COMiT: Learning Structured Visual Tokens through Sequential Communication

Anonymous authors
Paper under double-blind review

Abstract

Discrete image tokenizers provide a sequential interface for vision and multimodal models, but are typically optimized for reconstruction or compression and therefore tend to encode local appearance rather than object-level structure. We introduce COMiT, a communication-inspired framework for learning structured discrete visual representations. COMiT constructs a fixed-length latent message through sequential visual observations: at each step, a transformer processes a localized image crop and updates, refines, and reorganizes the existing token sequence. After several iterations, the resulting message conditions a flow-matching decoder that reconstructs the complete image. The encoder and decoder are implemented within a single transformer and trained end-to-end using flow-matching reconstruction and semantic representation-alignment objectives. COMiT substantially improves compositional generalization and relational reasoning over prior methods. Our experiments show that, while semantic alignment helps ground the representation, attentive sequential tokenization is critical for inducing more interpretable, object-centric token structures.

1 Introduction

Modern multimodal systems increasingly process vision through the lens of sequence modeling (Sun et al., 2024; Chen et al., 2025a): images are converted into discrete token sequences that can be consumed by transformer-based architectures. This token-based interface enables scalable training, efficient compression, and unified reasoning over visual and textual inputs. Consequently, learning effective image tokenizers has emerged as a central problem in visual representation learning.

Conventional discrete encoders (Van Den Oord et al., 2017; Esser et al., 2021) typically represent images as two-dimensional grids of tokens and are trained primarily for reconstruction under a compression constraint. As a result, the learned tokens often capture local texture and patch statistics rather than object-level semantic structure, limiting their interpretability and usefulness for downstream image understanding tasks.

Recent progress in image tokenization has renewed interest in *one-dimensional* discrete bottlenecks, which better match the sequential format expected by transformer-based models and can improve the semantic organization of token sequences (Yu et al., 2024a; Wang et al., 2025; Bachmann et al., 2025; Duggal et al., 2024). However, despite this syntactic compatibility, most existing approaches remain primarily optimized for compression trade-offs. As a result, semantic information is often entangled and poorly localized across tokens, limiting performance on downstream tasks that require compositional, object-centric structure.

In this work, we shift the focus from compression trade-offs to the *semantic organization* of visual token sequences. While semantic information can be encouraged through alignment with pretrained visual representations, we argue that such supervision alone is insufficient to induce structured and interpretable tokens. Instead, structured visual tokenization requires coupling semantic learning objectives with an encoding procedure that explicitly encourages compositional organization.

Our approach is inspired by how humans communicate about visual scenes. When describing a scene containing multiple objects, a speaker typically attends to one region at a time, sequentially incorporating salient information into the message. This incremental process allows the listener to integrate each observation into a coherent mental model of the scene while gradually reducing

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

uncertainty (Hassabis & Maguire, 2007). Under limited communication bandwidth, descriptions naturally prioritize high-level entities and their relations over fine-grained details, yielding a certain coarse-to-fine hierarchy in the space of possible descriptions. Notably, while so far we have considered the speaker and the listener to be two separate roles, humans are able to simultaneously perform both tasks. For instance, one could imagine the setting where the speaker is trying to memorize the scene for a certain period of time and then is asked to recall it.

Motivated by this perspective, we study how a communication-inspired encoding process can shape the structure of learned visual token sequences. In particular, we adopt two key design principles:

- **Attentive and sequential tokenization.** The encoder processes the image as a sequence of localized observations, attending to different regions at each step and incrementally integrating information into a discrete latent message.
- **Homogeneous communication.** In contrast to traditional autoencoders that use separate encoder and decoder networks, we adopt a unified design in which the same network acts as both "speaker" and "listener", mirroring the symmetry in human communication.

Building on these principles, we propose *COMmunication inspired Tokenization* (COMiT), which formulates image encoding as an iterative communication-and-reconstruction game. At each step, the model observes a new image crop and updates its discrete latent message. After several steps, the final message serves as a compact representation of the scene, from which the same network reconstructs the full image. The whole pipeline is trained as a generative model within the flow-matching framework, so that reconstruction acts as a learning signal rather than the end goal: we position COMiT as an understanding-oriented tokenizer with reconstruction capability rather than as a state-of-the-art reconstruction codec. To encourage semantic grounding, we incorporate a semantic representation alignment objective that distills high-level features from a frozen self-supervised vision model. Crucially, while alignment provides semantic signal, the attentive sequential tokenization determines how this information is distributed and localized across tokens.

