

Is Visual in-Context Learning for Compositional Medical Tasks within Reach?

Simon Reiß¹ ✉ Zdravko Marinov¹ Alexander Jaus¹ Constantin Seibold²
M. Saquib Sarfraz^{1,3} Erik Rodner⁴ Rainer Stiefelhagen¹

¹Karlsruhe Institute of Technology ²Heidelberg University Hospital ³Mercedes-Benz Tech Innovation ⁴University of Applied Sciences Berlin

✉ simon.reiss@kit.edu

Abstract

In this paper, we explore the potential of visual in-context learning to enable a single model to handle multiple tasks and adapt to new tasks during test time without re-training. Unlike previous approaches, our focus is on training in-context learners to adapt to sequences of tasks, rather than individual tasks. Our goal is to solve complex tasks that involve multiple intermediate steps using a single model, allowing users to define entire vision pipelines flexibly at test time. To achieve this, we first examine the properties and limitations of visual in-context learning architectures, with a particular focus on the role of codebooks. We then introduce a novel method for training in-context learners using a synthetic compositional task generation engine. This engine bootstraps task sequences from arbitrary segmentation datasets, enabling the training of visual in-context learners for compositional tasks. Additionally, we investigate different masking-based training objectives to gather insights into how to train models better for solving complex, compositional tasks. Our exploration not only provides important insights especially for multi-modal medical task sequences but also highlights challenges that need to be addressed.

1. Introduction

In recent years, deep learning-based models trained for a wide range of visual perception tasks have enabled the digitization of applications, that previously required manual intervention. These models are typically developed through a rigorous process involving the collection of large datasets, manual annotations, and supervised training of neural networks. However, the need for extensive image data and labor-intensive annotations remains a significant challenge.

A variety of research areas have addressed these challenges, including transfer learning [30, 31], few-shot learning [40, 42], zero-shot learning [49], semi-supervised learning [34, 35], weakly supervised learning [25, 36], and self-

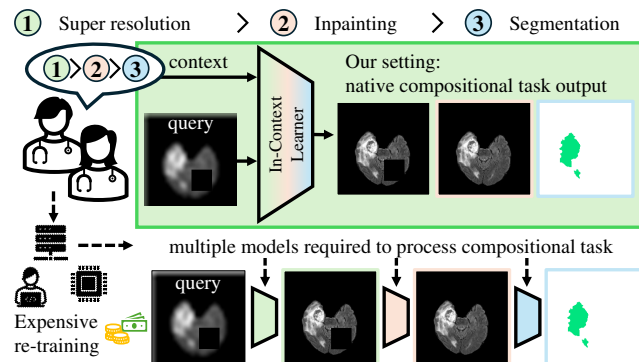


Figure 1. Top: In green, our scenario of visual in-context learning for compositional tasks using one general model. Bottom: Task-specific image processing with multiple specialized models.

supervised learning [1, 5]. Recently, a new approach has emerged: in-context learning [3, 4, 46]. In this paradigm, a model is trained to derive the correct output \mathcal{O} for a query image Q given a task-specific context \mathcal{C} . This approach is akin to few-shot learning but extends to more diverse tasks, combining few-shot and multi-task learning.

In this work, we investigate whether visual in-context learning can be adapted to handle not only individual tasks but also entire task sequences. The ability to instruct a model to process images through a structured sequence of tasks at test time – such as denoising, artifact removal, and target structure segmentation – would enable users, e.g., medical doctors, to define highly individualized image processing pipelines without expensive re-training of a specialized model for each subtask. This flexibility would allow for the dynamic redefinition of image processing pipelines using a single model while upholding transparency, as step-by-step intermediate results can be inspected (Figure 1). Our exploration addresses several fundamental questions:

- **Codebook Limitations (Sec. 4):** We examine the limitations of codebooks in visual in-context learning and propose improvements to better capture diverse task outputs.

- **Dataset Enrichment (Sec. 5.1 & Sec. 5.2):** We present a method to enrich datasets with segmentation labels, creating informative task sequences for in-context training.
- **Training Objectives (Sec. 5.3):** By training codebooks and in-context learners with diverse data and objectives, we gain insight into what works and identify areas needing advancements for compositional medical tasks.

Our findings contribute to the development of more flexible and efficient visual in-context learning models capable of handling complex, multi-step tasks in medical imaging.

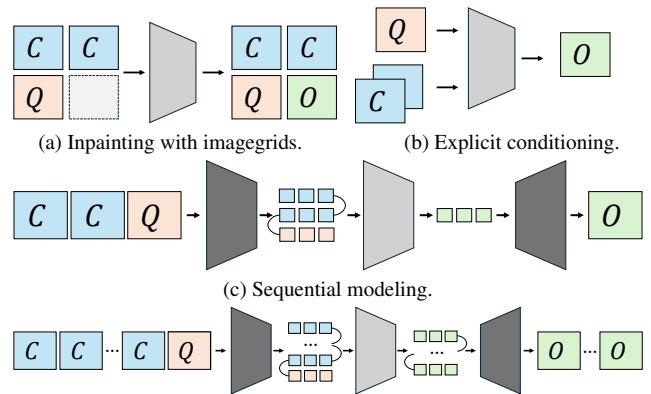
2. Related work

In literature, different choices on how to design visual in-context learners are proposed. One branch of approaches [4, 37, 46, 47] encodes the set of reference images associated with annotations directly as a single input image by forming a grid with them (see Figure 2a). The task to be addressed is described in a contextual image-task pair \mathcal{C} in the first row, while in the last row, the query image \mathcal{Q} for which the task specified must be solved is inserted. The corresponding position of the output to \mathcal{Q} is left blank. The visual in-context learning model is then trained to inpaint the missing output image \mathcal{O} , thereby enabling prompting with new contexts \mathcal{C} to solve new tasks for new queries.

In the medical domain, a different design of the in-context learning process has been proposed [9, 12]. There, networks encode each image-task pair in the context set together with the representation of the query image pairwise and aggregate the information between all image-task pair-query combinations with a pooling operation, enabling arbitrarily many context pairs to be used (see Figure 2b).

