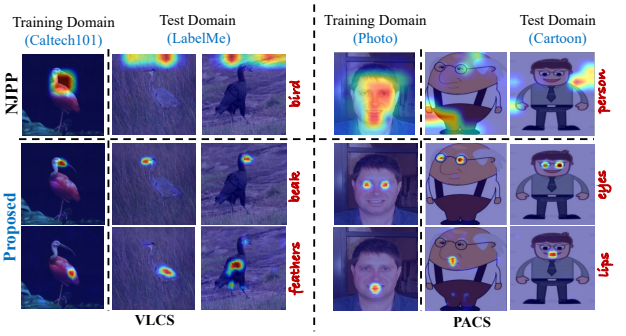

Figure 9. Samples from VLCS (left) and PACS dataset (right) across domain shifts, corresponding to bird and person classes. The first row shows GradCAM [43] saliency maps computed using the baseline NJPP [59], while the second and third row refers to the proposed method where the model highlights key concepts across domains consistently.

A.3. Additional Concept Localization Results

We show additional results showcasing the model's ability to localize key concept-specific regions across diverse datasets and domains. For instance, consider the examples in Figure 10, where the model highlights the `shell` and `flippers` of sea turtle in paintings (DomainNet), the `windows` and `roof` of house in sketch domain (PACS), and the `wheels` and `window` of car in a cluttered background (VLCS). Additionally, it maintains consistent focus

across varying visual styles, such as abstract representations in cartoons or pencil strokes in sketches.

A.4. Additional Annotation Results

We show additional concept-level annotations generated across various datasets and domains in Figure 11 using the pipeline proposed in Section 3.1 in the main paper.

A.5. LLM Prompting Details

As mentioned in Section 3.1 in the main paper, we used GPT-3.5 [2] to generate a list of distinctive, stable features that are semantically relevant for classification. We instruct the LLM to focus on features that are not specific to any particular domain. This means it avoids features that might appear in some domains but not in others. For example, `fur` may be present in photos of cats but is less likely to appear in sketches of cats. Below, we provide the exact prompt used:

> **Prompt:** List the most visually distinctive and static features of a `classname` that a classification model would rely on for accurate identification. Focus only on domain-agnostic features that are intrinsic to the object itself and truly discriminative for the class, avoiding any features that may be related to the environment or context in which the object is typically found.

Example Outputs:

- *cat*: `whiskers`, `eyes`, `ears`
- *dog*: `snout`, `ears`, `tail`
- *bird*: `beak`, `feet`, `feathers`
- *squirrel*: `tail`, `ears`, `claws`
- *hammer*: `claw`, `cheek`, `face`

Once we identified the key concepts for each class as above, we formulated a structured prompt to guide the diffusion model in generating synthetic images that highlight these concepts e.g. the synthesized exemplar image in Figure 3 and 4 in the main paper. The template for the prompt is as follows:

Template Text: *Generate a photo of a* `classname` *with its* `concept 1`*,* `concept 2`*, ..., and* `concept n`*.*

Example Prompts:

- For the class **dog**: *Generate a photo of a* `dog` *with its* `snout`*,* `ears` *and* `tail`*.*
- For the class **chair**: *Generate a photo of a* `chair` *with its* `seat`*,* `legs`*, and* `backrest`*.*