To evaluate these properties, we introduce a suite of benchmarks that probe not only semantic content, but also compositional generalization and relational reasoning. Across these evaluations, COMiT consistently outperforms existing one-dimensional discrete image encoders by a substantial margin. Ablations further reveal complementary effects: semantic alignment improves token meaning, while communication-inspired sequential encoding induces more interpretable, object-centric tokens.

2 Related Work

Attentive Encoding. The idea of aggregating information from partial observations of a scene has a long history. One early example is the Recurrent Attention Model (RAM) proposed by Mnih et al. (2014), where a model iteratively selects regions of an image to attend to and accumulates a latent state to solve classification. In that work, the primary motivation was computational efficiency on large images, rather than representation learning or reconstruction. In other early approaches (Eslami et al., 2016; 2018) the reconstruction serves as the main training signal. However, often the latent state in such models explicitly encodes inductive biases, such as object-level bounding box coordinates (Eslami et al., 2016), rather than abstract concepts. Moreover, the early methods were only tested on rather toy data, and, to the best of our knowledge, there is no evidence on scaling these to real images.

Image Tokenization. Early works such as VQ-VAE (Van Den Oord et al., 2017; Razavi et al., 2019) introduced discrete visual codebooks learned via vector quantization. Subsequent approaches, including MaskGIT (Chang et al., 2022) and VQ-GAN (Esser et al., 2021), improved generative fidelity through masked prediction and adversarial training. More recently, several works have explored *one-dimensional* tokenization schemes, with Yu et al. (2024a) pioneering this direction. Many of these methods, however, continue to treat tokenization primarily as a compression and reconstruction objective, producing tokens in a single encoding pass without explicit mechanisms for sequential refinement. Other works have investigated how token ordering can reflect properties of the underlying visual signal, such as semantic hierarchies (Bachmann et al., 2025) or frequency structure (Wang et al., 2025). Closest to our approach, Duggal et al. (2024) also introduce an

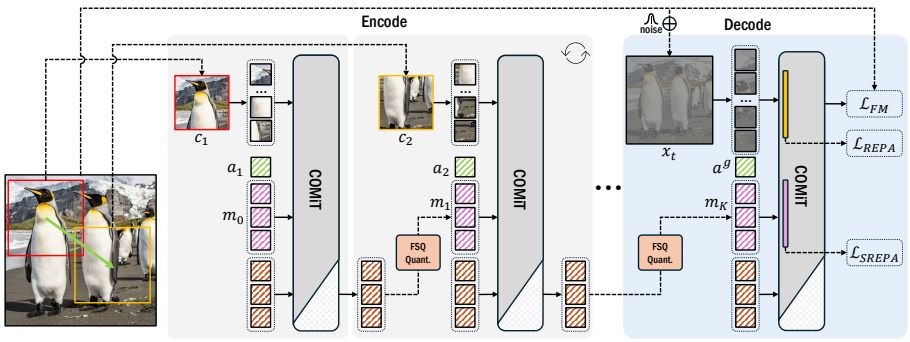


Figure 1: The overall training pipeline of COMiT. A sequence of K random crops is extracted from the input image and iteratively embedded into the latent message m_K that is discretized via FSQ Mentzer et al. (2024). The latter is decoded by the same model using the flow matching objective Lipman et al. (2023). Additionally, we use REPA Yu et al. (2024b) to speed up the training and SREPA to inject more semantic priors into the latent message.

iterative recurrent procedure for refining the latent representation. In their setting, however, the first encoding step is designed to capture most of the image content, while subsequent refinement steps use additionally allocated capacity. Overall, existing tokenizers remain largely optimized for reconstruction-compression trade-offs rather than structure and semantics. Appendix E expands this discussion with more recent and concurrent tokenizers.

Flow Matching and Diffusion Autoencoders. Flow-matching generative models (Lipman et al., 2023; Liu et al., 2022) offer stable and efficient training of continuous data distributions. These frameworks have also been used to train decoders in autoencoders (Preechakul et al., 2022; Guo & Schwing, 2025; Bachmann et al., 2025; Chen et al., 2025b). However, to the best of our knowledge, we are the first to leverage this formulation to train both encoding and decoding stages through a unified, differentiable flow objective, integrating representation learning and generation into a single network.

