Exemplar Masking for Multimodal Incremental Learning
Supplementary Materials

Yi-Lun Lee[1,3] Chen-Yu Lee[2] Wei-Chen Chiu[1] Yi-Hsuan Tsai[3]
[1]National Yang Ming Chiao Tung University [2]Google [3]Atmanity
{yllee10727, walon}@cs.nycu.edu.tw, chenyulee@google.com, yhtsai@atmanity.io

1. Ablation Studies

Masking threshold. We have shown the effects of different masking thresholds of our exemplar masking framework on the MM-ImageNet dataset in Table 4 of the main paper. To validate the same conclusion from different datasets, we further evaluate on the UPMC Food-101 dataset as shown in Table 1. As shown in the results, the performance trend on UPMC Food-101 is similar to that on MM-ImageNet-R, demonstrating that our exemplar masking framework is not sensitive to datasets. Moreover, the trend also reveals that when the threshold is beyond the specific amount (i.e., $\mu - 0.25 \times \sigma$), the performance does not show significant variance, indicating that the framework is also not sensitive to the thresholds.

Table 1. Analysis of the masking thresholds on UPMC Food-101.

Masking Threshold τ	Preserved Ratio	# of Exemplars	$\bar{A}$
$\mu + 0.5 \times \sigma$	0.25	19.09	84.60
$\mu + 0.25 \times \sigma$	0.28	16.89	**85.10**
μ	0.36	13.16	84.51
$\mu - 0.25 \times \sigma$	0.60	7.88	84.39
$\mu - 0.5 \times \sigma$	0.86	5.47	80.72
$\tau_i = 0.005, \tau_t = 0.001$	0.46	10.29	84.40
$\tau_i = 0.004, \tau_t = 0.001$	0.55	8.64	84.32

Masking methods. The design choices for the masking method involve two factors: the order of masking (i.e., which modality to be masked first) and the cross-attention strategy (i.e., what information should the masking strategy capture in the secondary modality). Specifically, there are two ways of cross-attention we investigate: whether the second modality captures 1) the contextual information from the discarded tokens (denoted as **complementary**), or 2) the information of the preserved tokens (denoted as **relevant**) of the first-masked modality. **Complementary** method preserves the information from discarded regions, completing the information of entire image contents with different modalities. **Relevant** method keeps the auxiliary information related to the preserved regions, enhancing the information of target objects for recognition.

In Table 2, we show the results of considering various designs. In the first two rows, we observe that when masking the image first, using the different cross-attention designs for masking texts has competitive results, indicating that both designs encourage the masked texts to preserve different helpful information to assist models in replaying old knowledge. Particularly, the model with the complementary method reaches the best performance since it can obtain more diverse training samples via multimodal data augmentation and thus result in better replaying of exemplars. Moreover, the performance gain between the models applying these two designs increases when the number of incremental learning phases increases, validating the better effectiveness of using the complementary method for masking the second modality. In the case of masking texts first, the masked text may not provide as much discriminative information as the masked image does, leading to suboptimal performance.

2. More Quantitative Results

We have shown the comparisons with state-of-the-art methods on the MM-ImageNet and UPMC Food-101 datasets in Table 2 of the main paper. Moreover, we provide a significance analysis of the results. We conduct experiments three times with different random seeds and report standard deviations to demonstrate the effectiveness of our exemplar masking framework, as shown in Table 3.

In addition to the comparisons with state-of-the-art class-incremental learning methods, we also compare with the recently proposed multimodal class-incremental learning approach, GMM [1], which considers MCIL as a generative task. However, they use the LLM as the backbone, which benefits from the powerful prior knowledge gained from large pre-trained datasets. This makes a direct comparison unfair to our method, where we utilize a smaller, transformer-based ViLT model. Despite this difference, to demonstrate the effectiveness of our method, we show results on Tiny-ImageNet as GMM used in their paper. In

Table 2. Ablation study of different design choices for the masking method (see the second paragraph of Section 1 for details).

First-masked Modality	Second-masked Modality	Cross-Attention Strategy	$\bar{A}$ ($L = 10$)	$\bar{A}$ ($L = 20$)
Image	Text	Complementary	**80.55 (-0.00)**	**78.64 (-0.00)**
Image	Text	Relevant	80.34 (-0.21)	78.24 (-0.40)
Text	Image	Complementary	79.83 (-0.72)	77.57 (-1.07)
Text	Image	Relevant	80.05 (-0.50)	77.83 (-0.81)

Table 3. Average incremental accuracy on our MM-ImageNet-R dataset with different numbers of incremental phases L=5, 10, 20.

Methods	# of Parameters	MM-ImageNet-R		
		L=5	L=10	L=20
L2P [12]	85K	65.488±3.344	69.357±0.153	67.620±1.833
DualPrompt [11]	(15+46 × L) K	78.058±0.756	74.859±1.458	70.299±0.415
EASE [13]	1152K	77.421±0.558	76.414±0.465	73.674±0.418
SSF	206K	80.859±0.265	77.193±0.333	75.393±0.179
SSF + MRDC [10]		81.325±0.485	78.758±0.476	77.020±0.803
SSF + CIM [8]		81.071±0.089	77.494±0.641	75.003±2.193
SSF + Ours		**83.405±0.574**	**80.545±0.159**	**77.850±0.701**

Table 4, our method performs comparably in different incremental learning settings.

