If you can describe it, they can see it: Cross-Modal Learning of Visual Concepts from Textual Descriptions

Supplementary Material

A. Impact and Limitations

We believe that Knowledge Transfer has the potential to be an impactful technique for introducing novel concepts in pre-trained models. Overall, Knowledge Transfer is quite cheap in terms of computational requirements, as it works by only fine-tuning on just a handful of synthesized samples. Thus, it is very quick and does not need a large amount of memory. In this sense, it may be comparable to parameter-efficient fine-tuning (PEFT) techniques, such as low-rank adaptation (LoRA) [8], which minimize the amount of memory required for fine-tuning. However, compared to PEFT, Knowledge Transfer does not require any real data besides a single textual description for each novel concept.

From this point onward, we will refer to the KT algorithm described in the main paper as **Explicit** Knowledge Transfer, namely the variant that relies on an inversion step to synthesize a visual example before fine-tuning. In contrast, we use the term **Implicit** Knowledge Transfer to denote approaches that avoid this inversion step and instead rely on shared parameters between modalities (e.g., multimodal neurons), enabling transfer through purely textual objectives such as MLM.

The main limitation of Explicit Knowledge Transfer lies in the inversion step, which takes the most time compared to fine-tuning. If this step could be avoided, we could achieve near real-time learning of novel concepts with minimal computational requirements. This could enable the development of rapidly improving intelligent agents in many real-world applications. We hypothesize this is possible with Implicit Knowledge Transfer, for example by using Masked-Language Modeling (MLM) as a proxy for knowledge transfer. However, in this work, we do not focus on this topic, as preliminary experiments (shown in Sec. D.3) did not achieve satisfactory results compared to Explicit Knowledge Transfer.

Another limitation lies in the limited comparison with state-of-the-art approaches; however, to the best of our knowledge, we are not aware of other works sharing the same goal as ours.

B. Knowledge Transfer

A general overview of explicit Knowledge Transfer can be found in Fig. 1.

B.1. Examples of inverted images

Examples of inverted images can be found in Fig. 2.

B.2. Possible improvements of Explicit Transfer

We start from the inversion equation:

$$\begin{aligned}\hat{X}_V^* &= f_V^{-1}(f_T(X_T)) \\ &\approx \max_{\hat{X}_V^*} sim(A(f_V(\hat{X}_V^*)), f_T(X_T)) + \alpha R(\hat{X}_V^*) \quad .\end{aligned} \tag{1}$$

B.2.1. Relaxation of Eq. 1

Computing $\hat{X}_V^*$ as in Eq. 1 might produce images that are widely different from the training distribution of natural images, as shown in Fig. 2. So, instead of inverting the whole visual encoder f_V, we can invert just a subset of layers $\Psi_V \subset f_V$, starting from the top of the model:

$$\begin{aligned}\hat{Z}_V^* = \Psi_V^{-1}(f_T(X_T)) \approx \max_{\hat{Z}_V^*} sim(\Psi_V(\hat{Z}_V^*), f_T(X_T)) \\ + R(\hat{Z}_V^*)\end{aligned} \tag{2}$$

where R could be a regularization similar to style transfer [6] to encourage $\hat{Z}_V^*$ to be similar to the intermediate representations of natural images.

B.3. Implicit Knowledge Transfer

Although in this work we focus on explicit knowledge transfer, we briefly present the idea behind Implicit Knowledge Transfer for the sake of completeness. It has been shown how multi-modal neurons can be found in multi-modal models [7, 24]. These neurons exhibit high activation on the same concepts in either modality, meaning that they are able to capture cross-modal representations. We hypothesize that in a shared-parameter architecture (e.g. early-fusion transformers [15, 20]) it should be possible to exploit these neurons for knowledge transfer, for example with simple masked language modeling on the novel concept description, effectively eliminating the need for model inversion. For this purpose, early-fusion architectures that can process single modalities independently would be required. However, to the best of our knowledge, we are not aware of many large pre-trained models satisfying these requirements at the time being, hence we leave an in-depth exploration of this path for future research. Hints that training on different modalities independently can help can be found in the literature, for example during pre-training of U-VisualBERT [18]. Even more relevant to our research, authors in [28] report some capabilities of cross-modal transfer on SimVLM, however,

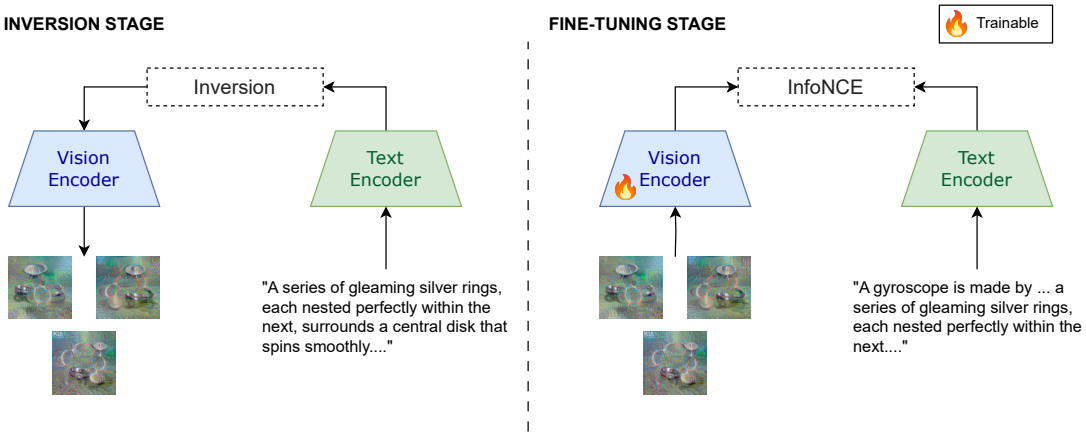


Figure 1. **Graphical overview of Knowledge Transfer.** Starting from a textual description of the target concept, we synthesize images via model inversion (left) then, using an image-text matching loss, we fine-tune the visual encoder to match the concept (right). In this way, we leverage prior knowledge contained in the model (from pre-training) to learn novel concepts.

(a) Moongate. Caption: *A perfectly circular archway built from uniformly cut stones or bricks, set into a larger wall. It forms a smooth circle, framing views of gardens or landscapes beyond, creating a picturesque portal.*

(b) Tonometer. Caption: *A slender, pen-like probe attached to a small base equipped with precise dials and gauges. This tool is often part of a larger medical apparatus, featuring a metallic finish and a refined, professional appearance.*

Figure 2. Example of inverted images (top) and real images (bottom) from rare concepts that CLIP struggles to classify correctly.

the model is proprietary and we are unable to reproduce their claims. Thus, here we focus on ViLT and we report some preliminary analysis in Sec. D.3.

B.4. Open questions

Q1 Domain Gap. Inverted images, as shown in Fig. 2, appear widely different from natural images. However, as shown by the results in the paper, fine-tuning the models on them leads to improved results. Is a domain gap present between inverted and real images? Or is it indicative of a fundamental difference in which deep models process visual information? This phenomenon may be linked with adversarial attacks [29].

Q2 Generalizability of inversion An interesting point to analyze, which could provide some insights for Q1, is the generalizability of the inverted images. For example, can images inverted with a certain model (e.g. CLIP) be used for training some other model from scratch? Or are they "fitted" to only work with the specific model used for inversion?

Q2 Catastrophic Forgetting To what extent can we prevent catastrophic forgetting when applying Knowledge Transfer? In this work, we show that lower learning rates generally achieve a good trade-off between learning novel concepts and preserving previous information. However, there is still room for improvement. For example, LoRA [8] has been shown to help in avoiding catastrophic forgetting during fine-tuning, hence applying it during Knowledge Transfer could further improve the results. Also, Implicit Transfer (on shared-parameter models) might avoid catastrophic forgetting better than Explicit Transfer, for example by focusing on multi-modal neurons.

