A. Qualitative Case Study on Hallucination Reduction

We present a qualitative case study on object hallucination using the MSCOCO dataset [1]. Captions generated using LLaVA-OV [2] and using our proposed CAFE are compared. Text in red indicates objects that are not present in the original image but are generated by the model. According to the case study, LLaVA-OV [2] generated hallucinated objects, *e.g.*, 'sink', 'bird'. Our proposed CAFE demonstrates superior performance by generating fewer object hallucinations. Additionally, it produces longer and more detailed captions compared to LLaVA-OV [2].

LLaVA-OV [2] Generated Caption: In the heart of a bustling city, a large, vibrant mural adorns the side of a brick building. The mural is a lively depiction of a yellow clock with black hands and numbers, standing out against the red background. The clock is not alone; it's accompanied by a red bird perched on its top and a gold trumpet resting on its bottom, adding an element of whimsy to the scene. The building itself is situated on a busy street corner. A yellow taxi cab and a red van are parked in front of it, hinting at the hustle and bustle of city life. Above, the sky is a clear blue, dotted with white clouds, providing a serene backdrop to this urban landscape. The image captures the essence of city life - the blend of art and architecture, the constant movement of vehicles, and the ever-changing sky. It's a snapshot of a moment in time in a city that's always on the move.

CAFE (Ours) Generated caption: The image captures a large, ornate clock mounted on the side of a building. The clock face is circular and prominently features Roman numerals for the hours, with the hands indicating a time of approximately 12:30. Above the clock face, there is a decorative element that includes a red shield with the Vodafone logo in white, flanked by two golden trumpets on either side. Below the clock face, there are two banners; the upper banner reads "İşte ben de bir kral olur" which translates to "I too want to be a king" in Turkish, while the lower banner states "Vodafone 24 saat hizmet" meaning "Vodafone 24-hour service." In the foreground, there are several vehicles parked along the street, including a yellow taxi cab, a white van, and a maroon SUV. The sky above is clear with a few clouds, suggesting it might be late afternoon or early evening based on the lighting. The building itself appears to be constructed with red brick, and there is another advertisement visible on the right side of the building, though it is partially obscured by the clock structure. The overall style of the image is a real-life photograph taken during daylight hours.

LLaVA-OV [2] Generated Caption: This image captures a modern kitchen bathed in soft light. The dominant color scheme is white, punctuated by the sleek silver of the appliances and the vibrant red of the kettle. The kitchen is equipped with a stainless steel refrigerator standing tall on the right, its surface reflecting the ambient light. Adjacent to it, a black stove waits ready for culinary adventures, while a microwave sits atop it, offering convenience at arm's reach. The heart of the kitchen is a sink nestled between the stove and a window. The window allows natural light to filter in, illuminating the space and creating a warm atmosphere. Above the sink, white cabinets offer ample storage space, their pristine surfaces adding to the overall clean aesthetic of the kitchen. A red kettle rests on the counter next to the sink, its bold color contrasting beautifully with the surrounding whites. Its presence suggests many mornings started with a comforting cup of tea. Overall, this kitchen exudes a sense of modern elegance and functionality, with every element thoughtfully placed for ease of use and visual appeal.

CAFE (Ours) Generated caption: The image depicts a modern kitchen with a clean and organized appearance. The kitchen features white cabinetry that includes both upper and lower cabinets, providing ample storage space. The upper cabinets have two doors each, while the lower cabinets have one door each, all fitted with silver handles. On the countertop, there is a stainless steel microwave oven positioned to the left of a red kettle. The microwave has a digital display and control panel on its front side. Next to the microwave, there is a red kettle with a black handle, which adds a pop of color to the otherwise neutral-toned kitchen. To the right of the microwave and kettle, there is a stainless steel refrigerator with a freezer compartment at the bottom. The refrigerator has a sleek design with a handle located on the right side for accessing the freezer and a separate handle for the main compartment. Below the countertop, there is a built-in oven with a glass window and a red oven mitt hanging from the oven handle. The oven appears to be part of the cabinetry, seamlessly integrated into the design. The backsplash behind the countertop consists of dark red tiles, which contrast with the white cabinetry and add depth to the space. The floor is covered with light-colored carpeting, which complements the overall color scheme of the kitchen. The lighting in the kitchen is soft and ambient, coming from recessed lights in the ceiling, which creates a warm and inviting atmosphere. There are no visible windows or other sources of natural light in the image. Overall, the kitchen is designed with functionality and aesthetics in mind, featuring modern appliances and a cohesive color palette that combines white, silver, and red accents.

B. Zero-shot and Linear Probe Classification

We present additional results on zero-shot and linear probe classification over 7 datasets in Table 1. For all the experiments, we follow the setup of CLIP [3]. CAFE consistently demonstrates comparable performance to CLIP.

	Zero-Shot							Linear Probe						
Model	CIFAR10	CIFAR100	MNIST	STL10	SUN397	SST2	GTSRB	CIFAR10	CIFAR100	MNIST	STL10	SUN397	SST2	GTSRB
CLIP [3] (ViT-B/32)	89.8	**65.1**	48.2	**97.1**	**63.2**	59.6	32.2	95.1	80.5	**99.0**	98.3	76.6	70.8	85.3
CAFE-0.5B	90.5	60.0	**48.4**	93.9	38.0	**60.3**	**45.6**	97.0	**82.7**	98.1	99.0	76.0	76.3	**87.6**
CAFE-7B	**93.1**	63.7	33.0	96.9	56.7	54.1	43.6	**97.4**	82.0	98.6	**99.1**	**76.9**	**86.6**	87.4

Table 1. Zero-shot and linear probe classification results. CAFE achieves comparable performance to CLIP [3].

References

[1] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *ArXiv*, abs/1504.00325, 2015. 1

[2] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 2

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3