Table 4. Average incremental accuracy (**Avg**) and the accuracy after the last incremental phase (**Last**) on Tiny-ImageNet with different numbers of incremental phases L=5, 10.

Methods	Tiny-ImageNet			
	L=5		L=10	
	Avg	**Last**	**Avg**	**Last**
GMM [1] - LLM	84.16	78.46	**83.95**	78.64
Ours - ViLT	**84.28**	**79.42**	83.65	**78.92**

3. More Qualitative Results

We provide more examples of the proposed exemplar masking on the MM-ImageNet-R dataset, in which the captions are generated by querying intstructBLIP, as shown in Figure 1, 2, 3, 4, 5, 6. We also provide examples on the UPMC Food-101 dataset, as shown in Figure 7, 8, 9, 10.For the masked images, we visualize the masked regions (i.e., $M_I \otimes x_I$) and the discarded regions (i.e., $(1 - M_I) \otimes x_I$) with the corresponding mask maps. For the masked texts, we highlight the words related to the target objects (with the yellow color box) and the words related to the contextual information from the discarded regions (with the melon color box).

As shown in these examples, the image masks correctly bound the class-related regions that contain the most important information. Moreover, the preserved regions are quite smaller than the discarded regions, meaning that in the image modality, there is a large proportion of redundant information that requires large storage space but is not helpful for model learning recognition. On the other hand, the masked texts preserve both the words related to the corresponding class and the words indicating the contextual information from the discarded regions. These preserved words not only provide complementary information related to the class object but also preserve the information from the discarded image regions.

4. Training Procedure

Algorithm 1 shows the whole training procedure of our exemplar masking framework for multimodal incremental learning. In each incremental phase, we first train the model with the available data, including new samples and exemplars. During training, we adopt multimodal data augmentation on the exemplars to replay the old knowledge effectively. After finishing the training step, we use the learned model to obtain the attention maps of the new training samples. Then we calculate the masking thresholds and obtain the resultant masks for both modalities. Finally, we apply herding algorithm [9] to select k samples as the exemplars, where the size of k samples does not exceed the budget limit.

5. Implementation Details

Inputs. In our experiment, we validate the proposed method on the vision and language dataset, which consists of image and text modality. For the image modality, we follow [3] to resize the shorter side of input images to 384 and constrain the longer side to under 640 while keeping the aspect ratio. Following [2], we decompose images into patches of size 32 × 32. For the text modality, the text in-

Algorithm 1: Training procedure of the exemplar masking framework.

for *each incremental phase* **do**
  Data: Training set contains new training data D^{new} and exemplars D^{exp} in the l incremental phase.
  Train the model with training set $D^{\text{new}} \cup D^{\text{exp}}$
  for *Each epoch* **do**
    for (x_T, x_I, y) in $D^{\text{new}} \cup D^{\text{exp}}$ **do**
      if (x_T, x_I, y) in D^{exp} **then**
        Random select x'_T, where $y' = y$
        Create the augmented sample (x'_T, x_I, y)
      Update learnable parameters via $\mathcal{L}_{\text{CE}}$ in Eq. 6
  Build exemplars
  for (x_T, x_I, y) in D^{new} **do**
    Mask image tokens
    Calculate the attention map $A_{CLS \to I}$, and obtain the image threshold τ_I via Eq. 1
    Obtain image masks M_I via Eq. 2
    Mask text tokens
    Calculate the cross-attention map $A_{I \to T}$, and obtain the text threshold τ_T via Eq. 3
    Obtain text masks M_T via Eq. 4
    Obtain the masked samples via Eq. 5
  Select exemplars under the memory limit
  for c *in* $C^{\text{new}(l)}$ **do**
    Select the k masked samples with the features nearest to the class mean μ_c as exemplars without exceeding the storage space.

put is tokenized by the bert-base-uncased tokenizer with the maximum length of text inputs set to 128.

Model Configurations. In the multimodal incremental learning framework, we have two components, including the multimodal backbone and the incremental classifier. We adopt the pre-trained multimodal transformer ViLT [3] as our backbone for feature extraction since it is widely used in various transformer-based methods for multimodal learning. Based on Vision Transformers [2], ViLT advances to process multimodal inputs with the tokenized texts and patched images, and is pre-trained on several large vision-language datasets (e.g., MS-COCO [6] and Visual Genome [4]) via objectives such as Image Text Matching and Masked Language Modeling. The incremental classifier is a single linear layer that maps the features to the class prediction. As the new classes come in sequentially, in each incremental phase, we extend the classifier with additional normal-initialized parameters for the new classes.

Model Training Details. In our experiments, we apply our exemplar masking framework on two learning methods, including finetuning and parameter-efficient tuning (PET) methods (i.e., SSF [5]). For the finetuning, to prevent the dramatic overriding of the weights of pre-trained ViLT backbone meanwhile learning recognition of new classes, we set the learning rate of the ViLT backbone and the classifier to 1×10^{-5} and 1×10^{-3} respectively. For the SSF, we freeze all the parameters of the ViLT backbone and only train the learnable parameters (i.e., scales and shifts vectors for SSF) in each layer as well as the parameters of the classifier. We set the learning rate for all learnable parameters to 1×10^{-3}. We use the AdamW optimizer [7] in all experiments and weight decay is set to 2×10^{-2}. The learning rate is warmed up for 10% of the total training epochs and is then decreased linearly to zero. For each incremental phase, we train the models by 30 epochs.