C. Experimental Setup

C.1. Training Details

Captioning To produce descriptive captions for the new concepts, we employ a LLM-based approach. Specifically, for natural images, we use Llama-3 Instruct (with 8B parameters) [2] with the following prompt: *"Generate a small*

description of the ImageNet class <class name> without using the word itself. The description must contain visual cues useful for recognizing the subject with low-level and accurate details. Please don't insert anything else in the response except the description.", where we insert the appropriate class name for each new concept. Note that we employ an LLM only for the sake of convenience (e.g. captioning all 1000 ImageNet classes), but this is not a requirement. For medical data, we actually employ a mix of hand-crafted captions based on Radiopaedia [27] augmented with some elements from ChatGPT-4 [21]. All captions can be found in the supplementary material.

Inversion We run inversion for 5k steps, using a cosine learning rate annealing schedule. For the regularization term, we use the default value $\alpha = 0.005$ [14]. The augmentation we employ is composed of random affine transformations (rotation comprised between -30 and +30 degrees, a translation of 10%, and a scaling comprised between 70% and 100% of the image size), with a probability of 0.5. An example of inverted images can be found in Fig. 2. For each concept, we generate ten inverted samples.

Fine-tuning Fine-tuning is performed using the InfoNCE loss to achieve alignment between the inverted images and the textual descriptions. We only fine-tune the visual encoder, while keeping the text encoder frozen. The motivation is that we wish to align features extracted from the visual encoder to those extracted from the text encoder. For most experiments, we perform a quick fine-tuning consisting of only one single epoch, with small learning rates between 10^{-6} and 10^{-4}. For CLIP-based models, we generally employ a weight decay of 0.2 as in [22]. More details are provided in the description of each experiment.

C.2. Datasets

We employ a variety of datasets in different domains and for different downstream tasks. Here we provide a complete list, divided by task. Note that we do not use any training data from these datasets, as we only use them for testing. All improvements come from the textual description.

Natural images classification
1.) RareConcepts is a collection of images of rare concepts gathered from the web. We release the dataset as part of this work. In our experiments, we focus on concepts that are relatively unknown to different large multi-modal architectures: Moongate, Gyroscope and Tonometer, together with four additional fine-grained or deformable categories and geometric patterns (fabric patterns: Bengal Strip, Madras, Floral; parquet pattern: Chantilly). For each concept, we collect 10 images.

2.) ImageNet-1k [4] is a large-scale benchmark for visual recognition, with 1000 classes and 3.2M natural images.

Medical images classification
3.) CheXpert-2x500c [10] is a dataset of Chest X-Rays obtained from the large-scale CheXpert dataset [11] by considering 200 examples for the classes Atelectasis, Cardiomegaly, Edema, Consolidation, and Pleural Effusion.
4.) JSRT [26] is a Chest X-Ray dataset containing 154 conventional chest radiographs with a lung nodule of different types (malignant and benign nodules).

Medical images segmentation
5.) UnitoChest [3] is a collection of 306,440 chest CT slices coupled with nodules segmentation masks. We consider slices where nodules are present, for a total of 4179 images.
6.) UDIAT [30] is a dataset of breast masses in ultrasound images, containing 110 benign and 54 malignant cases.
7.) SIIM Pneumothorax [33] is a Chest X-ray dataset for pneumothorax segmentation, released as a challenge in 2019. We consider a total of 500 images.
8.) BraTS23 Glioma [1] is a brain MRI dataset of adult patients with brain gliomas. We consider all slices where a tumor is present for a total of 14,746 images.

Image-Text retrieval and captioning
9.) Flickr30k [31] is a dataset of 31,783 images from Flickr, each one associated with 5 captions provided by human annotators. For our experiments, we used Karpathy's test split [13], which contains 1000 images and 5000 captions.
10.) MSCOCO [19] is a large-scale dataset of more than 330k images with textual captions. We use Karpathy's test split [13], containing 5000 images.

C.3. Controlled Experiments

C.3.1. Rare Concepts (CLIP and ViLT)

We train using the Adam optimizer, with a batch size of 4, a weight decay of 0.2, and learning rates between 1e-5 and 5e-5 as reported in the table in the main text. We train using 10 inverted images for each concept. The captions used for inversion can be found in Tab. 10.

Details about image inversion for ViLT For ViLT we use a slightly different approach, in order to accommodate the different architecture. To run inversion, we start from a pair of input $< x_t, \hat{x}_v^* \sim N(0;1) >$ composed by the textual caption and random noise. We then optimize $\hat{x}_v^*$ by optimizing the image-text matching (ITM) score computed on the ITM head of ViLT [15]. This head outputs two values: one indicating no match and the other indicating a match. To optimize this, we use the cross-entropy loss during inversion, aiming to maximize the output corresponding to a match

while minimizing the output for no match. The rest of the setup is the same as CLIP. Furthermore, we disabled the random affine augmentation, as it produced noisy inverted images. Additionally, we use a weight decay value of 0.01, which is consistent with the one used by the authors of ViLT. The captions used for inversion with ViLT can be found in Tab. 11.

C.3.2. KT on Medical Images (MedCLIP)

For MedCLIP, we use the same setup as CLIP on rare concepts, see Sec. C.3.1. Namely, we employ Adam with a batch size of 4 and a weight decay of 0.2, using 10 inverted images for each concept. The descriptions used for inversion with MedCLIP can be found in Tab. 12.

C.3.3. CLIP on medical images (out of domain KT)

For ViT-B/32, we use a learning rate of 5e-5 with a batch size of 8, and we train for 5 epochs; for ViT-L/14 we use a learning rate of 1e-5, a batch size of 4, and we train for 2 epochs. The captions used for inversion are reported in Tab. 13.

C.4. Real-world experiments

C.4.1. Captioning (CoCa)

In these experiments, we deal with two types of captions: the first is the *concept caption*, that we use for inversion and fine-tuning with InfoNCE as in all other experiments (listed in Sec. G), the second is the *target caption* that we use to fine-tune the autoregressive captioning decoder of CoCa with $\mathcal{L}_{cap}$.

Captioning Loss When fine-tuning on inverted images, we apply an autoregressive captioning loss, as defined in [32]:

$$\mathcal{L}_{cap} = -\sum_{t=1}^{T} \log P_\theta(y_t | y_{<t}, x) \tag{3}$$

which aims at predicting the next token y_t given the previous tokens $y_{<t}$ and the image x. The final objective function that we optimize is the combination of the InfoNCE loss and the captioning loss:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{CLIP} + \lambda_2 \mathcal{L}_{cap} \tag{4}$$

where $\lambda_1, \lambda_2 \geq 0$. In our fine-tuning, we use $\lambda_1 = 1$ and $\lambda_2 = 0.1$.

