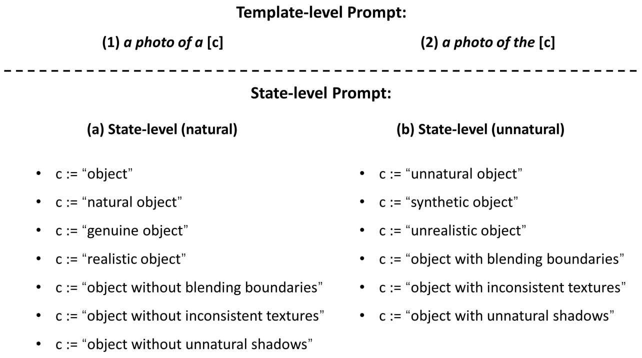
Supplementary Materials for “KA-Owl: Advancing Multimodal Fake News Detection through Knowledge-Augmented LVLMs”

Anonymous Author(s)

A CONTRASTIVE CLASS PROMPTS

Fig. A1 presents a detailed list of prompts utilized for implementing the contrastive class prompt. Following WinCLIP [1], we employ the compositional prompt ensemble to generate texts representing natural and unnatural states. Specifically, we curate prompt templates involving *a photo of a* `[c]` and *a photo of the* `[c]`. A complete prompt can be composed by replacing the token [c] in the template-level prompt with one of the state-level prompts, either from the natural or unnatural states.



Template-level Prompt:

(1) ***a photo of a*** **[c]** **(2)** ***a photo of the*** **[c]**

State-level Prompt:

(a) State-level (natural)

- c := “object”
- c := “natural object”
- c := “genuine object”
- c := “realistic object”
- c := “object without blending boundaries”
- c := “object without inconsistent textures”
- c := “object without unnatural shadows”

(b) State-level (unnatural)

- c := “unnatural object”
- c := “synthetic object”
- c := “unrealistic object”
- c := “object with blending boundaries”
- c := “object with inconsistent textures”
- c := “object with unnatural shadows”

Figure A1: Lists of two state-level prompts considered in this paper to construct contrastive class prompts.

B LOCALIZATION LOSS

B.1 Pixel-wise Localization Loss

The manipulated segmentation map M_s is utilized to calculate focal loss [2] and dice loss [3] supervised by the grounding mask. In the multimodal fake news detection task, where the majority of regions in fake images remain pristine, employing focal loss can alleviate the issue of class imbalance. The Focal loss is computed as follows:

$$\mathcal{L}_{\text{focal}} = -\frac{1}{n}\sum_{i=1}^{n}(1-p_i)^{\gamma}\log(p_i), \tag{1}$$

where n denotes the total number of pixels, p_i represents the predicted probability of positive classes, and γ is a hyperparameter for adjusting the weight of hard-to-classify samples. In our implementation, we set γ to 2.

Dice loss is based on the dice coefficient and can be computed as follows:

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2\sum_{i=1}^{n} y_i\hat{y}_i}{\sum_{i=1}^{n} y_i^2 + \sum_{i=1}^{n}\hat{y}_i^2}, \tag{2}$$

where n denotes the total number of pixels, y_i is the pixel value in the segmentation map and $\hat{y}_i$ is the ground truth value.

B.2 Path-level Localization Loss

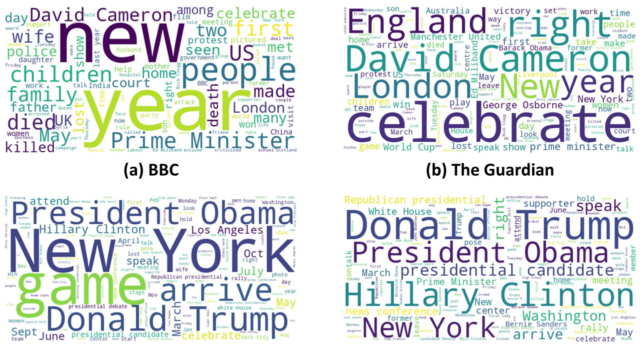
To regress the predicted bounding box, the aggregated token u_{agg} is fed into the BBox Detector D_v, which comprises two multi-layer perception (MLP) layers. Then we compute the patch-level Localization loss by combining normal L1 loss and generalized Intersection over Union (IoU) loss as follows:

$$\begin{aligned}\hat{b} &= D_v\left(u_{\text{agg}}\right), \\ \mathcal{L}_{\text{patch}} &= \mathcal{L}_{\text{L}_1}(b,\hat{b}) + \mathcal{L}_{\text{giou}}(b,\hat{b}),\end{aligned} \tag{3}$$

where $\hat{b}$ denote the predicted bounding boxes and b denote the ground-truth box.

C REAL-WORLD DISTRIBUTION DIVERGENCE

In Fig. C2, we present word clouds to illustrate the distribution divergence in real-world context across regional perspectives and thematic focus. Notably, the BBC predominantly reports British news, encompassing various subjects such as culture and entertainment. Conversely, the Washington Post tends to emphasize American political news. This disparity leads to variations in word usage across distinct domains. For instance, the commonly used words in BBC news include “prime minister", “children", and "family" etc, while in the Washington Post news, prevalent terms are “president", “Donald Trump", and “Obama", etc.



(a) BBC (b) The Guardian

(c) USA Today (d) The Washington Post

Figure C2: Word clouds of training data sourced from subsets of BBC, The Guardian, USA Today, and The Washington Post, where the size of terms corresponds to the word frequency.

D IMPLEMENTATION DETAILS

Here we provide detailed implementation details for training on the remaining three subsets: The Guardian, USA TODAY, and The Washington Post. The hyperparameters are the same for all subsets but differentiate the learning rate, batch size, and maximum epoch by dataset scale, detailed in Table D1. Additionally, to ensure a fair comparison, the hyperparameters for the PandaGPT using soft prompt tuning are set to be consistent with the values used in our method. For the state-of-the-art method HAMMER, the hyperparameters are set to default values used in [4].

Parameters	Train set		
	Guardian	USA	Wash.
Optimizer	AdamW	AdamW	AdamW
Learning Rate	4.1e-5	1.1e-5	1e-5
LR Scheduler	Linear	Linear	Linear
Batch Size	128	16	16
Maximum Epoch	12	10	10

Table D1: Hyperparameters used in The Guardian, USA TODAY, and The Washington Post subsets.

E CASE ANALYSIS

Fig. E3 and Fig. E4 depict cases from the testing set. The former illustrates comparisons between the vanilla LVLM and FKA-Owl, while the latter showcases cases where either the compared methods or the LVLM using soft prompt tuning predictions are incorrect.

In Fig. E3, Case 1 and Case 2 show cases where the vanilla LVLM model predicts all labels as “Fake” while the FKA-Owl makes correct predictions. This implies that the vanilla LVLM lacks the necessary knowledge to effectively detect fake news when learning from the general corpus, making it challenging to accurately characterize the concept of forgery and provide precise responses. Consequently, the vanilla LVLM relies heavily on common prompt words, such as “face”, resulting in all labels being predicted as “Fake”.

Figure E4 illustrates comparisons between our method with the state-of-the-art model HAMMER, and PandaGPT using soft prompt tuning (PandaGPT+SPT). In Case 3, the absence of forgery-specific knowledge impedes the ability of PandaGPT to perform manipulation reasoning, especially when confronted with fake images containing subtle artifacts. In Case 4, the powerful model HAMMER struggles to handle agnostic instances due to the lack of prior information, resulting in incorrect judgments regarding unseen fake news. Conversely, Case 5 demonstrates the effectiveness of our FKA-Owl, which integrates inherent world knowledge from LVLMs and incorporates forgery-specific information to make accurate predictions.