References

[1] Xusheng Cao, Haori Lu, Linlan Huang, Xialei Liu, and Ming-Ming Cheng. Generative multi-modal models are good class incremental learners. In *CVPR*, 2024. 1, 2

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 3

[3] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 2, 3

[4] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 3

[5] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. *NeurIPS*, 2022. 3

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018. 3

[8] Zilin Luo, Yaoyao Liu, Bernt Schiele, and Qianru Sun. Class-incremental exemplar compression for class-incremental learning. In *CVPR*, 2023. 2

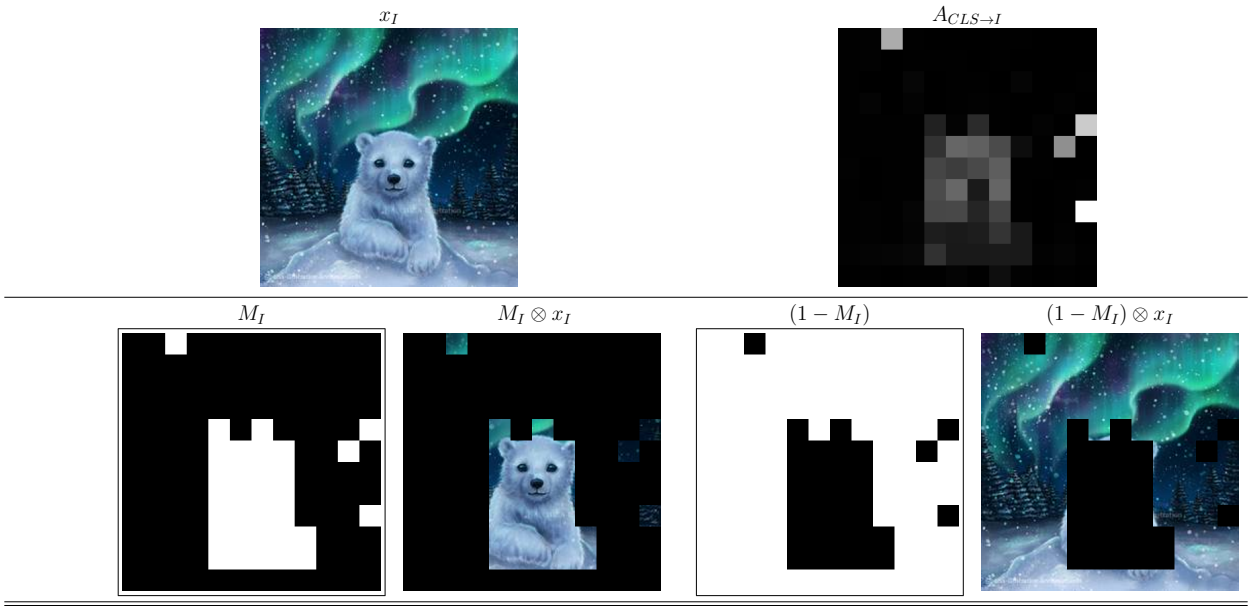
[9] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017. 2

[10] Liyuan Wang, Xingxing Zhang, Kuo Yang, Longhui Yu, Chongxuan Li, Lanqing Hong, Shifeng Zhang, Zhenguo Li, Yi Zhong, and Jun Zhu. Memory replay with data compression for continual learning. In *ICLR*, 2022. 2

[11] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *ECCV*, 2022. 2

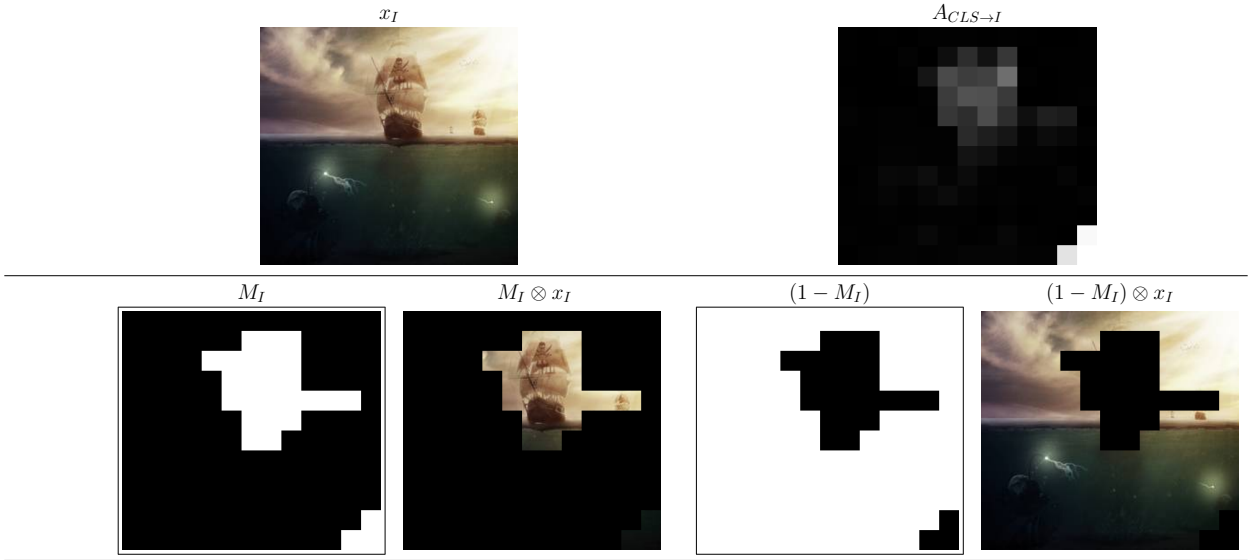
[12] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *CVPR*, 2022. 2