The third design choice for visual in-context learning frames the setting as a masked- or next token prediction task on input sequences [3, 21, 26] as common in natural language processing [15, 32]. To obtain sequences of discrete tokens, the context image-task pairs and query image are arranged into a sequence and processed to discrete tokens via a codebook, more explicitly, the quantization layer [44] of a pre-trained VQ-GAN [10, 11, 17]. The masked token prediction is done in the latent space, spanned by the codebook (see Figure 2c). To get the in-context prediction for a given query image \mathcal{Q} in inference, the learner is given the contextual image-task pairs \mathcal{C} as token sequence, then the tokens of the query image \mathcal{Q} are appended, and through the training objective of reconstructing masked portions of the sequence, the model predicts tokens corresponding to the output, conditioned on $\mathcal{C}\mathcal{Q}$. Finally, the output tokens can be decoded back into the image space with the codebook.

Further related studies include the design of visual prompting strategies [41, 53], making image tokenization more concise [51], addressing in-context segmentation [9, 33, 47], in-context learning on video [52] and visual in-context learning with diffusion models [19, 50].



(d) Composite tasks as sequences: Context encodes sequential tasks, the generated output is of sequential nature as well.

Figure 2. Visual in-context learning architectures and paradigms.

In contrast to previous approaches, we explore visual in-context learning on compositional tasks, *i.e.*, our context and our outputs encompass sequences of tasks. To do this, we build on top of the sequential modeling paradigm and extend it to such compositional tasks (see Figure 2d).

3. Preliminaries

3.1. Compositional in-context learning

We define the task of in-context learning as training a model $\theta(\cdot)$ which, based on a context \mathcal{C} and a query input \mathcal{Q} predicts the contextual output \mathcal{O} for the query. In visual in-context learning, all constituents in \mathcal{C} , \mathcal{Q} , \mathcal{O} are images of shape $\mathbb{R}^{3 \times W \times H}$. The context \mathcal{C} contains a task specification, *i.e.*, a small set of images and corresponding task-specific annotations $\mathcal{C} = \{c_0, \dots, c_n\}$, that exemplify input-output relations, where n is the number of example input-output relations that specify the task to be solved. As such, the model needs to capture these relations and apply the task-specific transformation from the query to the output: $\theta(\mathcal{C}, \mathcal{Q}) = \mathcal{O}$. The model is trained to generalize to novel contexts \mathcal{C} , which encode new tasks at test time.

We investigate a specific type of visual in-context learning, where the constituents c_i of the context set are not pairs of image and annotation, but a long sequence of interdependent tasks $c_i \in \mathbb{R}^{t \times 3 \times W \times H}$, where t indicates the number of tasks in the composite task. This adaptation also entails changes to \mathcal{O} , which previously was a singular output image, but now becomes a sequence $\mathcal{O} \in \mathbb{R}^{(t-1) \times 3 \times W \times H}$. As such, models need to produce an image sequence of $(t-1)$ intermediate task outputs for a query image \mathcal{Q} .

3.2. Visual codebooks

To cope with the high dimensional nature of visual in-context learning on compositional tasks, with input $\mathcal{C}\mathcal{Q} \in \mathbb{R}^{(n \cdot t + 1) \times 3 \times W \times H}$ and output $\mathcal{O} \in \mathbb{R}^{(t-1) \times 3 \times W \times H}$, a

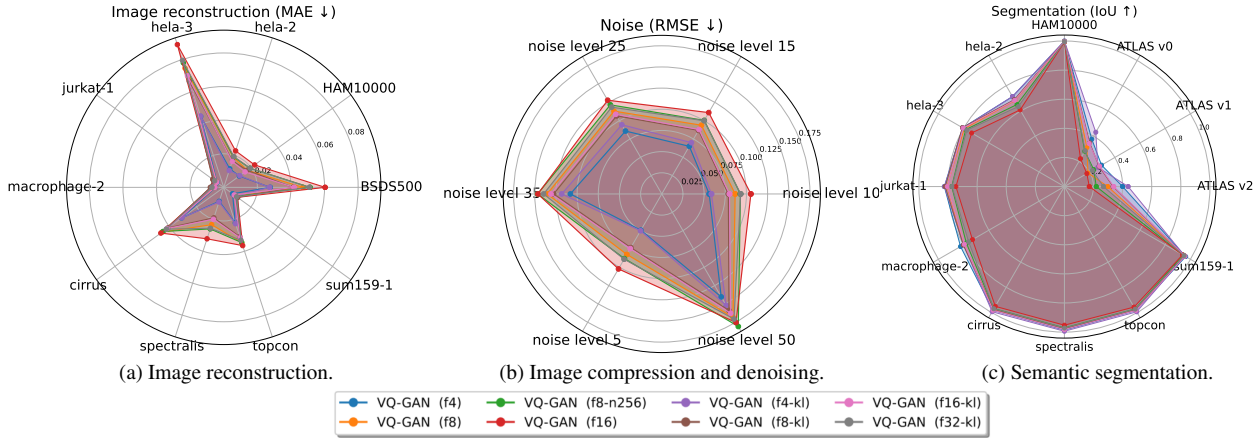


Figure 3. Investigation on how well pre-trained codebooks represent task data. Left to right: Evaluation of image reconstruction, evaluation of reconstruction of different noise levels in BSDS500 and reconstruction efficacy on segmentation maps of different datasets.

codebook $\phi(\cdot)$ is an essential tool to reduce the dimensionality. Often parameterized with an auto-encoding neural network with quantization layer [17, 44], a codebook transforms an image x into a sequence of q discrete tokens $\phi(x) \in [0, \dots, N - 1]^q$, where N is the vocabulary size, *i.e.*, the number of discrete tokens in the codebook. With such a codebook processing of $\mathcal{CQ} \in \mathbb{N}^{(n \cdot t + 1) \times q}$ to $\mathcal{O} \in \mathbb{N}^{(t-1) \times q}$ is computationally palpable, as $\phi(\cdot)$ is built to produce $q \ll 3 \times W \times H$. The reverse direction from discrete token space back into the image space is done with a decoder $\phi^{-1}(\cdot)$, both $\phi(\cdot)$ and $\phi^{-1}(\cdot)$ are trained jointly.

The number of tokens per image q in common codebooks, *e.g.*, VQ-GAN [17] codebooks, is influenced by the image size and the network architecture, as spatially smaller latent representations lead to fewer tokens.