Target captions template We use a set of 26 different templates as target captions during fine-tuning. At each optimization step, we select a random template for each sample in the following manner:

```
TEMPLATES = (
 lambda c: f'a bad photo of a {c}.',
 lambda c: f'a low resolution photo of the {c}.',
 lambda c: f'a rendering of a {c}.',
 lambda c: f'a bad photo of the {c}.',
 lambda c: f'a cropped photo of the {c}.',
 lambda c: f'a photo of a hard to see {c}.',
 lambda c: f'a bright photo of a {c}.',
 lambda c: f'a photo of a clean {c}.',
 lambda c: f'a photo of a dirty {c}.',
 lambda c: f'a dark photo of the {c}.',
 lambda c: f'a photo of my {c}.',
 lambda c: f'a bright photo of the {c}.',
 lambda c: f'a cropped photo of a {c}.',
 lambda c: f'a photo of the {c}.',
 lambda c: f'a good photo of the {c}.',
 lambda c: f'a rendering of the {c}.',
 lambda c: f'a photo of one {c}.',
 lambda c: f'a close-up photo of the {c}.',
 lambda c: f'a photo of a {c}.',
 lambda c: f'a low resolution photo of a {c}.',
 lambda c: f'a photo of a large {c}.',
 lambda c: f'itap of the {c}.',
 lambda c: f'a jpeg corrupted photo of the {c}.',
 lambda c: f'a good photo of a {c}.',
 lambda c: f'itap of a {c}.',
 lambda c: f'a photo of the large {c}.',
)

template_idx = torch.randint(
    0, len(TEMPLATES), (1,)
).item()
template = TEMPLATES[template_idx]
return tokenize(template(class_name))
```

These templates are inspired by OpenAI prompt ensembling for zero-shot classifiers[1]. We use these captions although they are not in the exact style of MSCOCO, as we do not want to leverage information from MSCOCO besides the concept classes. Target captions crafted specifically for MSCOCO might further improve the results.

D. Additional Results

D.1. Controlled Experiments

D.1.1. Rare Concepts (CLIP and ViLT)

Here we report results across differen learning rate values, in Fig. 3. Results in numerical forms can be found in Tab.1.

D.1.2. KT on Fine-grained and deformable rare-concepts

Tab. 2, we perform additional experiments on fine-grained categories, including both deformable (fabric) and non-deformable (parquet) patterns.

D.1.3. KT on medical images (MedCLIP)

The full results on JSRT with MedCLIP can be found in Tab. 3.

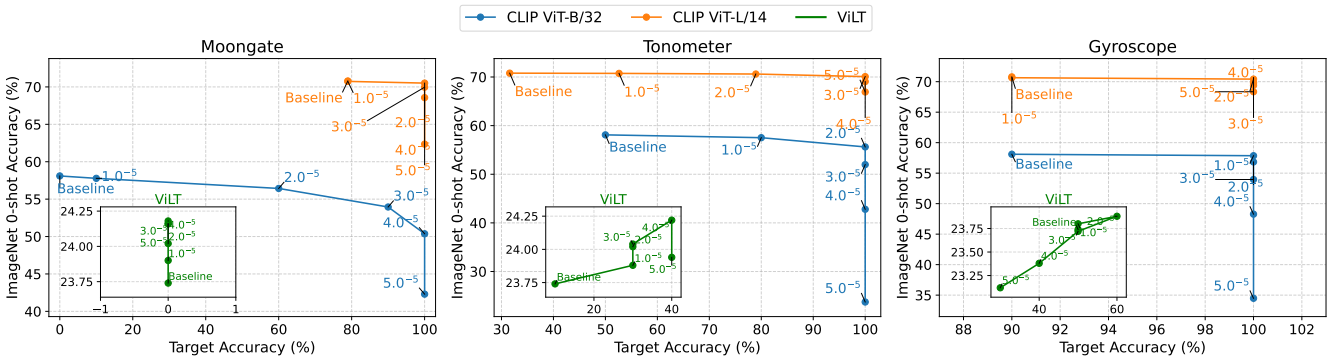


Figure 3. **Knowledge Transfer (KT) on novel and rare concepts** (CLIP and ViLT) across different learning rates. In most instances, we achieve improvement (even notable) in the target accuracy on the novel concept, preserving original knowledge (measured as accuracy on ImageNet). We also observe that on ViLT the accuracy on ImageNet generally improves when performing KT.

				Learning Rate				
Model	**Concept**		**Baseline**	**1e-5**	**2e-5**	**3e-5**	**4e-5**	**5e-5**
CLIP ViT-B/32 [22]	Moongate	Target Acc.	0%	10%	**60%**	90%	100%	100%
		ImageNet 0-shot	58.10%	57.78%	**56.43%**	53.95%	50.37%	42.30%
	Tonometer	Target Acc.	50%	**80%**	80%	100%	100%	100%
		ImageNet 0-shot	58.10%	**57.52%**	55.62%	51.98%	42.80%	23.73%
	Gyroscope	Target Acc.	90%	**100%**	100%	100%	100%	100%
		ImageNet 0-shot	58.10%	**57.86%**	56.84%	53.96%	48.28%	34.48%
CLIP ViT-L/14 [22]	Moongate	Target Acc.	78.95%	78.95%	**100%**	100%	100%	100%
		ImageNet 0-shot	70.79%	70.74%	**70.51%**	69.96%	68.57%	62.35%
	Tonometer	Target Acc.	31.58%	52.63%	**78.95%**	100%	100%	100%
		ImageNet 0-shot	70.79%	70.74%	**70.61%**	70.08%	69.06%	66.92%
	Gyroscope	Target Acc.	90%	90%	**100%**	100%	100%	100%
		ImageNet 0-shot	70.79%	70.65%	**70.42%**	69.84%	69.39%	68.35%
ViLT [15]	Moongate	Target Acc.	0%	0%	0%	0%	**0%**	0%
		ImageNet* 0-shot	23.74%	23.90%	24.02%	24.16%	**24.18%**	24.16%
	Tonometer	Target Acc.	10%	30%	30%	30%	**40%**	40%
		ImageNet* 0-shot	23.74%	23.88%	24.02%	24.04%	**24.22%**	23.94%
	Gyroscope	Target Acc.	50%	**60%**	50%	50%	40%	30%
		ImageNet* 0-shot	23.74%	**23.80%**	23.88%	23.72%	23.38%	23.12%

Table 1. **Knowledge Transfer on novel and rare concepts** (CLIP and ViLT) in terms of accuracy. * for VilT, we employ ImageNet-100 [12] due to the computational requirements of evaluating every possible image-caption pair for zero-shot classification.