[13] Da-Wei Zhou, Hai-Long Sun, Han-Jia Ye, and De-Chuan Zhan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In *CVPR*, 2024. 2



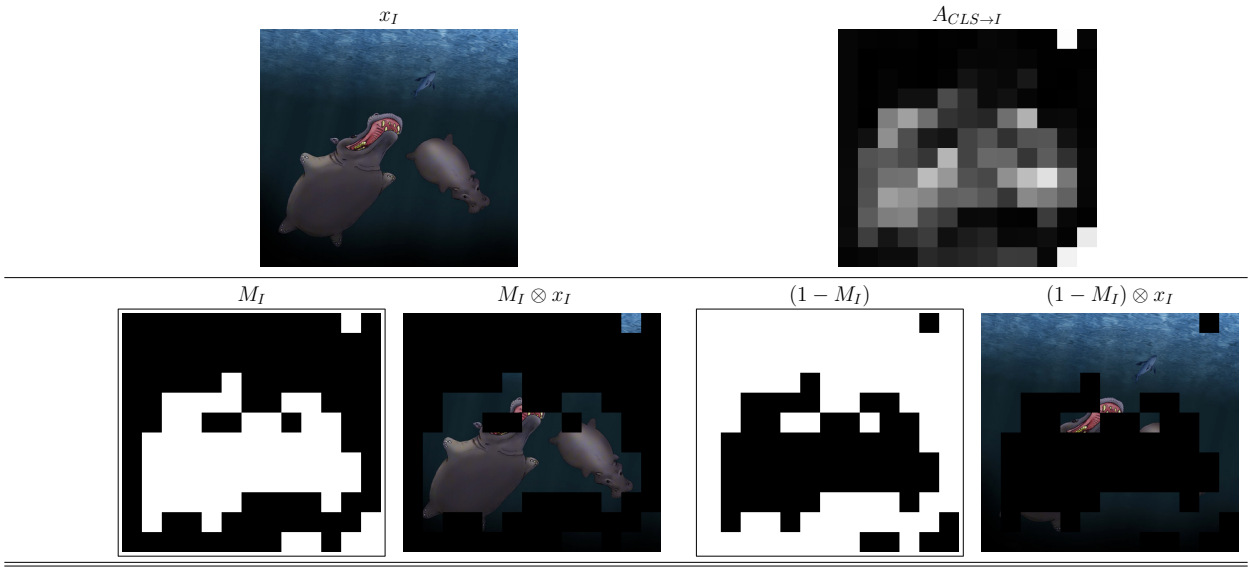
$M_T \otimes x_T$: [CLS] the image depicts a white polar bear sitting on top of a snow - covered surface during the aurora borealis, creating a mesmerizing and serene atmosphere. this is a painting made by an artist who captures the beauty of the natural world in her artworks. the picture features a white polar bear sitting on a snow - covered surface, with a vibrant and colorful northern lights backdrop illuminating the scene. the white and blue color scheme complements the aurora borealis, making it a stunning painting that showcases the majesty of this natural phenomenon while highlighting the beauty of the polar bear species as well. [SEP]

Figure 1. An example of exemplar masking on the MM-ImageNet-R (texts generated by InstructBLIP) for the class "**ice bear**".