4. Analysis of task recovery in codebooks

Codebooks are an enabling factor for visual in-context learning on compositional tasks, but they also constrain what can be learned when transferring \mathcal{CQ} to \mathcal{O} , as this transfer now happens in token space. To investigate how much codebooks limit in-context learning, we measure how well they can recover task-specific outputs in \mathcal{O} .

4.1. Codebooks for task-specific reconstruction

Therefore, we run a small set of preliminary experiments, where we simply take pre-trained codebooks (model zoo of [17, 38]), map different task outputs into the token space and directly reconstruct them back into the image space to evaluate them with respective task metrics. In Figure 3, we consider the tasks image reconstruction, image denoising, and semantic segmentation and determine the upper bounds for any in-context model working with these codebooks.

Image reconstruction: The codebooks were all trained for the task of image synthesis, a very similar task to this is

pixel-wise image reconstruction. When we look at the capabilities of accurately reconstructing images in Figure 3a, which can be measured in mean absolute error (MAE), we see, that the best models consistently are under 0.05 MAE, an acceptable bound for this metric. This outcome is not too surprising, as the objective function of all the trained codebooks involved a regression loss term, to lead to small color-value deviations [17]. What’s more interesting is, that the errors of the best models stay low for non-natural image domains, on which the models were not specifically trained, such as optical coherence tomography (cirrus, spectralls, topcon [7]), electron microscopy (helia, jurkat, macrophage [23]) or skin-lesion images (HAM10000 [43]).

Image compression and denoising: When denoising is the task to be addressed a question that arises is, whether codebooks can even represent image noise, *i.e.*, whether the tokens allow for noise reconstruction. When increasing noise-intensity in natural images (BSDS500 [29]), codebooks can not recover this induced noise. In Figure 3b, more noise leads to a higher Root-mean-square error (RMSE), which is unsurprising, as neither the training data nor the loss functions of the codebooks are tuned for recovering noise and thereby, as a side-product of the compression via auto-encoding, noise is ignored during reconstruction. For the task of image denoising, the Peak signal-to-noise ratio (PSNR) is a commonly reported metric, which is bound to 27.46 on BSDS500 when using the best pre-trained codebook to reconstruct the clean data. As such denoising can not exceed this threshold for this data.

Semantic segmentation: To evaluate the upper bound for image segmentation, we tokenize segmentation maps of different datasets and reconstruct them. To compute the Intersection over Union (IoU), we map the reconstructed color values to the closest valid color in the input segmentation map. We notice, that for datasets where binary seg-

mentation is done, such as for skin lesion segmentation (HAM10000 [43]), the upper IoU bound is almost pixel-perfect, which might be due to the simple topology of this data, *i.e.*, blob-like structures. When moving from binary to multi-class segmentation, the upper bound successively degrades to 98% for three classes (cirrus, spectralis, top-con [7]) and when increasing to 36 classes (hela-2 [23]) the upper bound is 71% IoU. This likely has two causes, first, more classes lead to finer segmentation structures which are harder to reconstruct, and second, more classes lead to more colors in the segmentation maps which in turn leads to a fuzzier assignment to the valid colors. We exemplify this to the extreme, with experiments on the ATLAS dataset [24] with over 142 classes. There, the codebooks are not capable to precisely reconstruct colors such that they are mapped to the correct classes, leading to an upper IoU of merely 29%. Yet, we notice that if codebooks are prompted with segmentation maps utilizing different color schemes ($v_0 - v_2$) for these 142 classes, the upper bound changes drastically to an IoU of 44%. As such, codebook-based token spaces for in-context learning are sensitive to visual prompting [41].

Summary on recovering task information Building an in-context learner on top of a codebook that was trained for image synthesis propagates its capabilities, but also its limitations. Concretely, when the task includes image reconstruction or recovering simple binary structures, near perfect performance may be achieved by the learner using the codebook. Yet, when the tasks get highly discriminative, *i.e.*, discrimination between a high number of semantic concepts, or when fine-grained structures need to be identified, codebooks pre-trained on natural images produce errors.

4.2. Task-informed codebooks

To address this gap in representing task-related outputs, such as segmentation maps, we propose a simple adaptation to codebook training: train the codebook on imaging data *and* on task outputs. While this data-centric solution for codebooks to better represent task data sounds simple, in practice we experiment with two important aspects.

Data balancing: To capture multiple modalities, *i.e.*, image data and different task output types (*e.g.*, segmentation maps, bounding boxes, etc.), we explore how to best balance the training data. One aspect is the balance between different task output types, the second aspect is to balance sampling from different datasets when multiple are in use.

Color remapping augmentation: We noticed in our preliminary investigation on codebooks, that different color schemes can have a major effect on the upper performance bound under a given codebook. To overcome this, we utilize the idea of augmenting the semantic task outputs. During training of the codebook, we resample the colors in segmentation maps, bounding boxes, etc. to random different colors. This ensures, that the codebook is not only capa-

ble to represent a specific color map, but can generalize to prompts with new class-color definitions.

With a codebook that tokenizes images and task outputs, we move to in-context learning on compositional tasks next.

5. Compositional visual in-context learning

First, we obtain training data for compositional tasks via synthetic data generation, then we outline pre-training objectives for visual in-context learning. In Figure 4, we show our compositional visual in-context learning pipeline.

5.1. Synthetic task enrichment

To advance towards the setting of compositional in-context learning, the basic requirement is access to compositional task sequences $\mathcal{C}\mathcal{Q}\mathcal{O}$ to train with. As such, we start with describing a data-centric pipeline to obtain multiple vision tasks from pre-existing semantic segmentation datasets. In order to generalize to new task sequences at test time, the training task sequences need to be highly diverse and cover different tasks. We first enrich the dataset by either utilizing the images of the dataset to setup additional generative- and transformation tasks or the segmentation annotation to derive annotation variants for additional discriminative tasks.

Generative Tasks: The first type of tasks we formulate are generative tasks. Here the task is to restore the original image from a degraded version. For example, we lower the resolution of an image which serves as input and take the original image as prediction target to form the task of *super resolution*. Similarly, we formulate an *inpainting* task, and *denoising* task by either cutting out portions of the original image or adding Gaussian noise. Lastly, for other types of degradations or photometric adaptations to form input-output task pairs we add *color jitter* to remove, *invert* the intensity values of the original image or alter its *brightness*.