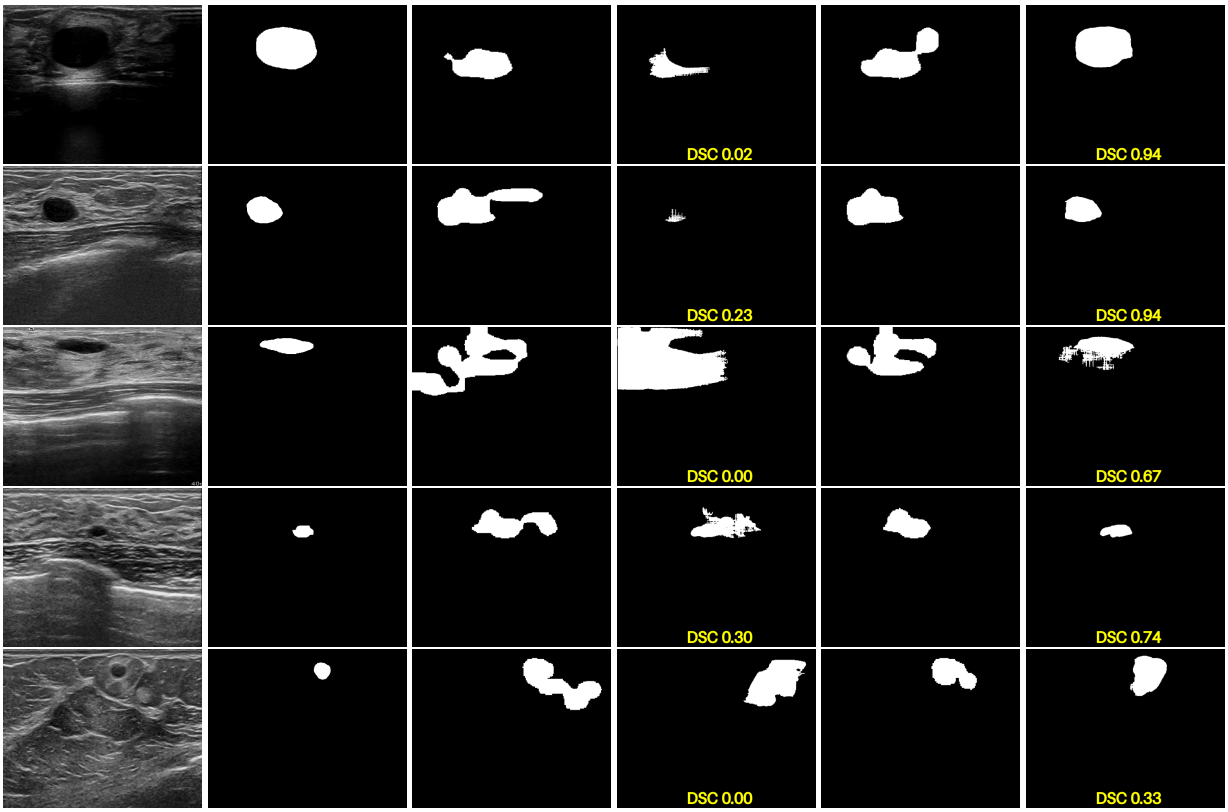
D.2. Real-world Experiments

D.2.1. Segmentation

Results of knowledge transfer on MedCLIP-SAMv2 with different values of learning rate are shown in Tab. 4. We report illustrative examples of the improvements achieved by knowledge transfer in Fig. 4 and Fig. 5. The captions used for inversion for segmentation can be found in Tab. 14.

[1] https://github.com/mlfoundations/open_clip/blob/main/src/open_clip/zero_shot_metadata.py

Differences in downstream tasks As said in the main text, lung nodules and pneumothorax segmentations are novel tasks on which MedCLIP-SAMv2 was not pre-trained. Regarding brain tumors, we employ the BraTS 2023 glioma dataset, which contains brain gliomas in adult patients. With respect to the original performance reported in [16] on brain tumors, we notice a significant gap. However, the preprocessing of the images is quite different, as data from BraTS 2023 is more heavily preprocessed (e.g. skull stripping) than in [16]. We were not able to compare MedCLIP-SAMv2 on



(a) Image (b) Ground Truth (c) M2IB Map (Baseline) (d) Final Segmentation (Baseline) (e) M2IB Map (Knowledge Transfer) (f) Final Segmentation (Knowledge Transfer)

Figure 4. Qualitative evaluation of knowledge transfer on breast tumor segmentation (UDIAT dataset). We report the top ten most illustrative examples in which knowledge transfer improved segmentation, in terms of DSC.

	Fabric			Parquet
Model	**Bengal Stripe**	**Floral**	**Madras**	**Chantilly**
CLIP ViT-B/32	0%	10%	0%	20%
CLIP ViT-B/32 + KT (ours)	**20%**	**30%**	**10%**	**70%**
ImageNet (KT)	57.08%	57.63%	57.33%	56.12%
CLIP ViT-L/14	70%	60%	80%	0%
CLIP ViT-L/14 + KT (ours)	**90%**	**80%**	**90%**	**10%**
ImageNet (KT)	70.82%	70.73%	70.79%	69.90%

Table 2. **KT applied to fine-grained and deformable categories** (fabric and parquet patterns), expanding the Rare Concepts dataset.

the original data, as, at the time of writing, details about the data split are missing.

D.2.2. Text-image retrieval

In this section, we show the full results for text and image retrieval tasks on Flickr30k with ViLT. Tab. 5 is the extended version of the results in which we report the huggingface's pre-trained baseline, along with the results of the experiments we performed while tuning the learning rate and the batch size. We report the best batch size for each learning rate. As can be seen our method works best with smaller learning rates in this setting. The captions used for inversion (mscoco) can be found in Tab. 15.

D.2.3. Captioning

We report additional results of captioning, with and without the use of captioning loss in Tab. 6. Even without captioning loss, we are achieve improvements over both version CoCa (pre-trained on LAION-2B) and CoCa FT (fine-tuned on MSCOCO). By applying $\mathcal{L}_{cap}$ we achieve a further increase in the reported metrics. We showcase some improvements in captioning on the CoCa model pre-trained on LAION-2B in Fig. 7. The concept captions used for inversion can be found in Tab. 15.

D.3. Preliminary results with Implicit Knowledge Transfer

In this section, we show preliminary results about Implicit Knowledge Transfer, presented in Sec. B.3. In Implicit Knowledge Transfer, the objective is to teach the model a novel concept by only training it on text, without using inverted images. To do so with ViLT, we no longer use the image-text matching objective as it requires images, instead,

Concept		Baseline	Learning Rate (multiplier)				
			×1	×2	×3	×4	×5
Benign Nodule	Target Acc. (base lr 1e-5)	54.55%	54.55%	**54.55%**	54.55%	54.55%	54.55%
	CheXpert-5x200c 0-shot	62.10%	61.80%	**62.30%**	62.10%	62%	62.20%
Lung Cancer	Target Acc. (base lr 1e-4)	83.93%	87.50%	**92.86%**	94.64%	92.86%	92.86%
	CheXpert-5x200c 0-shot	62.10%	62.20%	**61.50%**	53.70%	48.20%	44.50%

Table 3. **Knowledge Transfer on MedCLIP on the JSRT dataset** (accuracy). Full results across learning rates.

Model	Lung Nodules†			Lung Pneumothorax†			Breast Ultrasound			Brain MRI		
	DSC	NSD	IoU	DSC	NSD	IoU	DSC	NSD	IoU	DSC	NSD	IoU
MedCLIP-SAMv2	**14.83%**	17.30%	8.64%	6.30%	7.61%	3.75%	56.25%	59.44%	47.81%	17.20%	20.97%	12.05%
Transf. (1e-5)	13.95%	17.45%	8.75%	6.28%	7.59%	3.77%	**58.23%**	**61.56%**	**49.52%**	15.90%	19.36%	11.10%
Transf. (2e-5)	14.10%	17.65%	8.83%	**6.41%**	**7.76%**	**3.83%**	54.36%	57.30%	46.30%	**18.13%**	**22.26%**	**12.62%**
Transf. (3e-5)	14.10%	17.65%	8.85%	6.25%	7.55%	3.73%	55.70%	59.00%	47.49%	15.47%	18.85%	10.78%
Transf. (4e-5)	14.25%	17.85%	8.94%	6.24%	7.57%	3.71%	53.86%	56.82%	45.61%	15.26%	18.63%	10.62%
Transf. (5e-5)	14.20%	17.78%	8.92%	6.20%	7.51%	3.70%	54.90%	57.97%	46.09%	16.22%	19.81%	11.34%
Transf. (1e-4)	14.35%	**18.03%**	**9.04%**	6.02%	7.29%	3.59%	-	-	-	-	-	-
Transf. (2e-4)	10.74%	13.64%	6.66%	4.71%	5.54%	2.86%	-	-	-	-	-	-

Table 4. Full results on zero-shot segmentation with MedCLIP-SAMv2.

we employ masked language modeling (MLM) [5], using as input a pair composed of the textual description of the concept and random noise instead of the image. The assumption is that, in a model with parameters shared between modalities, fine-tuning on one modality (text), will also benefit the other modality. Here, our hypothesis is that during fine-tuning, multi-modal neurons [24] can help in transferring knowledge across modalities.