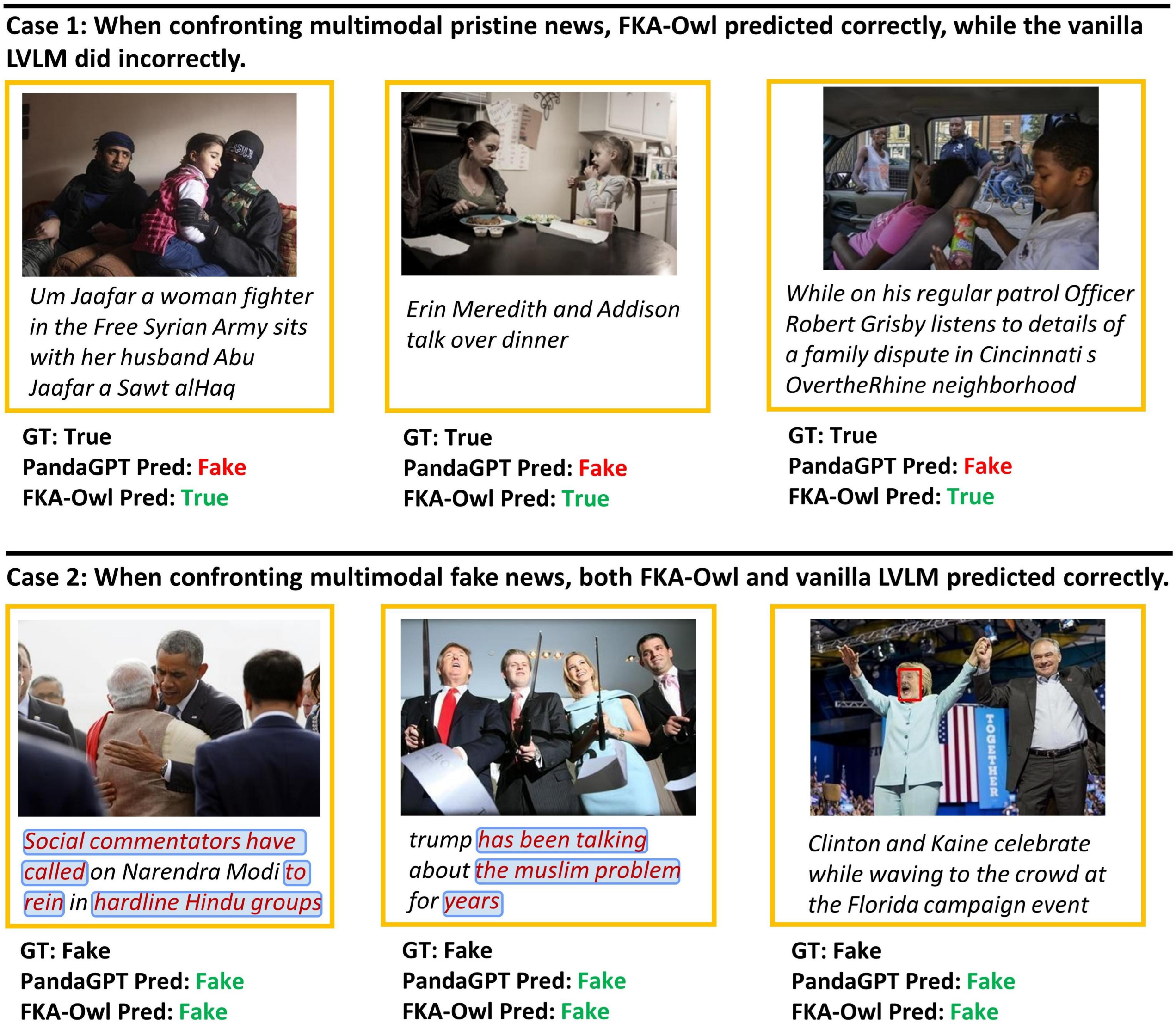
REFERENCES

[1] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. 2023. Winclip: Zero-/few-shot anomaly classification and segmentation. In *CVPR*.

[2] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *ICCV*.

[3] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*.