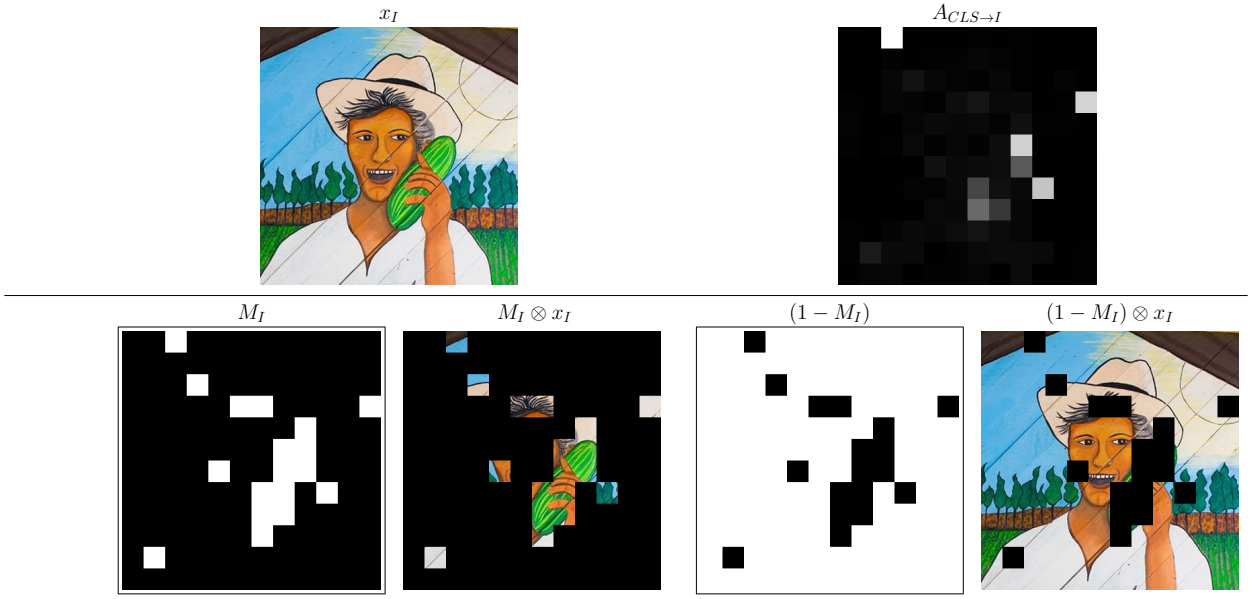
$M_T \otimes x_T$: [CLS] in the image, there is an enormous pirate ship floating in the ocean at sunset. the ship is positioned under a dark cloud and surrounded by various sea creatures, such as squids, fishes, and sharks. the scene oozes a sense of wonder and fear due to the ominous atmosphere and the presence of the fearsome creatures. the ship, along with the vast ocean and cloudy sky, forms a dramatic and intimidating backdrop for the various sea creatures and adds to the eerie tone of the scene. [SEP]

Figure 2. An example of exemplar masking on the MM-ImageNet-R (texts generated by InstructBLIP) for the class "**pirate ship**".



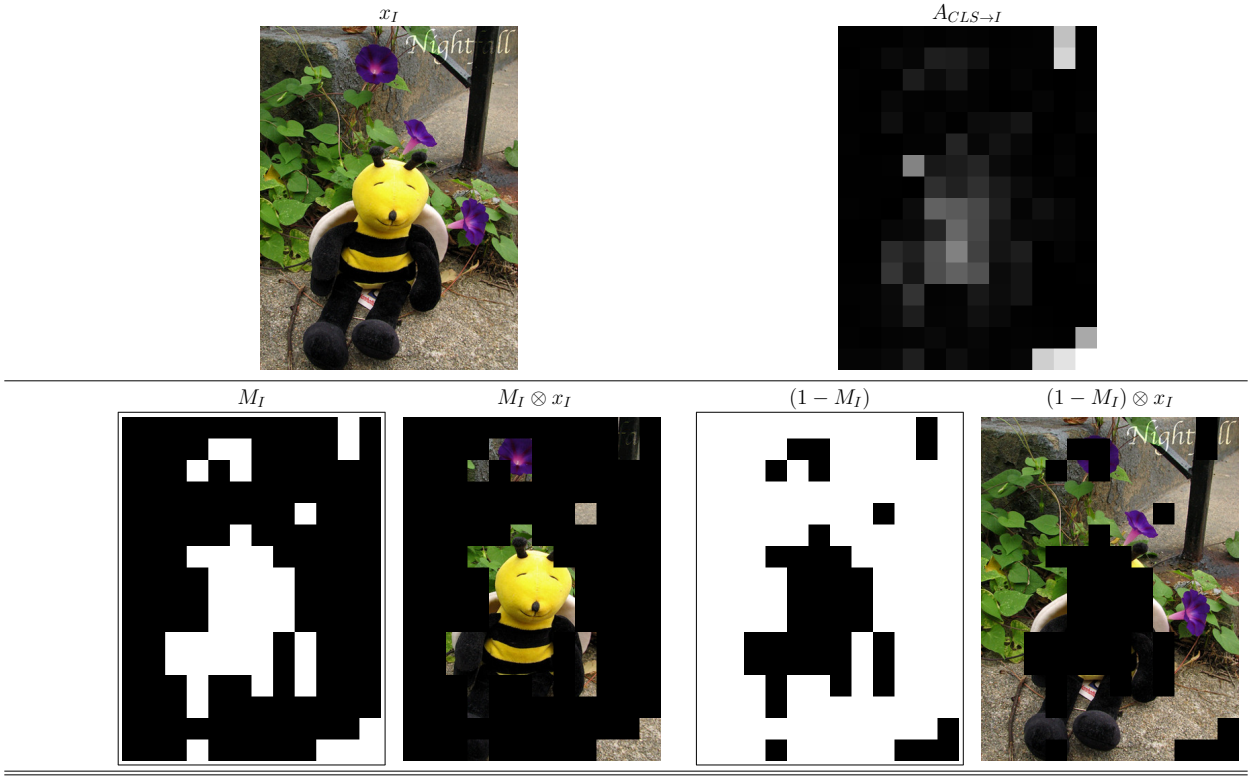
$M_T \otimes x_T$: [CLS] the image is of a peaceful scene featuring a shark and two hippos, all submerged underwater. two large hippos are present underwater, one near the background and the other nearer to the foreground. one hippo has a visible open mouth, almost as if about to start swimming or perhaps looking for food. also, a shark can be seen beneath the waves in the foreground. the colors in the image are deep and contrasting, with neutral blue tones emphasizing the underwater environment and the creatures swimming and living within it. [SEP]

Figure 3. An example of exemplar masking on the MM-ImageNet-R (texts generated by InstructBLIP) for the class "**hippopotamus**".