D.3.1. Implicit Knowledge Transfer with MLM

For Implicit Knowledge Transfer we used the same masked language modeling setup as in ViLT [15], which means that we use whole-word masking and a masking probability of 15%. We use 10 examples for fine-tuning, each of which is composed by a random noise image and a masked caption. The masked captions are generated starting from the same caption by masking differently each time. For the caption we use the template "A X is Y", where X is the name of the concept and Y is the concept's description (from Tab. 11). We use a batch size of 4 with different learning rates, for a total of 3 train steps. Weight decay is set, as in the other experiments, to 0.01.

Explicit Knowledge Transfer baseline with MLM For comparison, we also evaluate the results of explicit knowledge transfer with the masked language modeling objective instead of the image-text matching objective. We use the same setup as the implicit one, with the only exception that instead of random noise images, we use inverted images. In particular, we use the same inverted images we used for the explicit knowledge transfer with the image-text matching objective.

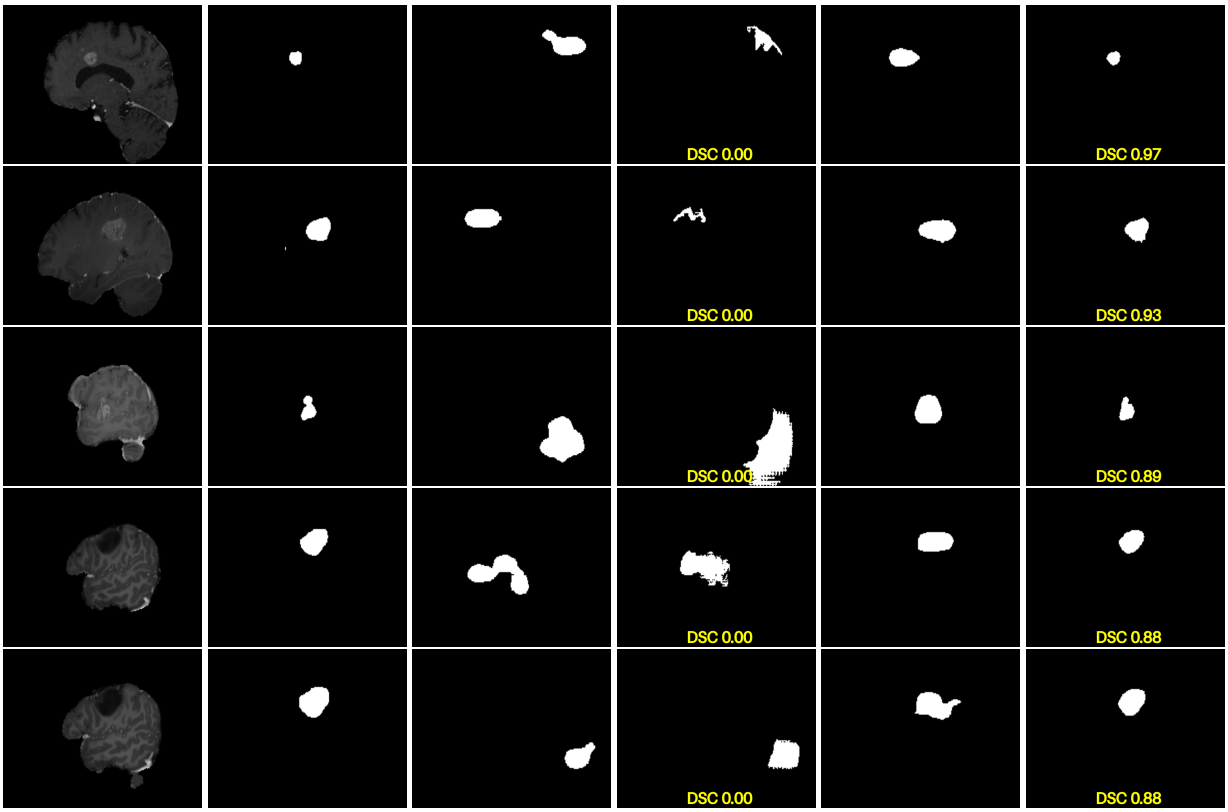
D.3.2. Results discussion

Tab. 8 reports the results for both implicit and explicit knowledge transfer with masked language modeling. In both cases, no improvements are observed for the moongate concept, whose accuracy stays at 0%. For tonometer, explicit knowledge transfer seems to work better since with the implicit one, there is a loss of performance, while for gyroscope the opposite is true. In all cases, we observe an increase in the accuracy over the ImageNet-100 classes, as observed when using image-text matching objective. The only improvement is registered for the gyroscope concept in the implicit transfer setting, from 50% to 60%. Overall we can say that implicit knowledge transfer with masked language modeling does not work for the ViLT model, this is probably due to the fact that ViLT was pre-trained on image-text pairs, which means that it expects both modalities in input. Regarding explicit knowledge transfer with MLM, more experiments are needed to determine the correct algorithm and set of hyperparameters to make it work, for example, we may have to use more examples generated from different textual descriptions.

E. Ablation studies

E.1. Fine-tuning strategy

We perform an ablation study on our fine-tuning strategy. In our experiments, during fine-tuning, we freeze the text encoder and only train the visual encoder. Here we evaluate fine-tuning with different configurations. The results are illustrated in Fig. 6. When fine-tuning both encoders, we



(a) Image (b) Ground Truth (c) M2IB Map (Baseline) (d) Final Segmentation (Baseline) (e) M2IB Map (Knowledge Transfer) (f) Final Segmentation (Knowledge Transfer)

Figure 5. Qualitative evaluation of knowledge transfer on brain tumor segmentation (BraTS 2023 glioma dataset). We report the top ten most illustrative examples in which knowledge transfer improved segmentation, in terms of DSC.

			Flickr30k (1K)					
			Text Retrieval			Image Retrieval		
Model	LR	Batch Size	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
ViLT-B/32 (huggingface)	-	-	73.8%	93.5%	96.5%	57.3%	83.9%	90.4%
ViLT-B/32	8e-7	32	74.5%	93.8%	96.4%	57.7%	84.0%	90.4%
ViLT-B/32	9e-7	32	**74.6%**	93.8%	96.4%	**57.8%**	84.0%	90.5%
ViLT-B/32	1e-6	16	74.4%	93.8%	96.5%	57.7%	**84.1%**	90.5%
ViLT-B/32	2e-6	128	**74.6%**	93.7%	96.5%	**57.8%**	84.0%	90.5%
ViLT-B/32	3e-6	256	74.5%	**93.9%**	96.5%	57.7%	83.9%	90.5%
ViLT-B/32	4e-6	32	73.8%	93.6%	96.5%	57.4%	84.0%	90.5%
ViLT-B/32	5e-6	256	74.5%	**93.9%**	96.5%	57.6%	84.0%	90.5%
ViLT-B/32	8e-6	32	73.2%	93.7%	96.1%	57.4%	83.7%	90.4%
ViLT-B/32	1e-5	128	74.4%	93.8%	**96.8%**	56.8%	83.7%	**90.6%**
ViLT-B/32	2e-5	32	71.8%	93.2%	96.4%	56.7%	83.6%	90.4%
ViLT-B/32	3e-5	32	70.8%	92.1%	95.7%	56.0%	82.9%	90.2%

Table 5. Full results for text and image retrieval on Flickr30k with ViLT. The first section reports baseline results, while the second shows the outcome of each tested learning rate and its optimal batch size (chosen among 16, 32, 64, 128, and 256). Recall scores at top 1, 5, and 10 are reported.