Geometric transformation tasks: In addition to altering the image content with artifacts and degradation and successively reconstructing the obstructed information, we formulate transformation tasks. These tasks are based on the geometric transformations: *horizontal flip*, *vertical flip* and *rotation* by $\{90^\circ, 180^\circ, 270^\circ\}$. Here, the input is the original image and the output is the geometrically transformed image, resulting in tasks that require global information exchange when formulated as image-to-image tasks.

Discriminative tasks: The last group of tasks we derive from the original dataset are based on segmentation annotations. We extract the edges of all semantic classes in the annotations to get *semantic edge* annotations. Akin to object detection, we obtain detection targets by drawing *semantic boxes* around individual segments in the images. Lastly, we simulate two exotic annotation types for more diversity, so called *skeletons* (medial axes) of segments which are thin lines describing their topology as well as central *points* on

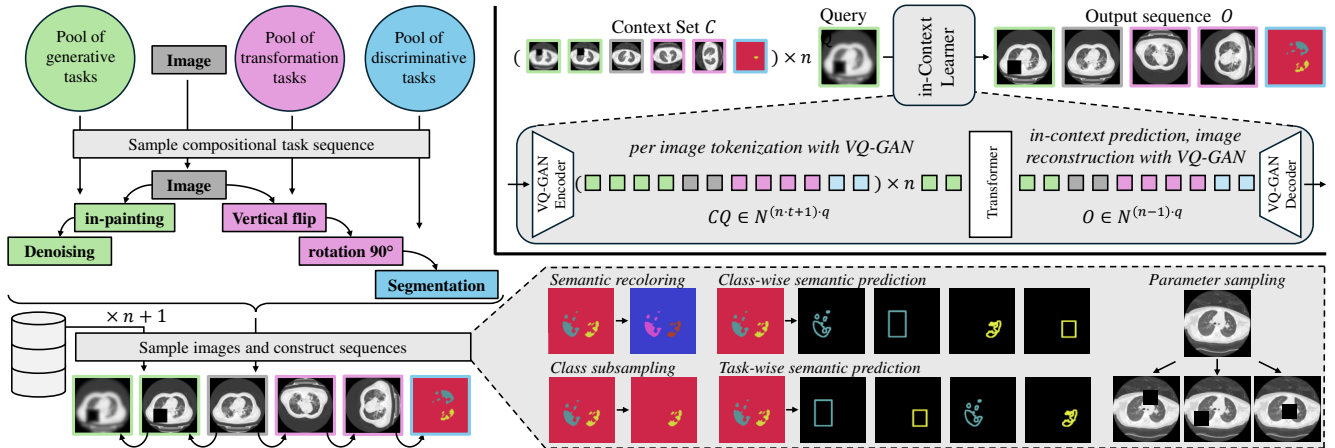


Figure 4. On the left, we present the sequence structure, which we enforce to construct compositional tasks, with some guardrails shown on the bottom right. The top right shows the general processing pipeline of compositional tasks through a visual in-context learner.

segments. Skeletons and points are related to scribbles [28] from interactive- or weakly-supervised learning [36].

The extracted generative-, transformation- and discriminative task information will form the basis of our compositional task sequences. As in-context learning has been mainly done with large heaps of data [3, 8, 21], this enrichment is done for 38 medical segmentation datasets [27].

5.2. Compositional task sampling

With the enriched dataset comprised of diverse vision tasks, we are equipped to build task sequences. Crucially, when sampling task sequences from the dataset, all $n + 1$ task sequences in \mathcal{CQO} , *i.e.*, all c_i and the concatenation \mathcal{QO} need to obey the same task-sequence logic. While randomly drawing tasks, putting them into a sequence and based on it drawing samples from the dataset successively leads to the highest diversity in task sequences, this approach fails in reality. This is due to task sequences becoming extremely poor in information. For discriminative tasks, this is due to the fact, that randomly drawing images often leads to sequences that do not share the same classes, and as such \mathcal{O} may contain classes that are not present in any c_i , and thereby the task to predict them from \mathcal{C} is ill-posed. Similarly, for generative tasks, sequencing is key, as randomly sampling an image into the sequence and thereafter its low resolution version from the *super resolution* task amounts to a rather simple task sequence encoding down scaling.

5.2.1. Controlling task sequence generation

With this insight, we formulate guardrails within which task sequences can be sampled randomly, to balance the information content of the sequences with their diversity.

We enforce a task sequence structure that follows: $\{\textit{generative}\}$ *image* $\{\textit{transformation}\}$ $\{\textit{discriminative}\}$. For each task family in brackets, multiple task in-

stances may be sampled, *e.g.*, an explicit task sequence of the structure *inpaint* \rightarrow *denoise* \rightarrow *image* \rightarrow *rotate 90* \rightarrow *segmentation* would be valid. With this structure, task sequences can be computed in a straightforward fashion. First, generative tasks, which obfuscate the original image x are applied from right to left $\varphi_{\text{inpaint}}(\varphi_{\text{denoise}}(x), \varphi_{\text{denoise}}(x), x)$, then geometric transformations are applied and added to the right of the sequence: $\varphi_{\text{inpaint}}(\varphi_{\text{denoise}}(x), \varphi_{\text{denoise}}(x), x, \varphi_{\text{rotate } 90^\circ}(x))$ and finally the discriminative tasks are added, *i.e.*, the semantic annotation x^{seg} , which has to be transformed with all geometric transformations first: $\varphi_{\text{inpaint}}(\varphi_{\text{denoise}}(x), \varphi_{\text{denoise}}(x), x, \varphi_{\text{rotate } 90^\circ}(x), \varphi_{\text{rotate } 90^\circ}(x^{\text{seg}}))$. We directly see, that some tasks are dependent on others, *e.g.*, the segmentation task operates on the rotated version of x , or, inpainting operates on the noised image. This highlights the interdependence and compositional nature of our synthetic task sequences.

On top of this task structure, we add a few more rules and guardrails for information-rich task sequences. For discriminative tasks, we increase the diversity by sub-sampling the available classes, such that not all classes will be present in all task sequences, but consecutive task sequence sampling may contain different class-subsets. This adds an incentive to adhere to the classes in the context \mathcal{C} and only produce outputs \mathcal{O} that contain the same class subset. This guardrail is designed to counteract a degradation towards always predicting all classes associated to a given image.