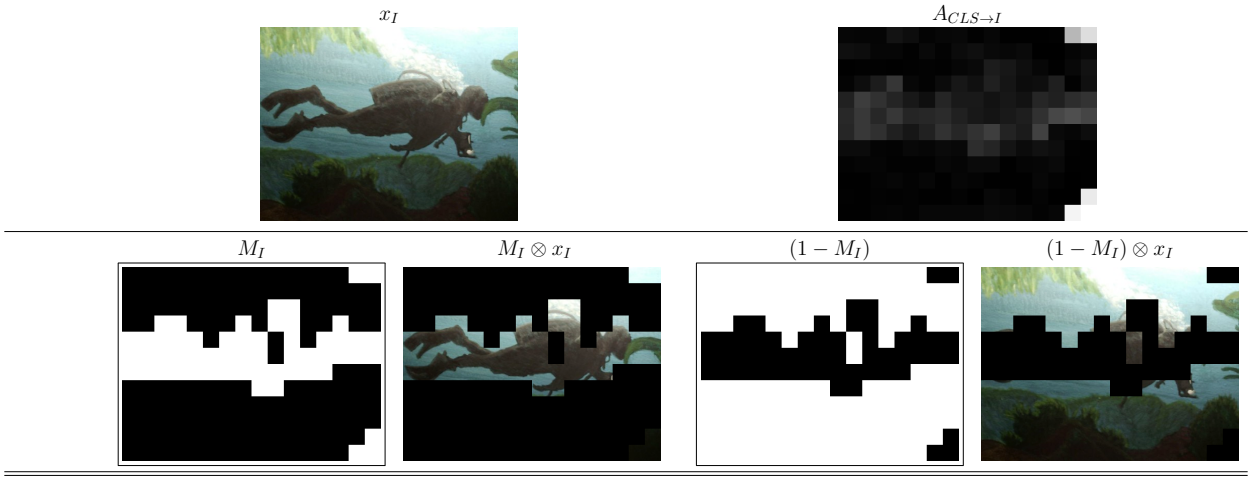
$M_T \otimes x_T$: [CLS] the image features a painted mural located in chiapas, mexico of a **farmer** holding a melon or cucumber in his hand while talking on a phone. the painting, which is located near a road or pathway, showcases a vivid scene of a local farmer using communication technology, reflecting how modern life is evolving within the traditional mexican farming setting. the painted backdrop behind the farmer has a wooden structure or fence, indicating the rural nature of this area and the importance of agriculture in the local community. overall, the image captures the essence of a busy farmer, who is balancing modern communication and traditional farm work, illustrating the unique blend [SEP]

Figure 4. An example of exemplar masking on the MM-ImageNet-R (texts generated by InstructBLIP) for the class "**cucumber**".



$M_T \otimes x_T$: [CLS] the image features a stuffed bee sitting on the steps of a small staircase, with purple flowering plants surrounding it. the bee is the centerpiece of the picture and resembles a cute and fun decorative element. the flowers and plant life add some vibrancy to the setting. to make this scene more interesting and fun, a stuffed raccoon wearing a hat is seen sitting by the steps, interacting with the bee. [SEP]

Figure 5. An example of exemplar masking on the MM-ImageNet-R (texts generated by InstructBLIP) for the class "**bee**".



$M_T \otimes x_T$: [CLS] the scene depicts a diver exploring underwater, withthe focus on the scuba gear he himself is wearing and a particular area of the underwater life. the painting displays the scuba gear and a pear, indicating this might be a diving experience, and possibly showcased a particularly detailed and well - executed scene of an underwater experience by the artist. while the painting is mostly of the scuba gear, it's important to note that the diver can be seen as well, suggesting a focus on the underwater environment that the diver explores. [SEP]

Figure 6. An example of exemplar masking on the MM-ImageNet-R (texts generated by InstructBLIP) for the class "**scuba diver**".