Model	MSCOCO (5K) BLEU@4	METEOR	CIDEr	SPICE
CLIP-ViL [25]	40.2	29.7	134.2	23.8
BLIP [17]	40.4	-	136.7	-
VinVL [34]	41.0	31.1	140.9	**25.4**
SimVLM [28]	40.6	33.7	143.3	25.4
LEMON [9]	**41.5**	30.8	139.1	24.1
CoCa [32] (proprietary)	40.9	**33.9**	**143.6**	24.7
CoCa	6.9	12.8	31.1	9.1
CoCa (transf. 6e-5)	13.6	18.5	47.3	13.6
CoCa† (transf. 9e-5)	**17.9**	**19.4**	**60.8**	**13.7**
CoCa FT	34.9	29.7	123.1	**23.5**
CoCa FT (transf. 2e-5)	35.2	29.8	123.1	23.2
CoCa FT† (transf. 5e-6)	**35.2**	**29.8**	**124.0**	23.3

Table 6. Image captioning on MSCOCO. † means the decoder is also fine-tuned. CoCa refers to the baseline model pre-trained on LAION-2B [23], while CoCa FT refers to the model fine-tuned for captioning on MSCOCO. We highlight in bold the best results overall and the improvements achieved by Knowledge Transfer.

observe a rapid collapse of target accuracy and ImageNet accuracy for all concepts. We also observe a similar trend when fine-tuning the text encoder only, while leaving the visual encoder frozen. This is, however, expected as our assumption is that the knowledge contained in the text encoder is already good enough to represent the target concept, and we just wish to align visual features to it. Moreover, if we alter the text encoder weights, correspondence between captions and inverted images may be lost, leading to degenerate cases.

E.2. Captions construction

We focus on the construction of the captions for fine-tuning. As explained in the main text, during fine-tuning we prepend each caption with the name of the concept, for example *"A moongate is [...]"*. Here we motivate why this is necessary by comparing captions prepended with the name and captions without the name. The results are shown in Fig. 7. As we can observe, using the name of the concept during fine-tuning is necessary in order to map visual features to its textual description.

E.3. Captions Quality

In Tab. 9 we reports the prompts used in the ablation study on prompt quality and length, with P1 indicating a short human-written prompt, P2 mid-sized human caption with some clear visual hints, and P3 the original LLM-generated description that we used in our experiments (which can be found in Sec. G).

F. Code

Code will be publicly released upon paper acceptance.

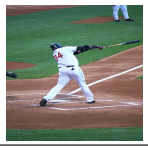
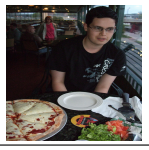
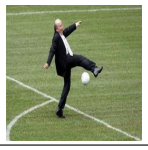
	MSCOCO				
Sample					
Actual	A shot of a clock in the train station.	A pianist in a suit and glasses playing a keyboard.	A baseball player hitting a ball in a professional game.	A baseball holding a baseball bat during a baseball game.	Two people that are sitting on a table.
Baseline	grand - central - station - new - york.jpg (METEOR 0.0)	Paul Kagasoff, president and chief executive officer of Intel corp., speaks during the 2012 Computex trade show in Taipei, Taiwan (METEOR 8.6)	Aaron judge 2016 new york Yankees (METEOR 5.0)	20080419 _ mariners _ 0001 | by Mike. Smith (METEOR 4.2)	Dinner in our tiny studio apartment in Amsterdam. (METEOR 0.0)
Knowledge Transfer	A black and white photo of a clock at Grand Central terminal. (METEOR 92.9)	A photo of a man in a suit sitting at a keyboard. (METEOR 86.1)	A baseball player swings his bat at a batter. (METEOR 83.6)	A baseball player takes a swing at a pitch. (METEOR 86.9)	A photo of a man and a woman sitting at a kitchen table. (METEOR 80.8)
Sample					
Actual	A baseball batter up at the plate that just hit a ball	A pair of red scissors sitting on a newspaper.	A young man sitting at a table with a pizza.	A man in a black suit kicking around a soccer ball.	The man in a black tie sits in a chair with his shirt sleeves rolled up.
Baseline	david - ortiz _ display _ image _ display _ image _ display _ image _ display _ image _ display _ image _ display _ image (METEOR 3.9)	Your scissors are now digitized (METEOR 9.5)	Pizza and beer in Chicago, Illinois. Photo via Flickr: David J. Abbott M.D. (METEOR 8.3)	blatter - foot - ball.jpeg (METEOR 5.6)	Avatar for Marc Anthony (METEOR 5.3)
Knowledge Transfer	A baseball player swings his bat at a pitch. (METEOR 80.3)	A photo of a pair of red scissors on a piece of paper. (METEOR 84.9)	A photo taken of a man sitting at a table with a plate of pizza. (METEOR 83.5)	A man in a suit kicking a soccer ball on a field. (METEOR 80.7)	A man in a white shirt and black tie sits on a chair. (METEOR 78.4)

Table 7. Visual example of captioning on MSCOCO. We report the top ten most illustrative examples in which knowledge transfer improved captioning, in terms of METEOR score.

Type	Concept		Baseline	Learning Rate 1e-5	2e-5	3e-5	4e-5	5e-5
Implicit	Moongate	Target Acc.	0%	0%	0%	**0%**	0%	0%
		ImageNet* 0-shot	23.74%	23.82%	23.90%	**23.98%**	23.94%	23.86%
	Tonometer	Target Acc.	10%	10%	**10%**	10%	10%	0%
		ImageNet* 0-shot	23.74%	23.84%	**23.86%**	23.70%	23.64%	23.60%
	Gyroscope	Target Acc.	50%	50%	**60%**	60%	60%	50%
		ImageNet* 0-shot	23.74%	23.74%	**23.62%**	23.42%	23.44%	23.46%
Explicit	Moongate	Target Acc.	0%	0%	0%	0%	0%	**0%**
		ImageNet* 0-shot	23.74%	23.80%	24.08%	24.02%	24.10%	**24.20%**
	Tonometer	Target Acc.	10%	**10%**	10%	10%	10%	10%
		ImageNet* 0-shot	23.74%	**23.80%**	23.74%	23.72%	23.70%	23.56%
	Gyroscope	Target Acc.	50%	50%	**50%**	50%	40%	30%
		ImageNet* 0-shot	23.74%	23.74%	**23.84%**	23.84%	23.84%	23.82%

Table 8. Knowledge Transfer on novel and rare concepts using masked language modeling with ViLT. In the Implicit Knowledge Transfer, we pass noise images along with a corresponding masked caption to ViLT; in the explicit one, we replace noise images with inverted images.

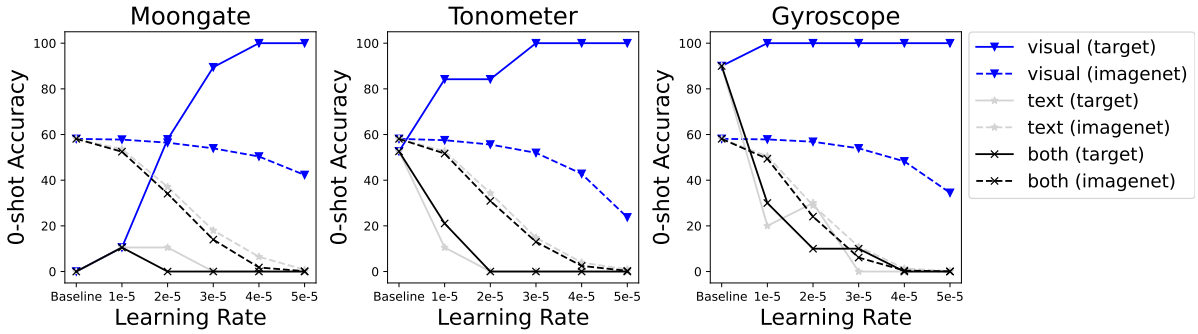


Figure 6. Comparison of fine-tuning strategies. Fine-tuning both the text and the visual encoders, or just the text encoder leads to a collapse in accuracy. Fine-tuning only the visual encoder correctly aligns prior visual features to the novel concept. A good choice of learning rate leads to higher accuracy on the novel concept (target) while limiting catastrophic forgetting on previous tasks (imagenet).