3 Communication Inspired Tokenization

In this section we start by describing the high-level training pipeline of COMiT that is illustrated in Figure 1. We then talk about some crucial details in Sections 3.1–3.5.

Encoding. Given an input RGB[1] image $x \in \mathbb{R}^{3\times H\times W}$, we define a sequence of random crops $\{c_k\}_{k=1}^K$, $c_k \in \mathbb{R}^{3\times h\times w}$ and their corresponding locations $\{l_k\}_{k=1}^K$, $l_k \in [-1,1]^2$. Here l_k is the absolute location of the c_k crop's center in the image x with each component normalized to $[-1,1]$. From these, we derive the sequence of offsets $\{a_k\}_{k=1}^K$, where $a_1 = (0,0)$ and $a_k = l_k - l_{k-1}, k \geq 2$. The offsets can be interpreted as actions the agent takes to scan the scene. We opt for relative offsets instead of global grounding to prevent parts of the latent message to specialize on certain constant regions.

In the beginning, the latent message $m_0 \in \mathbb{R}^{L\times d}$ is initialized with the same L d-dimensional vectors from a dictionary of N unique tokens. We refer to L as the message length and N as the vocabulary size. After that, at each step, the model f_θ takes as input the current latent message m_{k-1} and updates it with the information from the current crop:

$$m_k = f_\theta^m(c_k, t_k, a_k, m_{k-1}). \tag{1}$$

Here, the superscript m in f_θ^m denotes the readout part of the model's output that corresponds to the updated message. This is needed to support different output modalities that we introduce below. The second argument in f_θ^m corresponds to the denoising timestamp, which is $t_k = 1$ for the clean crops. This is needed to distinguish the encoding from the decoding regime, which we also introduce below.

[1] In fact, COMiT works in the latent space of a pretrained 2D VAE, but we omit the VAE in our notation as it does not affect any of the training steps.

The predicted message m_k is then quantized via FSQ Mentzer et al. (2024) to project the tokens to the vocabulary and fed back to the model along with the newly observed crop, forming a recurrent loop. After K steps, the final message m_K is obtained.

Decoding. The message derived through the above procedure of recurrent crop aggregation serves as conditioning for flow-based decoding of the full image. However, in contrast to the conventional autoencoder architectures, COMiT uses the same model for both encoding and decoding. That is, we start by creating a noisy version of the image $x_t = tx + (1-t)\varepsilon$ with $t \sim U[0,1]$ – random interpolation time and $\varepsilon \sim \mathcal{N}(0, I)$ – random noise instance. Then, the noisy image x_t is fed to the network along with the final message m_K and the offset relative to the last crop $a^g = -l_K$. The network is then supervised to predict the velocity of the marginal flow at x_t via the standard conditional flow-matching loss (Lipman et al., 2023):

$$\mathcal{L}_{\text{FM}} = \mathop{\mathbb{E}}_{\substack{t\sim U[0,1] \\ \varepsilon\sim\mathcal{N}(0,I)}} \left\| f_\theta^v(x_t, t, a^g, m_K) - (x - \varepsilon) \right\|_2^2 . \tag{2}$$

Here, as before, the superscript v in f_θ^v denotes the readout part of the model's output that corresponds to the estimated velocity.

Inference. Given a latent message $\hat{m}$, in order to visually probe its content, one may decode the message through numerically integrating the flow ODE Lipman et al. (2023). That is, starting at a noise sample $x_0 \sim \mathcal{N}(0, I)$, one obtains the clean image x_1 via iteratively denoising the current estimate:

$$x_{t+h} = x_t + h f_\theta^v(x_t, t, a^g, \hat{m}), \tag{3}$$

where h is the step size connected to the number of function evaluations (NFE) via $h = 1/\text{NFE}$.

3.1 Greedy Use of Tokens

As noted in the introduction, as the amount of information embedded in the latent message increases (in our case, the number of crops), it is natural to expect the model to discard irrelevant secondary details and retain only the most essential features for reconstruction. This process would naturally induce a hierarchical structure in the feature space. However, if during training the number of crops the model aggregates were fixed, the network would instead learn to pre-allocate parts of the latent message for future crops. This would effectively enforce a fixed capacity per crop, with each information chunk occupying a predetermined location in the message. To avoid this behavior, we randomize the number of crops the model processes during training. As a result, the model never knows whether additional crops will follow the current one and is therefore encouraged to use the available tokens greedily. This corresponds to the setting with fixed capacity and varying amount of information, which is precisely the regime we were aiming for (see Appendix D.4 for ablations on the number of crops).