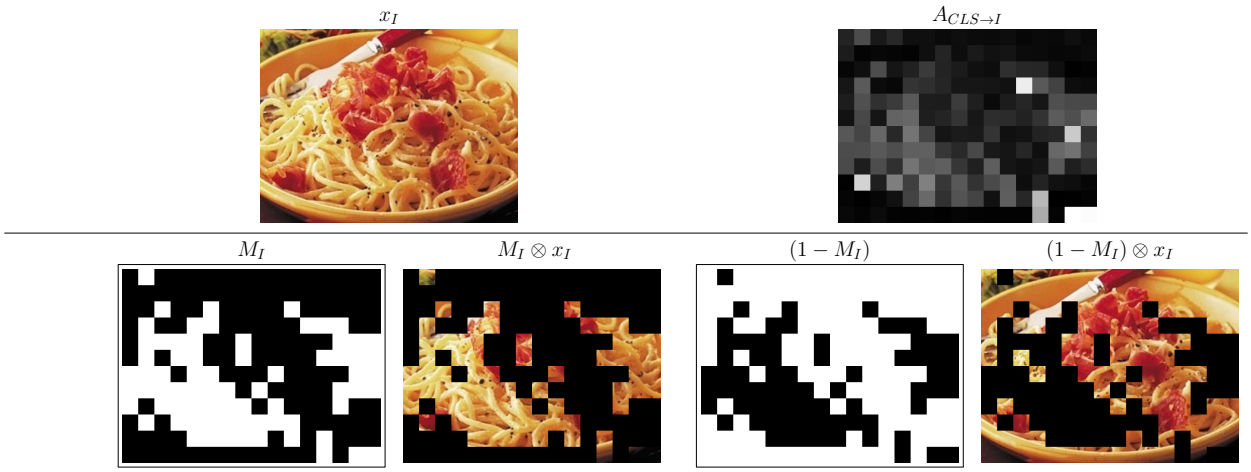


Figure 7. An example of exemplar masking on the UPMC Food-101 for the class "**spaghetti carbonara**".