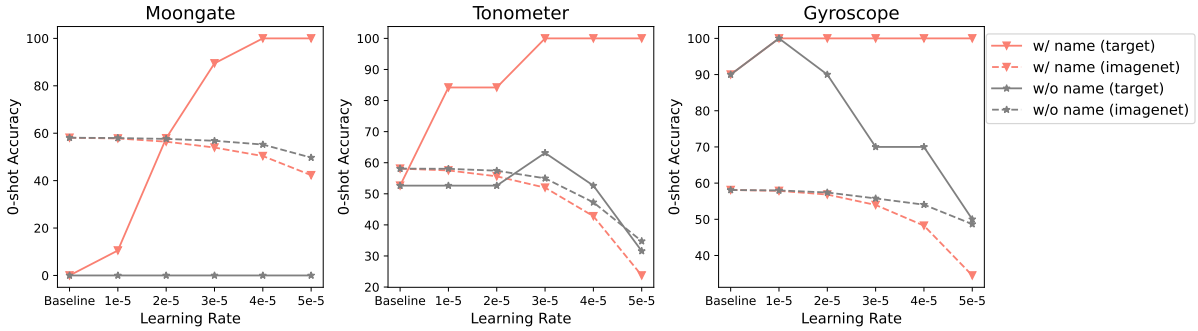


Figure 7. Ablation study on caption construction for finetuning.

Moongate	**P1** A circular stone archway **P2** A circular archway built from uniformly cut stones or bricks **P3** *original LLM caption*
Tonometer	**P1** An instrument to measure pressure **P2** A pen-like instrument with dials and pressure gauges **P3** *original LLM caption*
Gyroscope	**P1** A device to measure orientation and angular velocity **P2** A measuring device composed of a spinning wheel or disc to measure orientation and angular velocity **P3** *original LLM caption*

Table 9. **Prompt ablations** (P1 - short human caption; P2 - mid human caption; P3 - longer LLM caption). MOVE TO SUPP. [Carlo]

G. List of captions

Table 10. Descriptions for rare concepts (generated with Llama-3-8B-Instruct).

Moongate	A perfectly circular archway built from uniformly cut stones or bricks, set into a larger wall. It forms a smooth circle, framing views of gardens or landscapes beyond, creating a picturesque portal.
Tonometer	A slender, pen-like probe attached to a small base equipped with precise dials and gauges. This tool is often part of a larger medical apparatus, featuring a metallic finish and a refined, professional appearance.
Gyroscope	A series of gleaming silver rings, each nested perfectly within the next, surrounds a central disk that spins smoothly. The rings are connected by intersecting axes, allowing the disk to tilt and rotate freely while maintaining a sophisticated, mechanical look.

Table 11. Manually shortened descriptions for rare concepts (to fit into ViLT's 40 token input)

Moongate	A perfectly circular archway built from uniformly cut stones or bricks, set into a larger wall. It forms a smooth circle, framing views of gardens, creating a picturesque portal.
Tonometer	A slender, pen-like probe attached to a small base equipped with precise dials and gauges. This tool is often part of a larger medical apparatus.
Gyroscope	A series of rings each nested within the next, surrounds a central disk that spins. The rings are connected by intersecting axes allowing the disk to rotate freely.

Table 12. Descriptions for medical classes for JSRT (Mix with Radiopaedia and ChatGPT-4).

Benign Nodule	A small, round spots appearing in Chest X-Ray, typically well-defined with smooth, regular borders. These spots are often uniformly dense and do not cause distortion of surrounding structures.
Lung Cancer	A dense and irregular mass on Chest X-Ray images often with spiked or uneven edges. It may appear in the lung's periphery or near the airways.

Table 13. Descriptions for medical classes for CheXpert-5x200c (obtained with a mix of Radiopaedia and ChatGPT-4).

Atelectasis	A small areas of collapsed lung. It is usually seen on Chest X-Rays as small volume linear shadows, usually peripherally or at lung bases, appearing more opaque and shrunken.
Cardiomegaly	Enlargement of the heart usually seen in Chest X-Rays. The central shadow of the chest appears enlarged, extending beyond half the width of the entire chest cavity.
Pleural Effusion	A collection of fluid between the lungs and the chest, which makes the area appear white and smooth in Chest X-Ray images. The area does not present visible lung markings.
Consolidation	An area inside the lungs that appears as branching low attenuating (lucent) bronchi surrounded by high attenuating (dense) consolidated/opacified alveoli on Chest X-Ray images.
Edema	An abnormal accumulation of fluid in the extravascular compartments of the lung, which makes the area whiter in Chest X-Ray images. It is usually present on both lungs.

Table 14. Descriptions for medical classes for segmentation (Mix with Radiopaedia and ChatGPT-4).

Lung Nodules	Circular spots appearing within the lung fields, with clear and defined edges in CT images. They are denser than the surrounding tissue, often appearing in shades of gray or white, with varying size.

Breast Tumor	A dark, irregularly shaped area is visible against the lighter surrounding tissue. The borders may appear uneven or spiculated, and the area is typically less uniform in texture. Shadowing can often be seen beneath the mass.
Pneumothorax	An abnormal collection of air in the pleural space, which allows the parietal and visceral pleura to separate and the lung to collapse. The pleura edge is thin and no lung markings are visible.
Brain Tumor	An irregular bright mass in brain MRI, often with thick and irregular margins, surrounded by vasogenic-type edema or fluid accumulation. It may also have a hemorrhagic component.

Table 15. Descriptions for MSCOCO classes used for text and image retrieval experiments (With ChatGPT-4).

person	A human figure, typically with visible head, torso, arms, and legs, in various postures.
bicycle	A two-wheeled vehicle with a frame, handlebars, and pedals, usually ridden by a person.
car	A four-wheeled enclosed vehicle with windows and doors, commonly seen on roads.
motorcycle	A two-wheeled motorized vehicle with a seat and handlebars, typically ridden by one or two people.
airplane	A large flying vehicle with wings and a tail, often seen with windows along the sides for passengers.
bus	A large, rectangular vehicle with many windows and seating rows, designed to carry multiple passengers.
train	A long, linked series of vehicles running on tracks, often with a locomotive at the front.
truck	A large vehicle with a separate cab and an open or enclosed cargo area for transporting goods.
boat	A small to medium-sized watercraft with a hull and often visible sails or an engine.
traffic light	A vertical or horizontal post with red, yellow, and green lights, used to control vehicle flow at intersections.
fire hydrant	A small, red, metal cylinder with nozzles on the side, often found on sidewalks for fire emergencies.
stop sign	A red, octagonal sign with the word "STOP" in white, used to indicate where vehicles must halt.
parking meter	A tall, narrow post with a small display and slot, used to pay for parking time.
bench	A long seat, often with a backrest, typically found in parks or public areas.
bird	A small animal with feathers, wings, and a beak, often shown perched or flying.
cat	A small, furry animal with pointed ears, whiskers, and a long tail, often seen sitting or grooming.
dog	A furry, four-legged animal with a tail, usually seen with a collar or leash.
horse	A large, four-legged animal with a mane and tail, often depicted standing or galloping.
sheep	A woolly animal with a round body, small head, and short legs, often seen in groups in fields.
cow	A large animal with a boxy body, horns, and a long face, often shown grazing or with an udder.
elephant	A massive, gray animal with a long trunk, large ears, and tusks.
bear	A large, sturdy animal with thick fur, rounded ears, and a short tail, often shown standing or walking on all fours.
zebra	A horse-like animal with black and white stripes across its body.
giraffe	A very tall animal with a long neck and legs, spotted coat, and small horns on its head.
backpack	A bag with shoulder straps, typically worn on the back and used for carrying personal items.
umbrella	A foldable, rounded canopy on a stick, used for protection from rain or sun.
handbag	A small to medium-sized bag with handles, often carried by hand and used to hold personal items.
tie	A long, narrow piece of fabric worn around the neck, often knotted at the collar of a shirt.
suitcase	A rectangular, boxy container with a handle, used for carrying clothes and personal items when traveling.
frisbee	A flat, round disc often made of plastic, used for throwing and catching.
skis	Long, narrow pieces of equipment attached to boots, used for gliding on snow.
snowboard	A flat, wide board attached to boots, used for sliding on snow.
sports ball	A round object of varying sizes, such as a soccer ball or basketball, used in sports.
kite	A lightweight object with a string, often shaped like a diamond or triangle, designed to fly in the wind.
baseball bat	A smooth, cylindrical wooden or metal stick used to hit a baseball.
baseball glove	A padded, leather glove worn on one hand, used to catch baseballs.