To further increase diversity, we either present the discriminative task output for all classes in one image or adapt the discriminative task to a class-wise task sequence, where the task output is predicted for each class individually. In case a task sequence contains multiple discriminative tasks, the ordering can either be on a task basis, *i.e.*, different tasks in sequence, or on a class basis, *i.e.*, the task outputs are grouped by class, meaning all tasks are solved for one class

before the next class is handled.

To ensure flexibility, we recolor the semantic annotations consistently in \mathcal{CQO} and assign a new color for each class randomly to anticipate new class-colors that may occur in testing. Further, when sampling the context set \mathcal{C} we make sure that it contains all classes present in \mathcal{O} .

For more diversity in the different tasks, we sample their parameters, such as mean and variance for Gaussian noise, the cut-out positions for inpainting and the downscale factor for super resolution, line width of semantic edges, boxes or skeletons, and rotation angles. All generation guardrails are visualized on the bottom right of Figure 4.

5.2.2. Controlling task sequence length

A consideration when computing task sequences is the sequence length, as too long sequences make learning correlations between sequence elements difficult and computationally expensive. For adequate costs, we bound the total length $|\mathcal{CQO}|$, in our case to 30 images, which limits the number of tasks in a sequence to 15 (as the minimal compositional task is $|c_0\mathcal{QO}| = 30$, and thus $|c_0|, |\mathcal{QO}| = 15$).

Note, longer sequences are possible, yet, a sequence of 30 images equates to a sequence length of $30 \cdot q$ in token space. For common codebooks q ranges between 144 and 4,096 tokens per image, such long sequences require extensive compute infrastructure and may require model architectures tailored for excessive sequence lengths [6, 18].

5.3. Training compositional in-context learners

In our work, we mainly investigate *how* to enable neural networks to learn compositional vision tasks in context. The question which network architecture might be best suited for this setting is beyond this investigation. Thus, we resort to the tried and tested Transformer architecture [17, 32, 45], which we utilize as in-context learner.

The more fundamental question that needs to be answered is with which optimization objective to train the transformer. For this, we orient ourselves at masking- and subsequent reconstruction-based training strategies which have been successfully used in language modeling [8, 16, 39], image pre-training [1, 2, 22] and visual in-context learning [3, 21]. As such, we transfer the compositional task sequences \mathcal{CQO} into token space to form a long token sequence. Then we randomly switch out tokens in this sequence with random tokens, whereas the transformer is tasked with identifying those malicious tokens and predicting the correct ones. Specifically, the manipulated token sequence serves as input to the model which predicts a softmax output over all codebook entries for each sequence element. At each location of a malicious (*i.e.*, masked) token the cross-entropy loss between the prediction and the ground-truth sequence \mathcal{CQO} is back propagated.

The concrete masking strategy to best train the transformer for recovering the contextual output \mathcal{O} at test time

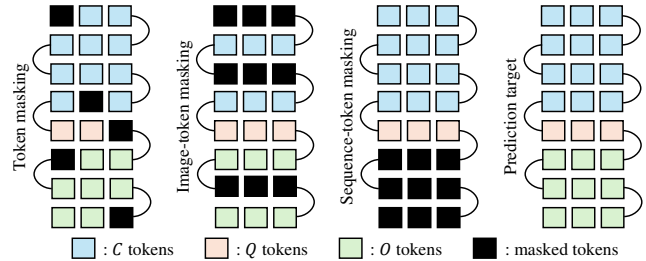


Figure 5. Token masking variants for training visual in-context learners, each row of tokens symbolizes the tokens of one image.

is not clear, three strategies are shown in Figure 5. Random *token masking* is one possible strategy, where in the sequence a portion p of tokens are switched out randomly as done in language modeling and vision transformer pre-training [8, 22]. Adapting the idea of masking out entire words in language modeling to our setting aligns better with switching out all tokens that encode individual images, which is what we refer to as *image-token masking*. Lastly, as we are able to produce the composite tasks of context set \mathcal{C} , query \mathcal{Q} and output \mathcal{O} , we can, in training, mirror the task of compositional in-context prediction by simply switching out all tokens in \mathcal{O} , which we term *sequence-token masking*.

Special tokens To inform the transformer about the end of tokens that correspond to one image, we add k special tokens $\langle i\text{-END} \rangle$ after the tokens of each image into the sequence. Similarly, we add k $\langle c\text{-END} \rangle$ tokens, when a context element c_i is complete. We add multiple such tokens, as transformers have been shown to benefit from additional special tokens to store information [13].

6. Experiments

In this section, we outline the training and evaluation of visual in-context learners for compositional medical tasks.

6.1. Datasets & evaluation setting

We utilize 38 datasets from the MedSAM dataset collection [27], which is comprised of diverse medical datasets, covering different imaging modalities such as computed tomography, optical coherence tomography, x-ray, ultrasound and images of skin lesions. We extract images and segmentation maps, encode them as RGB images, rescale them to 200×200 and enrich them as outlined in Section 5.1.

Training of codebooks is done on individual images and task outputs from our enriched MedSAM data. The codebook in turn is used to tokenize compositional tasks \mathcal{CQO} on an image-wise basis. Our codebook processes 200×200 images into $q = 144$ tokens, with at most 30 images in task sequences the max length is 4,320 (without special tokens).

The visual in-context learners are trained on compositional tasks as generated with our sampling process in Sec-