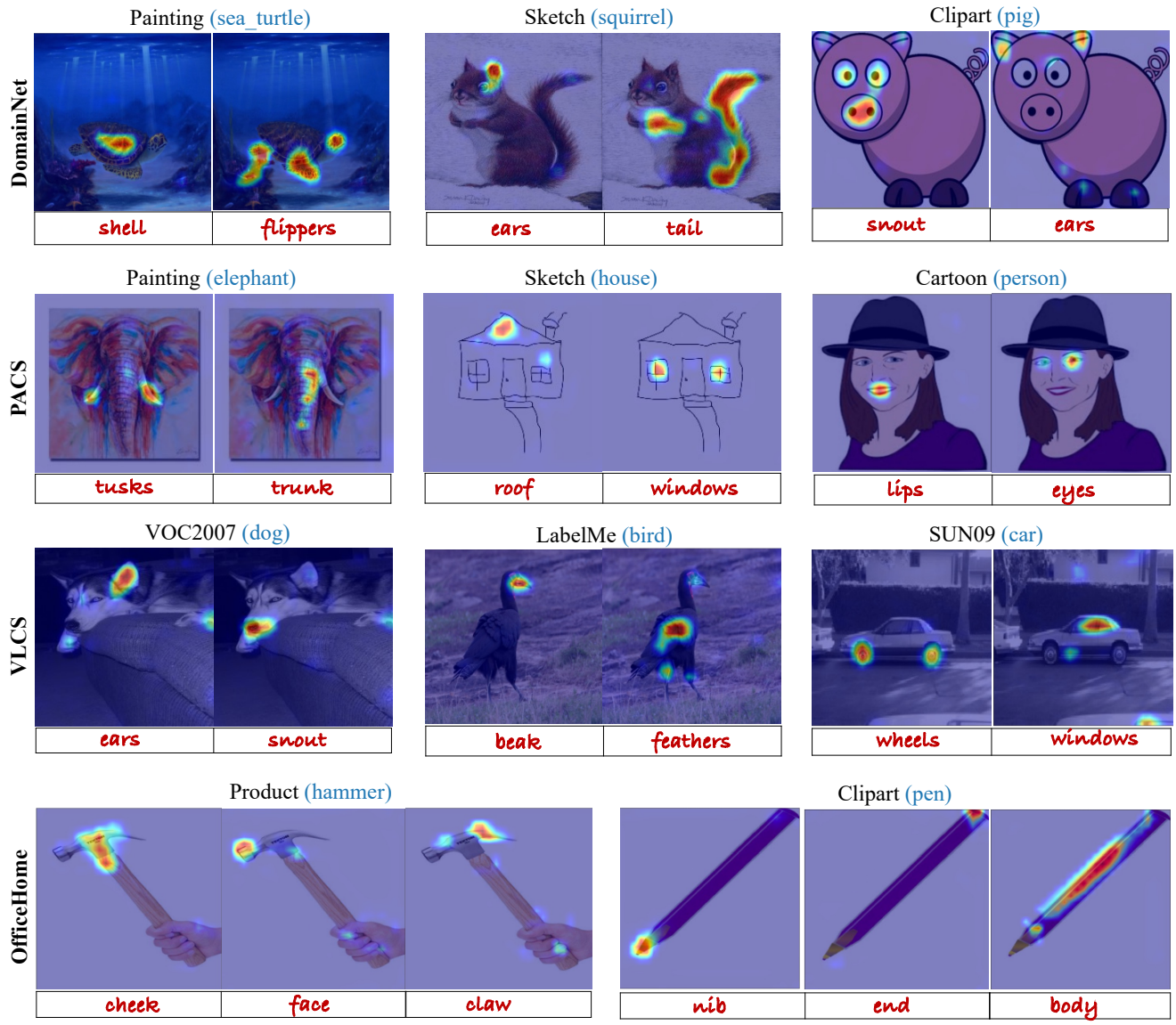

Figure 10. Additional Qualitative Results for Concept Localization.

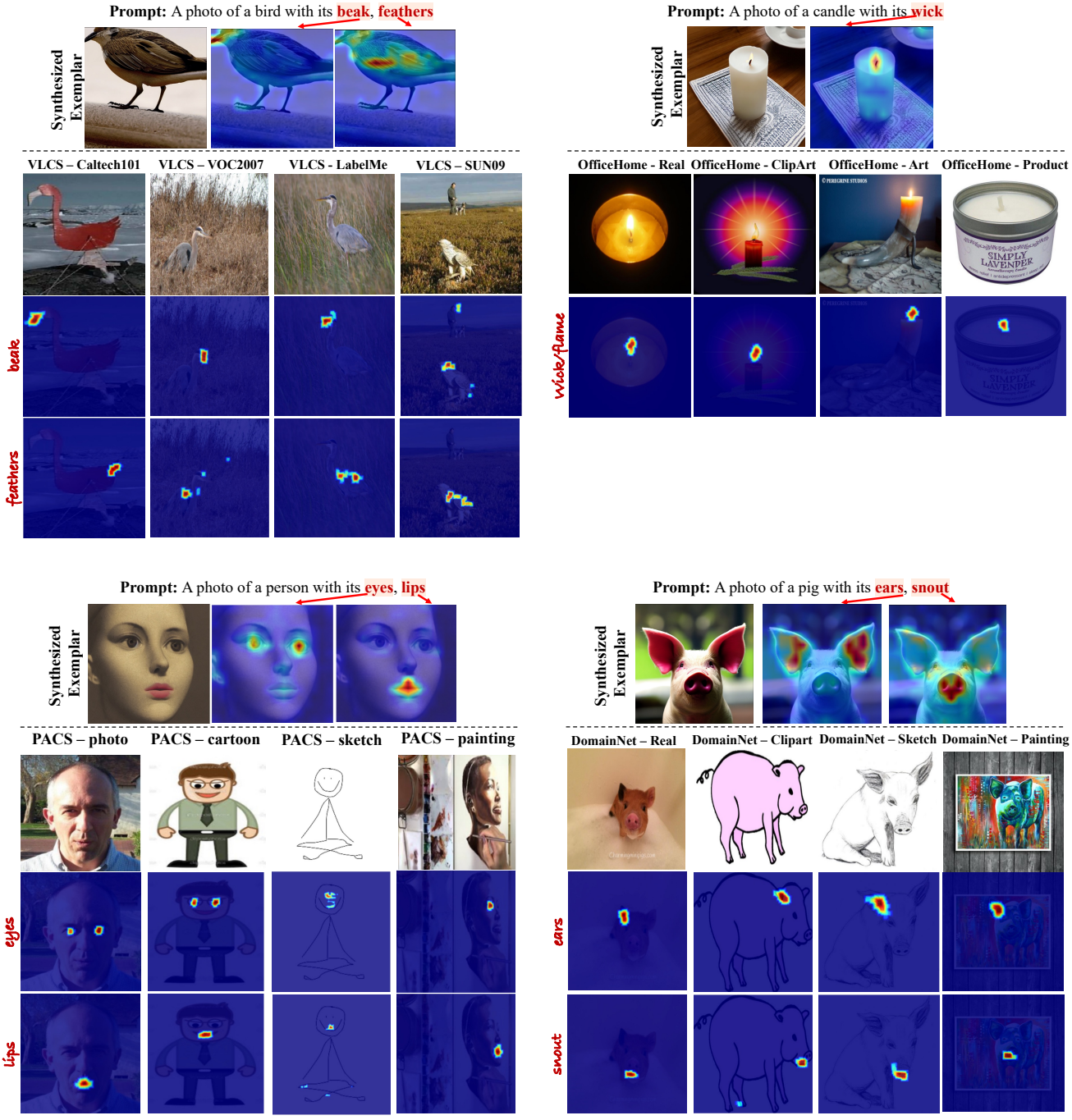

Figure 11. For each dataset (PACS, VLCS, OfficeHome, and DomainNet), we show an example of concept saliency maps (for `ear` and `mouth`) transferred from a single synthesized exemplar image to various target domain images. The figure includes one example per dataset, illustrating how the concept saliency maps align across different domains using diffusion features.