skateboard	A narrow board with wheels, used for rolling and performing tricks.
surfboard	A long, flat board used for riding waves in the ocean.
tennis racket	An oval-shaped frame with strings and a handle, used to hit a tennis ball.
bottle	A narrow-necked container with a cap, often used to hold liquids like water or soda.
wine glass	A stemmed glass with a wide bowl at the top, used for drinking wine.
cup	A small, handleless vessel used for drinking, usually made of ceramic or plastic.
fork	A utensil with multiple prongs, used to pick up food.
knife	A utensil with a long, sharp blade, used for cutting food.
spoon	A utensil with a shallow bowl at the end of a handle, used for eating or serving food.
bowl	A round, deep dish, often used to hold soup or other foods.
banana	A long, yellow fruit with a curved shape and soft interior.
apple	A round fruit, typically red or green, with a stem at the top.
sandwich	Two slices of bread with filling in between, such as meat, cheese, or vegetables.
orange	A round, orange-colored fruit with a thick, textured peel.
broccoli	A green vegetable with a tree-like shape, featuring a thick stalk and small florets.
carrot	A long, orange vegetable with a pointed end, often with green leaves at the top.
hot dog	A sausage in a bun, often with condiments like ketchup or mustard.
pizza	A round, flatbread topped with cheese, sauce, and various toppings, often cut into slices.
donut	A round, fried pastry with a hole in the middle, often glazed or topped with sprinkles.
cake	A sweet, layered dessert, often decorated with frosting or fruit.
chair	A piece of furniture with a backrest and four legs, designed for sitting.
couch	A large, cushioned seat with a backrest and arms, designed for multiple people.
potted plant	A plant growing in a container, often with green leaves or flowers.
bed	A large, rectangular piece of furniture for sleeping, with a mattress and pillows.
dining table	A flat, often rectangular surface with legs, designed for eating meals.
toilet	A porcelain fixture with a seat and flushing mechanism, used in bathrooms.
tv	A rectangular screen on a stand or wall, used for viewing shows and movies.
laptop	A portable computer with a hinged screen and keyboard.
mouse	A small, handheld device used to control a cursor on a computer screen.
remote	A small, rectangular device with buttons, used to control electronics like TVs.
keyboard	A flat, rectangular panel with keys, used for typing on computers.
cell phone	A handheld electronic device with a screen and buttons or touchscreen, used for communication.
microwave	A box-like appliance with a door, used for heating food quickly.
oven	A large appliance with a door and interior racks, used for baking or roasting.
toaster	A small appliance with slots, used to toast bread.
sink	A basin with a faucet, used for washing hands, dishes, or food.
refrigerator	A large, box-like appliance with doors, used to store perishable food at low temperatures.
book	A collection of pages bound together with a cover, containing text or images.
clock	A circular or rectangular device with hands or digital display, showing the current time.
vase	A decorative container, often made of glass or ceramic, used to hold flowers.
scissors	A handheld tool with two blades, used for cutting paper or fabric.
teddy bear	A soft, stuffed toy shaped like a bear, often used by children.
hair drier	A handheld device that blows warm air, used to dry hair.
toothbrush	A small brush with a handle, used for cleaning teeth.

References

[1] Maruf Adewole, Jeffrey D Rudie, Anu Gbdamosi, Oluyemisi Toyobo, Confidence Raymond, Dong Zhang, Olubukola Omidiji, Rachel Akinola, Mohammad Abba Suwaid, Adaobi Emegoakor, et al. The brain tumor segmentation (brats) challenge 2023: glioma segmentation in sub-saharan africa patient population (brats-africa). *ArXiv*, 2023. 3

[2] AI@Meta. Llama 3 model card. 2024. 2

[3] Hafiza Ayesha Hoor Chaudhry, Riccardo Renzulli, Daniele Perlo, Francesca Santinelli, Stefano Tibaldi, Carmen Cristiano, Marco

Grosso, Giorgio Limerutti, Attilio Fiandrotti, Marco Grangetto, et al. Unitochest: A lung image dataset for segmentation of cancerous nodules on ct scans. In *International Conference on Image Analysis and Processing*, pages 185–196. Springer, 2022. 3

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 7

[6] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 1

[7] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30, 2021. 1

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 1, 2

[9] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pretraining for image captioning. 2022 ieee. In *CVF Conference on computer vision and pattern recognition (CVPR)*, pages 17959–17968, 2021. 9

[10] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021. 3

[11] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. 3

[12] Kaggle. ImageNet100. https://www.kaggle.com/datasets/ambityga/imagenet100. [Accessed Nov. 2024]. 5

[13] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 3

[14] Hamid Kazemi, Atoosa Chegini, Jonas Geiping, Soheil Feizi, and Tom Goldstein. What do we learn from inverting clip models? *arXiv preprint arXiv:2403.02580*, 2024. 3

[15] Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021. 1, 3, 5, 7

[16] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Medclip-samv2: Towards universal text-driven medical image segmentation. *arXiv preprint arXiv:2409.19483*, 2024. 5

[17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 9

[18] Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. Unsupervised vision-and-language pre-training without parallel images and captions. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5339–5350, Online, 2021. Association for Computational Linguistics. 1

[19] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3

[20] Shentong Mo and Pedro Morgado. Unveiling the power of audio-visual early fusion transformers with dense interactions through masked modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27186–27196, 2024. 1

[21] OpenAI. Chatgpt, 2024. Nov 2024 Version. 3

[22] 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, 5

[23] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 9

[24] Sarah Schwettmann, Neil Chowdhury, Samuel Klein, David Bau, and Antonio Torralba. Multimodal neurons in pretrained text-only transformers. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2854–2859, 2023. 1, 7

[25] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can CLIP benefit vision-and-language tasks? In *International Conference on Learning Representations*, 2022. 9

[26] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American journal of roentgenology*, 174(1):71–74, 2000. 3

[27] Radiopaedia Team. Radiopaedia. https://radiopaedia.org/. [Accessed Nov. 2024]. 3

[28] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*, 2022. 1, 9

[29] Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International journal of automation and computing*, 17:151–178, 2020. 2

[30] Moi Hoon Yap, Gerard Pons, Joan Marti, Sergi Ganau, Melcior Sentis, Reyer Zwiggelaar, Adrian K Davison, and Robert Marti. Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE journal of biomedical and health informatics*, 22(4):1218–1226, 2017. 3

[31] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 3

[32] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. 4, 9

[33] Anna Zawacki, Carol Wu, Shih George, Julia Elliott, Mikhail Fomitchev, Hussain Mohannad, Paras Lakhani, Phil Culliton, and Shunxing Bao. Siim-acr pneumothorax segmentation challenge. *Kaggle*, 2019. 3

[34] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021. 9