Method	Masking			Generative						Transformation			Discriminative					
	p	#im	\odot	Super res. PSNR \uparrow	Inpaint PSNR \uparrow	Denoise PSNR \uparrow	Jitter PSNR \uparrow	Invert MSE \downarrow	Equal. MSE \downarrow	Bright MSE \downarrow	Flip (h) MSE \downarrow	Flip (v) MSE \downarrow	Rotate MSE \downarrow	Seg. IoU \uparrow	Boxes F-1 \uparrow	Points F-1 \uparrow	Skeletons F-1 \uparrow	Edges F-1 \uparrow
Copy baseline	-	-	-	11.477	12.347	11.814	11.585	0.0845	0.0728	0.0864	0.0789	0.0772	0.0802	50.79%	63.49%	51.52%	47.04%	48.59%
Token masking	15%	-	-	9.398	9.948	9.692	9.651	0.1341	0.1310	0.1296	0.1456	0.1484	0.1566	26.49%	38.17%	36.54%	36.49%	35.32%
Image-token masking	-	5	-	16.063	15.288	16.851	15.726	0.0387	0.0468	0.0425	0.0762	0.0839	0.0943	44.83%	61.01%	50.36%	46.17%	47.42%
Sequence-token masking	-	-	\checkmark	18.265	19.116	20.131	19.206	0.0173	0.0167	0.0174	0.0203	0.0200	0.0219	52.60%	62.26%	52.79%	47.38%	48.16%
\perp (END) $k = 2$	-	-	\checkmark	18.411	19.282	20.169	19.280	0.0176	0.0162	0.0171	0.0196	0.0193	0.0207	52.82%	62.32%	52.68%	47.17%	48.27%
\perp (END) $k = 4$	-	-	\checkmark	18.290	19.226	20.130	19.182	0.0181	0.0163	0.0175	0.0195	0.0194	0.0206	52.78%	62.36%	52.84%	47.16%	48.30%
Mixed masking	15%	3	\checkmark	17.870	18.431	19.393	18.554	0.0211	0.0200	0.0197	0.0236	0.0230	0.0252	50.39%	62.06%	52.68%	46.77%	47.82%
Longer training ($k = 2$)	-	-	\checkmark	18.513	19.652	20.531	19.711	0.0165	0.0147	0.0158	0.0165	0.0166	0.0169	61.21%	65.92%	55.66%	49.55%	50.85%
Upper Codebook bound	-	-	-	20.354	21.975	22.527	21.800	0.0107	0.0077	0.0109	0.0097	0.0096	0.0098	86.72%	85.33%	74.69%	60.21%	66.32%

Table 1. Evaluation of Visual in-Context Learning on compositional medical tasks under different masking pre-training strategies.

Method	Segment. IoU \uparrow	Edge detection F-1 Score \uparrow	Skeletons F-1 Score \uparrow	Bound. box F-1 Score \uparrow	Reconst. MAE \downarrow
<i>Fixed color semantics</i>					
VQ-GAN	59.89%	59.53%	55.74%	60.65%	0.09121
\perp task training	86.19%	95.54%	96.24%	99.45%	0.03756
\perp task balance	88.07%	95.54%	96.08%	99.74%	0.03745
\perp recoloring	84.07%	81.23%	84.27%	84.97%	0.03681
\perp task + data b.	86.76%	95.18%	96.06%	99.59%	0.03706
\perp recoloring	83.59%	85.04%	87.01%	91.58%	0.03742
<i>Random color semantics</i>					
VQ-GAN	59.12%	61.14%	56.62%	62.51%	0.09121
\perp task training	66.29%	72.78%	70.06%	73.28%	0.03756
\perp task balance	69.21%	74.65%	71.71%	74.58%	0.03745
\perp recoloring	83.00%	91.45%	91.19%	95.86%	0.03681
\perp task + data b.	71.15%	75.65%	72.73%	75.97%	0.03706
\perp recoloring	83.19%	91.93%	92.15%	97.46%	0.03742

Table 2. Upper performance bounds set by trained codebooks when encoding and reconstructing images and task outputs.

tion 5.2 and tokenization with the trained codebook.

We evaluate the efficacy of the codebook to reconstruct task-specific outputs and the performance of the visual in-context learner to predict task sequences with the metrics Intersection over Union (IoU), F-1 Score, Mean Average Error (MAE), Root-Mean-Square Error (RMSE) and Peak Signal-to-Noise Ratio (PSNR), depending on the task.

6.2. Implementation details

The visual codebooks we train are VQ-GANs [17] with 70M parameters that are pre-trained on ImageNet [14]. The visual in-context learners that we train are GPT2 transformers [32] as implemented in [17] with 195M parameters.

Codebook training: Our codebook is the VQ-GAN $f = 16, KL(d = 16)$ variant¹, the amount of tokens to be learned is $N = 16,384$. We fine-tune it on the enriched MedSAM data, while we omit the discriminator loss of the VQ-GAN setup, as we noticed diverging results when utilizing it. Training is done for one epoch with a batch size of 96 and a total of 5M images and task outputs combined.

In-context learner training: The GPT2 transformer is optimized 8 epochs on 222K synthetically generated compo-

sitional tasks. Due to memory constraints, we train with a batch size of four, a context window of 4, 500 and operate on 16,386 tokens, *i.e.*, codebook size plus two special tokens, our default number of inserted special tokens is $k = 1$. All models are trained on 4xNVIDIA A100 with 40GB.

6.3. Quantitative codebook evaluation

In Table 2, we show how well different codebooks can reconstruct MedSAM images and importantly the *discriminative task* outputs. First, we evaluate a pre-trained VQ-GAN codebook in row one, where we see that it struggles severely to represent *discriminative task* related information in MedSAM with scores between 55.74 – 60.65%, merely the image reconstruction error is low. When fine-tuning it on MedSAM task data, we see that for *discriminative tasks* the results improve drastically to scores ranging between 86.19 – 99.45% (row two). Balancing the training data to contain each task output and image to a similar extent leads to slightly higher upper codebook bounds for segmentation and boxes. Further, when balancing the data with respect to both tasks and oversampling small datasets in MedSAM for all datasets to be represented equally, we observe similar results (row five). In the first six rows, we evaluate on task outputs with fixed color schemes, we can see, that in this evaluation scenario adding the random color augmentation (rows four and six) from Section 4.2 lowers performance.

When switching to a more realistic setup, where *discriminative tasks* come with arbitrary colors during testing (rows seven to twelve), we see the ranking shift. The results of all codebooks trained without color remapping augmentation diminish severely on new class-colors. The clear winner is the VQ-GAN with task- and dataset balancing and color remapping augmentation (last row). For our visual in-context learning experiments, we utilize this codebook.