In addition, backpropagating gradients through the entire sequence of message updates would be computationally expensive and memory-intensive. To address this, we apply a stop-gradient operation to all updates except the final one. Combined with the randomized number of crop aggregations, this further promotes greedy token usage while enabling efficient training. In preliminary small-scale experiments, we found that restricting gradient backpropagation in this way has only a moderate impact on performance (see Appendix D.3 for a full-scale ablation).

3.2 Distilling Semantic Representation

As noted in the introduction, to enforce the semantic ordering of the information in the learned messages, we directly distill pretrained SSL features $\psi(x) \in \mathbb{R}^s$(namely, the `[CLS]` of DINOv2 Oquab et al. (2023)) into COMiT's intermediate representations that correspond to the message tokens. We call this procedure semantic representation alignment, or SREPA. More precisely, during the decoding stage, the intermediate representations in the j-th layer of the network $f_\theta^m[j](x_t, t, a^g, m_K) \in \mathbb{R}^{L\times r}$ are first projected with a small MLP and then averaged pooled into a single vector $\bar{f}_\theta^m[j](x_t, t, a^g, m_K) \in \mathbb{R}^s$. Then, the following loss is calculated:

$$\mathcal{L}_{\text{SREPA}} = \exp\left(-\text{Sim}\left(\psi(x), \bar{f}_\theta^m[j](x_t, t, a^g, m_K)\right)\right), \tag{4}$$

where $\text{Sim}(\cdot,\cdot)$ is the cosine similarity.

3.3 Reconstruction Fidelity

Although our main focus in this paper is on semantics of the token sequences, we apply several techniques to improve the reconstruction fidelity of COMiT. First, to enable classifier-free guidance (CFG) Ho & Salimans (2022) at inference, during training we skip the encoding part with probability $p_{\text{CFG}} \approx 0.2$, thereby training an unconditional decoder. Besides this, with probability $p_{\text{G}} \approx 0.5$ we replace the first crop with the full original image, allowing the model to encode the whole image in a single step. For clarity, we refer to the full image as the *global crop*, and to the smaller crops as *local crops*.

As a result, the training consists of three regimes: in some cases, COMiT first observes the global crop and then refines its latent message with local crops; in others, it observes only local crops; and the third portion of the training is devoted to the decoding of the image from an empty message.

3.4 Implementation Details

In practice, COMiT is implemented as a transformer encoder network Vaswani et al. (2017). Following the standard practices of DiT Peebles & Xie (2023), we use AdaLN Perez et al. (2018) layers to make the network timestamp-conditioned. Notably, different modalities (image, message) use separate projections in the AdaLN layers. All inputs to the network are flattened into a single sequence of tokens. Images are first encoded with a pretrained VAE (namely, the KL-regularized VAE of Stable Diffusion Rombach et al. (2022)) to reduce the input size and then patchified. The offset is embedded into a single token via linear projection. We found that the model performs slightly better when, instead of overwriting the previous message, one feeds both the previous message and newly initialized buffer tokens to predict the next message (see also Figure 1). This architectural design also allows us to apply causal masks in the attention maps of the buffer tokens, mimicking the causal nature of composing sentences (illustrated in Figure 1 as the tilted line of the gray shaded region in COMiT's main backbone). We train COMiT in three model sizes: B (12 layers, hidden size 768), L (24 layers, hidden size 1024) and XL (28 layers, hidden size 1152). All models are trained on ImageNet1k Deng et al. (2009). To speed up the convergence, we use REPA Yu et al. (2024b) that aligns the intermediate image representation in the network with the DINOv2's spatial features. The final loss thus takes the form:

$$\mathcal{L} = \mathcal{L}_{\text{FM}} + \lambda_{\text{REPA}}\mathcal{L}_{\text{REPA}} + \lambda_{\text{SREPA}}\mathcal{L}_{\text{SREPA}}. \tag{5}$$

In our experiments, we choose $\lambda_{\text{REPA}} = \lambda_{\text{SREPA}} = 0.5$.

The whole training procedure is summarized in Figure 1. Implementation details are in Appendix A.