[4] Rui Shao, Tianxing Wu, and Ziwei Liu. 2023. Detecting and grounding multi-modal media manipulation. In *CVPR*.



Case 1: When confronting multimodal pristine news, FKA-Owl predicted correctly, while the vanilla LVLM did incorrectly.

Um Jaafar a woman fighter in the Free Syrian Army sits with her husband Abu Jaafar a Sawt alHaq

GT: True
PandaGPT Pred: Fake
FKA-Owl Pred: True

Erin Meredith and Addison talk over dinner

GT: True
PandaGPT Pred: Fake
FKA-Owl Pred: True

While on his regular patrol Officer Robert Grisby listens to details of a family dispute in Cincinnati s OvertheRhine neighborhood

GT: True
PandaGPT Pred: Fake
FKA-Owl Pred: True

Case 2: When confronting multimodal fake news, both FKA-Owl and vanilla LVLM predicted correctly.

Social commentators have called on Narendra Modi to rein in hardline Hindu groups

GT: Fake
PandaGPT Pred: Fake
FKA-Owl Pred: Fake

trump has been talking about the muslim problem for years

GT: Fake
PandaGPT Pred: Fake
FKA-Owl Pred: Fake

Clinton and Kaine celebrate while waving to the crowd at the Florida campaign event

GT: Fake
PandaGPT Pred: Fake
FKA-Owl Pred: Fake

Figure E3: Cases in the testing set where vanilla LVLM and FKA-Owl confront with pristine news and fake news. (●) indicates wrong prediction and (●) indicates correct prediction.