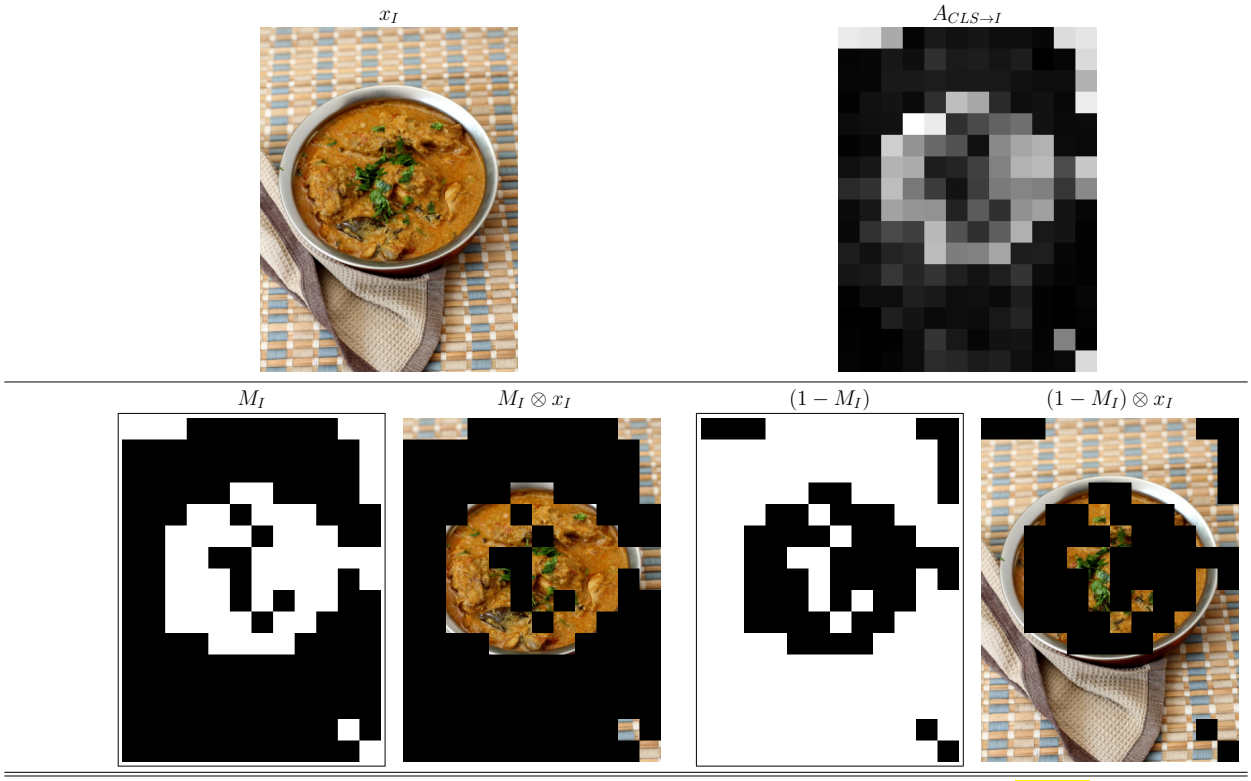
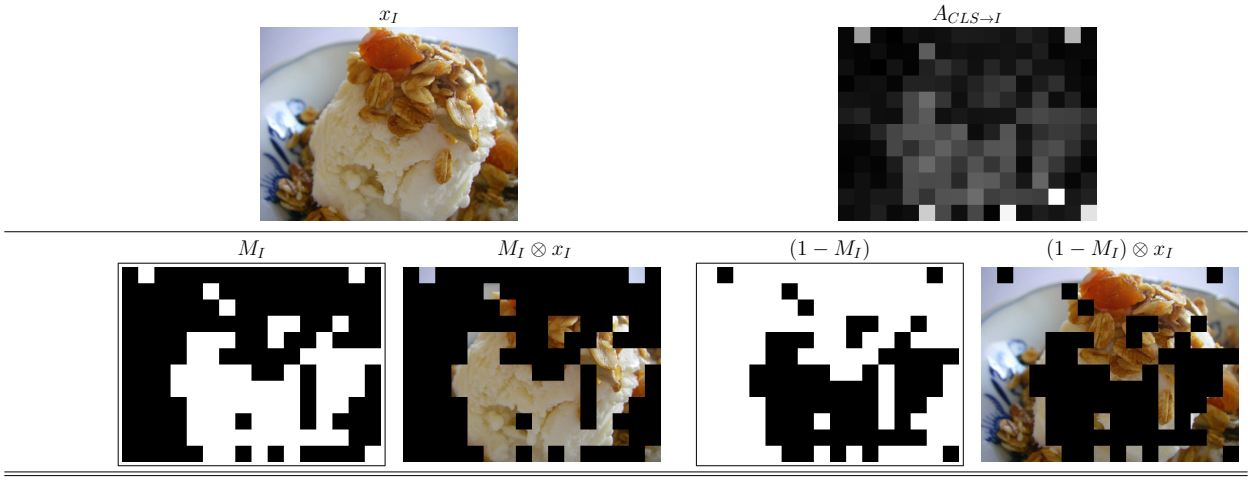
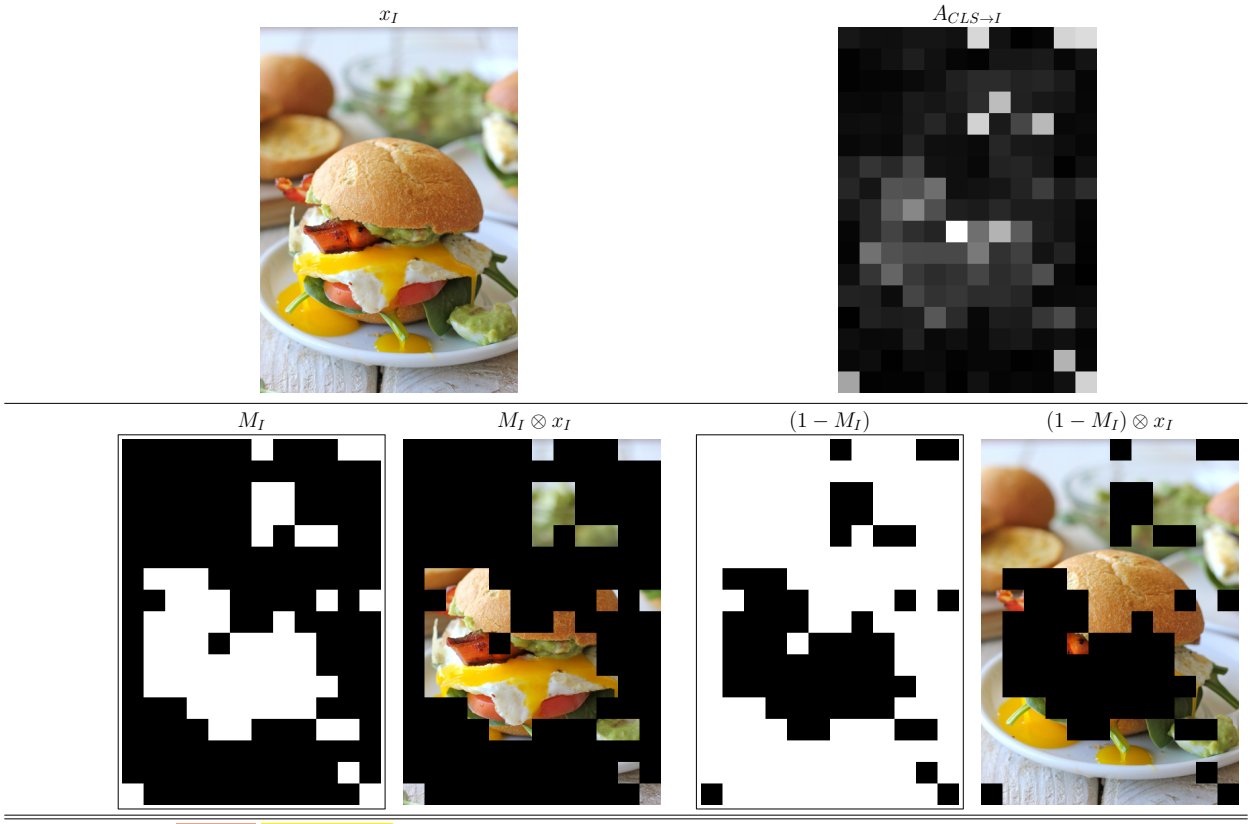


Figure 8. An example of exemplar masking on the UPMC Food-101 for the class "**chicken curry**".



$M_T \otimes x_T$: [CLS] ice cream flavor of the week : vanilla frozen yogurt with honey crunch granola — pink stripes [SEP]

Figure 9. An example of exemplar masking on the UPMC Food-101 for the class "**frozen yogurt**".



$M_T \otimes x_T$: [CLS] avocado club sandwich with spicy chipotle pepper spread - damn delicious [SEP]

Figure 10. An example of exemplar masking on the UPMC Food-101 for the class "**club sandwich**".