6.4. Quantitative results on compositional tasks

In Section 5.3, we define three different masking strategies for pre-training the visual in-context learner. We compare them in Table 1, where we train the in-context learner for 8 epochs either with switching out $p = 15%$ of the to-

¹<https://github.com/CompVis/latent-diffusion/>

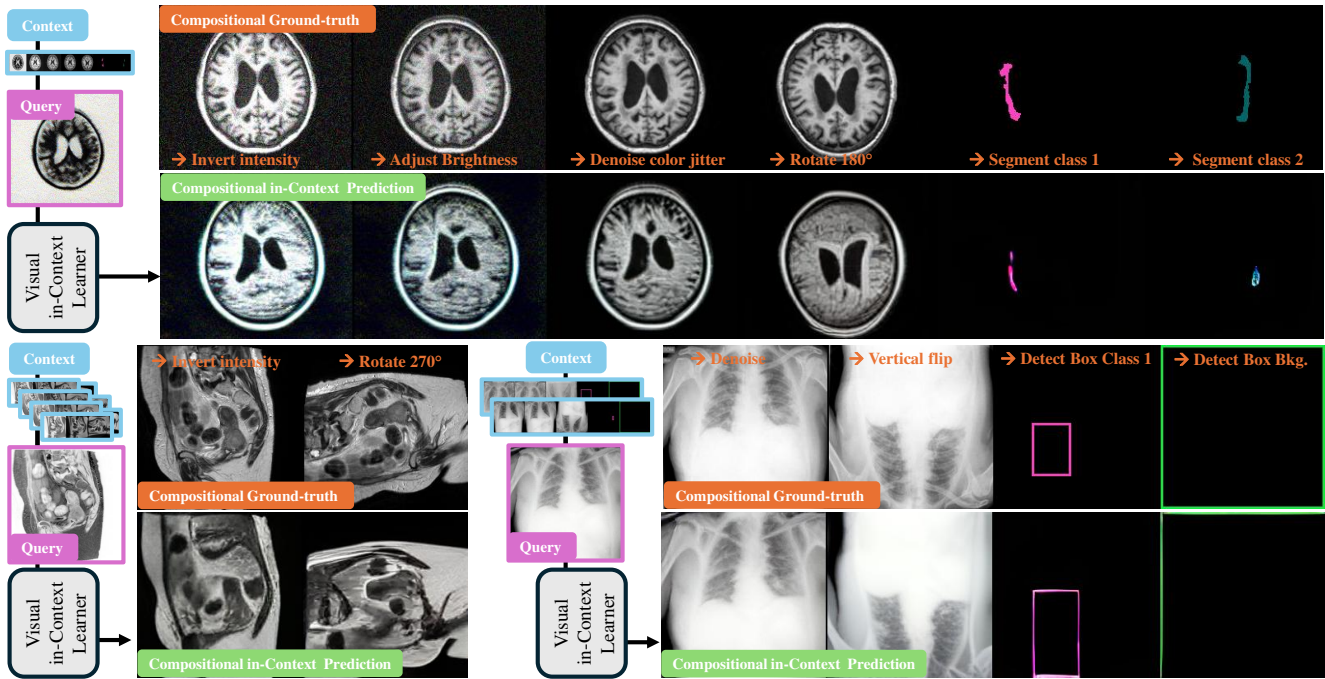


Figure 6. We show three different qualitative compositional medical task predictions by the best visual in-context learning variant.

kens randomly, or around 15% of image-tokens (*i.e.*, tokens of 5 full images), or all tokens in the output-sequence \mathcal{O} are switched out. The results are computed on 2,000 test compositional tasks formed from image- and task outputs excluded from the training set. The lower copy baseline, which simply predicts samples from the context set, and the upper codebook bound delimit the results.

The pre-training strategy with simple token masking (row two), yields the worst results when evaluating all intermediate task outputs in the predictions \mathcal{O} for unseen compositional tasks. Identifying and switching randomly switched tokens to the correct ones may be solved by attending to the local context of a token and does not require learning long range dependencies between contexts c_i and \mathcal{O} . When dropping out all tokens of images, we notice a substantial improvement in metrics (row three), the model starts to draw correlations between context- and output sequences, filling in coherent image-tokens. Lastly, we mirror the test scenario of switching out all tokens of \mathcal{O} and predicting the correct ones. This strategy leads to a substantial improvement in all metrics (row four), the network now captures the interdependence between task sequences in the context and patterns to be decoded for predicting the sequence \mathcal{O} . Mixing all strategies above does not improve results (row seven), increasing the number of $\langle \text{END} \rangle$ tokens to $k = 2$ (row five) and $k = 4$ (row six) yields slight improvements, with $k = 2$ coming out on top. By training the variant with $k = 2$ and *sequence-token masking* for 52 epochs (row eight) the results improve further. We see *generative-* and

transformation tasks can be better captured by the visual in-context learners, while they struggle with connecting image patterns with semantics in *discriminative tasks*, especially fine structures such as boxes, points, edges or skeletons.

6.5. Qualitative results on compositional tasks

In Figure 6, we display qualitative compositional task predictions. The first sequence prediction (top) indicates that the visual in-context learner can capture the structure of the tasks to be solved coherently, *i.e.*, follow the instructions from the context set. The bottom two outputs operate on different imaging modalities, all of which are captured well, highlighting the multi-modal capabilities of the model.

7. Conclusion

This paper explored visual in-context learning on compositional medical tasks. We identified prerequisites and opened a pathway to train transformer-based in-context learners on synthetic task sequences. The trained models are capable of solving complex vision problems step by step, thereby making their response interpretable through verifiable intermediate results – important in medical imaging. Such step-by-step processing may in the future enable visual thinking processes as in language modeling [20, 48]. Yet, as we only took a first step, challenges remain: codebooks need to represent fine structures better, training objectives need to be improved to better capture image-semantic relations, and capabilities on new task distributions need to be explored.

Acknowledgments This work was supported by funding from the pilot program Core-Informatics of the Helmholtz Association (HGF) and by the joint research school “HIDSS4Health – Helmholtz Information and Data Science School for Health. The authors gratefully acknowledge the computing time provided on the high-performance computer HoreKa by the National High-Performance Computing Center at KIT (NHR@KIT). This center is jointly supported by the Federal Ministry of Education and Research and the Ministry of Science, Research and the Arts of Baden-Württemberg, as part of the National High-Performance Computing (NHR) joint funding program (<https://www.nhr-verein.de/en/our-partners>). HoreKa is partly funded by the German Research Foundation (DFG). This work was performed with the help of the Large Scale Data Facility at the Karlsruhe Institute of Technology funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research. This research was also funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) via the Project Ap-IFM (528483508).