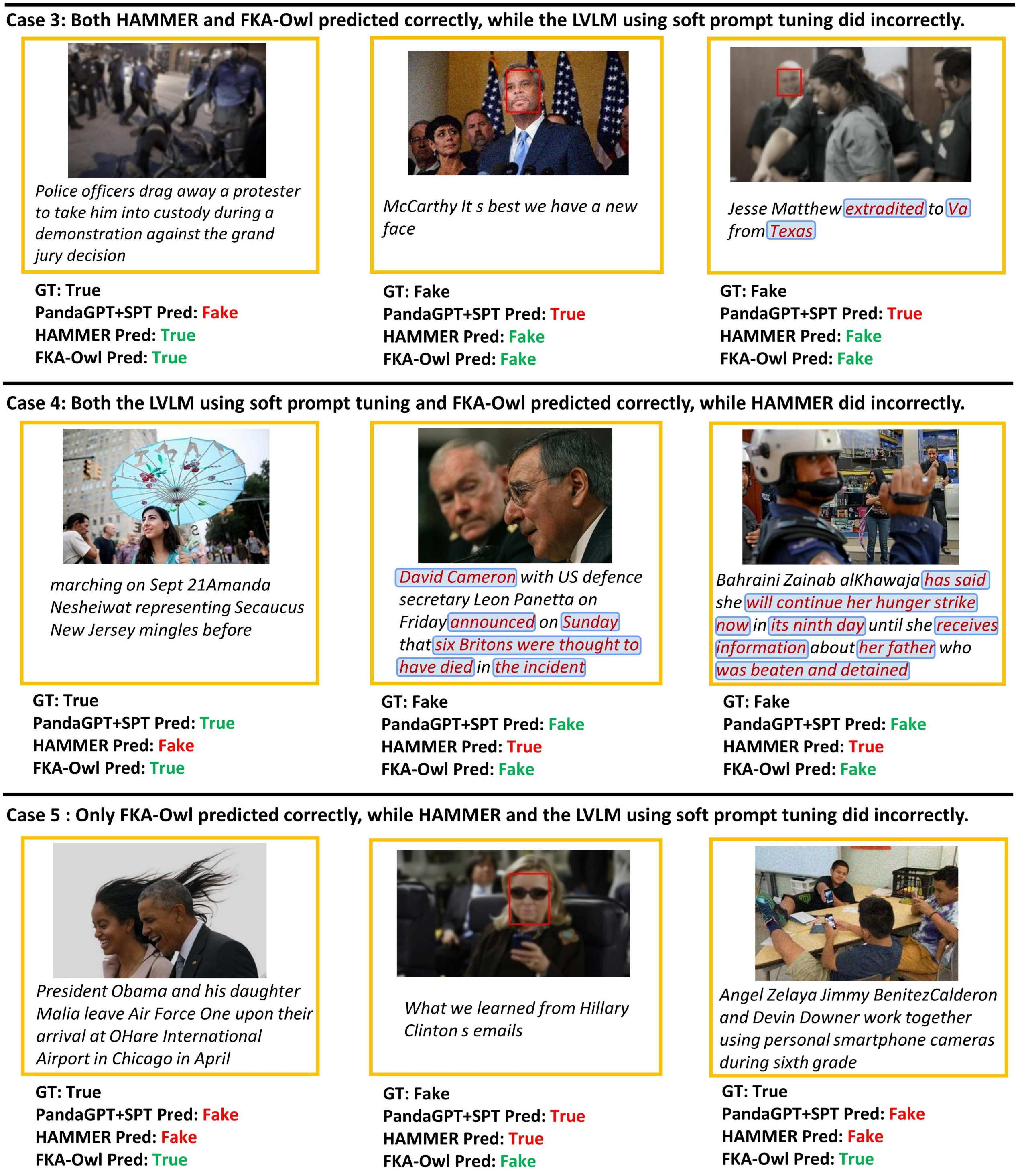

Case 3: Both HAMMER and FKA-Owl predicted correctly, while the LVLM using soft prompt tuning did incorrectly.
Police officers drag away a protester to take him into custody during a demonstration against the grand jury decision
GT: True
PandaGPT+SPT Pred: Fake
HAMMER Pred: True
FKA-Owl Pred: True
McCarthy It s best we have a new face
GT: Fake
PandaGPT+SPT Pred: True
HAMMER Pred: Fake
FKA-Owl Pred: Fake
Jesse Matthew extradited to Va from Texas
GT: Fake
PandaGPT+SPT Pred: True
HAMMER Pred: Fake
FKA-Owl Pred: Fake
Case 4: Both the LVLM using soft prompt tuning and FKA-Owl predicted correctly, while HAMMER did incorrectly.
marching on Sept 21Amanda Nesheiwat representing Secaucus New Jersey mingles before
GT: True
PandaGPT+SPT Pred: True
HAMMER Pred: Fake
FKA-Owl Pred: True
David Cameron with US defence secretary Leon Panetta on Friday announced on Sunday that six Britons were thought to have died in the incident
GT: Fake
PandaGPT+SPT Pred: Fake
HAMMER Pred: True
FKA-Owl Pred: Fake
Bahraini Zainab alKhawaja has said she will continue her hunger strike now in its ninth day until she receives information about her father who was beaten and detained
GT: Fake
PandaGPT+SPT Pred: Fake
HAMMER Pred: True
FKA-Owl Pred: Fake
Case 5 : Only FKA-Owl predicted correctly, while HAMMER and the LVLM using soft prompt tuning did incorrectly.
President Obama and his daughter Malia leave Air Force One upon their arrival at OHare International Airport in Chicago in April
GT: True
PandaGPT+SPT Pred: Fake
HAMMER Pred: Fake
FKA-Owl Pred: True
What we learned from Hillary Clinton s emails
GT: Fake
PandaGPT+SPT Pred: True
HAMMER Pred: True
FKA-Owl Pred: Fake
Angel Zelaya Jimmy BenitezCalderon and Devin Downer work together using personal smartphone cameras during sixth grade
GT: True
PandaGPT+SPT Pred: Fake
HAMMER Pred: Fake
FKA-Owl Pred: True

Figure E4: Cases in the testing set where at least one in the Baseline and the LVLM using soft prompt tuning made incorrect predictions. (●) indicates wrong prediction and (●) indicates correct prediction.