References

- [1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European Conference on Computer Vision*, pages 456–473. Springer, 2022. 1, 6
- [2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. *arXiv preprint arXiv:2301.08243*, 2023. 6
- [3] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. *arXiv preprint arXiv:2312.00785*, 2023. 1, 2, 5, 6
- [4] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017, 2022. 1, 2
- [5] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khruikov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations*, 2021. 1
- [6] Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. *arXiv preprint arXiv:2405.04517*, 2024. 6
- [7] Hrvoje Bogunović, Freerk Venhuizen, Sophie Klimscha, Stefanos Apostolopoulos, Alireza Bab-Hadiashar, Ulas Bagci, Mirza Faisal Beg, Loza Bekalo, Qiang Chen, Carlos Ciller, et al. Retouch: The retinal oct fluid detection and segmentation benchmark and challenge. *IEEE transactions on medical imaging*, 38(8):1858–1874, 2019. 3, 4
- [8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. 5, 6
- [9] Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Universeg: Universal medical image segmentation. *arXiv preprint arXiv:2304.06131*, 2023. 2
- [10] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [11] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2
- [12] Steffen Czolbe and Adrian V Dalca. Neuralizer: General neuroimage analysis without re-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6217–6230, 2023. 2
- [13] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 6
- [14] 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. 7
- [15] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2
- [16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. 6
- [17] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2, 3, 6, 7
- [18] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 6
- [19] Zheng Gu, Shiyuan Yang, Jing Liao, Jing Huo, and Yang Gao. Analogist: Out-of-the-box visual in-context learning with image diffusion model. *ACM Transactions on Graphics (TOG)*, 43(4):1–15, 2024. 2
- [20] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi

- Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 8
- [21] Jianyuan Guo, Zhiwei Hao, Chengcheng Wang, Yehui Tang, Han Wu, Han Hu, Kai Han, and Chang Xu. Data-efficient large vision models through sequential autoregression. *arXiv preprint arXiv:2402.04841*, 2024. 2, 5, 6
- [22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 6
- [23] Larissa Heinrich, Davis Bennett, David Ackerman, Woohyun Park, John Bogovic, Nils Eckstein, Alyson Petrucio, Jody Clements, Song Pang, C Shan Xu, et al. Whole-cell organelle segmentation in volume electron microscopy. *Nature*, 599(7883):141–146, 2021. 3, 4
- [24] Alexander Jaus, Constantin Seibold, Kelsey Hermann, Negar Shahamiri, Alexandra Walter, Kristina Giske, Johannes Haubold, Jens Kleesiek, and Rainer Stiefelhagen. Towards unifying anatomy segmentation: Automated generation of a full-body ct dataset. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 41–47. IEEE, 2024. 4
- [25] Zonglin Li, Ruiqi Guo, and Sanjiv Kumar. Decoupled context processing for context augmented language modeling. *Advances in Neural Information Processing Systems*, 35: 21698–21710, 2022. 1
- [26] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 2
- [27] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 5, 6
- [28] Zdravko Marinov, Paul F Jäger, Jan Egger, Jens Kleesiek, and Rainer Stiefelhagen. Deep interactive segmentation of medical images: A systematic review and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 2024. 5
- [29] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, pages 416–423, 2001. 3
- [30] Thomas Mensink, Jasper Uijlings, Alina Kuznetsova, Michael Gygli, and Vittorio Ferrari. Factors of influence for transfer learning across diverse appearance domains and task types. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9298–9314, 2021. 1
- [31] Basil Mustafa, Aaron Loh, Jan Freyberg, Patricia MacWilliams, Megan Wilson, Scott Mayer McKinney, Marcin Sieniek, Jim Winkens, Yuan Liu, Peggy Bui, et al. Supervised transfer learning at scale for medical imaging. *arXiv preprint arXiv:2101.05913*, 2021. 1
- [32] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2, 6, 7
- [33] Marianne Rakic, Hallee E Wong, Jose Javier Gonzalez Ortiz, Beth A Cimini, John V Guttag, and Adrian V Dalca. Tyche: Stochastic in-context learning for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11159–11173, 2024. 2
- [34] Simon Reiß, Constantin Seibold, Alexander Freytag, Erik Rodner, and Rainer Stiefelhagen. Every annotation counts: Multi-label deep supervision for medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9532–9542, 2021. 1
- [35] Simon Reiß, Constantin Seibold, Alexander Freytag, Erik Rodner, and Rainer Stiefelhagen. Graph-constrained contrastive regularization for semi-weakly volumetric segmentation. In *European Conference on Computer Vision*, pages 401–419. Springer, 2022. 1
- [36] Simon Reiß, Constantin Seibold, Alexander Freytag, Erik Rodner, and Rainer Stiefelhagen. Decoupled semantic prototypes enable learning from diverse annotation types for semi-weakly segmentation in expert-driven domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15495–15506, 2023. 1, 5
- [37] Sucheng Ren, Xiaoke Huang, Xianhang Li, Junfei Xiao, Jieru Mei, Zeyu Wang, Alan Yuille, and Yuyin Zhou. Medical vision generalist: Unifying medical imaging tasks in context. *arXiv preprint arXiv:2406.05565*, 2024. 2
- [38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3
- [39] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. 6
- [40] Mennatullah Siam, Boris N Oreshkin, and Martin Jagersand. Amp: Adaptive masked proxies for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5249–5258, 2019. 1
- [41] Yanpeng Sun, Qiang Chen, Jian Wang, Jingdong Wang, and Zechao Li. Exploring effective factors for improving visual in-context learning. *arXiv preprint arXiv:2304.04748*, 2023. 2, 4
- [42] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018. 1
- [43] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 3, 4
- [44] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2, 3
- [45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference*

- on *Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 6
- [46] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023. 1, 2
- [47] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023. 2
- [48] Tianhao Wu, Janice Lan, Weizhe Yuan, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. Thinking llms: General instruction following with thought generation, 2024a. URL <https://arxiv.org/abs/2410.10630>, 2023. 8
- [49] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5542–5551, 2018. 1
- [50] Yifan Yang, Houwen Peng, Yifei Shen, Yuqing Yang, Han Hu, Lili Qiu, Hideki Koike, et al. Imagebrush: Learning visual in-context instructions for exemplar-based image manipulation. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [51] Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *arXiv preprint arXiv:2406.07550*, 2024. 2
- [52] Wentao Zhang, Junliang Guo, Tianyu He, Li Zhao, Linli Xu, and Jiang Bian. Video in-context learning. *arXiv preprint arXiv:2407.07356*, 2024. 2
- [53] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What makes good examples for visual in-context learning? *Advances in Neural Information Processing Systems*, 36:17773–17794, 2023. 2