3.5 Cropping Policies

Because both the crops and their number are randomized during training, COMiT offers flexibility at inference time in how crops are selected when encoding an image. We refer to these choices as *cropping policies*. Random crops thus only serve as a training distribution that supplies diverse scan paths without annotations. The policy used at test time is a free design choice: it can cover the image exhaustively, follow a fixed order, or select regions based on the model's current uncertainty, and in principle it could also express approximations of human-like scan paths. Although many cropping policies are possible, in this paper, we study only a limited subset to illustrate the impact of the main design decisions, leaving a more exhaustive exploration to future work. For simplicity, we resize all images to a resolution of 256×256 and consider only 96×96 crops arranged on a 3×3 grid. We characterize cropping policies by three factors: (i) whether a global crop is included at the beginning of the crop sequence, (ii) the number of crops, and (iii) the order in which the crops are processed. With respect to the latter, we consider random ordering, a fixed raster-scan order, and an adaptive policy. The adaptive policy decodes the current latent message after each newly aggregated crop and selects the next crop to be the one that has the largest mean reconstruction error. The intermediate reconstructions are obtained using a single decoding step, resulting in blurry regions in areas where the model is uncertain about the content (see Figures 4 and 3).

Table 1: The effect of SREPA during training. Both models use the same cropping policy (single global crop) at test time. Parentheses: gain over the ablation.

Model variant	SREPA	IN100 top-1↑	CSSD mIoU↑
COMiT-B	✓	**82.91** (+10.65)	**0.527** (+0.020)
COMiT-B	–	72.26	0.507

Table 2: The effect of using recurrent local crops aggregation during training. At test time, both models use the same cropping policy with a single global crop. Parentheses: gain over the ablation.

Model	Local crops at training	IN100 top-1↑	CSSD mIoU↑
COMiT-B	✓	**82.91** (+1.97)	**0.527** (+0.189)
COMiT-B	–	80.94	0.338

4 EXPERIMENTS

In this section, we evaluate COMiT to showcase its improved semantic encodings and emergent properties. We design the test suite of three representative benchmarks.

Visual recognition (ImageNet100). To evaluate whether COMiT encodes high-level semantic information useful for visual recognition, we perform classification probing on ImageNet100 (IN100) Deng et al. (2009). Instead of standard linear probing, we adopt an attention-based probing strategy using a lightweight two-layer of self-attention operating over the 1D sequence of tokens produced by our model and the baselines. This design allows the probe to flexibly attend to and aggregate information across the entire token sequence, without assuming that class-discriminative features are localized at a fixed position. Attention probing is therefore better suited for the 1D tokenization setting, where semantic information may be distributed across tokens. It also allows comparing tokenizers that emit different numbers of tokens per image. As an alternative, attention-free probe, Appendix D.2 reports a single affine readout over the flattened token sequence.

Compositional generalization (MSCOCO). We evaluate compositional generalization using a benchmark derived from MSCOCO Lin et al. (2014), focusing on images that contain two object categories. The dataset is split into two disjoint partitions such that both splits share the same set of object categories, while specific object co-occurrences (i.e., object pairs) appear in only one split. We train the probing network on the first split and evaluate it on the second, ensuring that the probe must generalize to unseen object compositions. This setup tests whether object-related information is stored in a disentangled manner across the token sequence (e.g., distributed across different subsets of tokens corresponding to individual objects, rather than being collapsed into a single token that encodes the entire scene). Indeed, if the model mixed the information about two objects in a single token, the probe would have a more difficult job to effectively generalize to unseen object pairs.

Inter-object relations (Visual Genome). We evaluate the ability of models to encode relational semantics using relationship prediction on Visual Genome (VG) Krishna et al. (2017). Given a batch of images paired with candidate subject–object relations, we test whether the probing network can correctly assign each image to its corresponding relation. This benchmark assesses whether inter-object relationships are explicitly represented in the token sequence, beyond individual object semantics, and whether such relational information can be recovered through lightweight probing.

For more details on our testing suite, we refer the reader to the Appendix B.

4.1 ABLATIONS

Semantic distillation. We evaluate the impact of SREPA by comparing the full COMiT-B model against an ablated variant in which $\lambda_{\text{SREPA}} = 0$ during training. Both models are evaluated under identical settings on IN100. We observe that SREPA has significant contribution to our tokenizer's performance (see Table 1). Note that since FSQ uses the same codebook across model variants, all the differences in performance may be attributed to token co-occurrences.

Attentive tokenization. Next, we evaluate the influence of the presence of local crops during the training. To this end, we train COMiT-B with and without local crops. The variant trained without local crops lacks recurrent message refinement and effectively reduces to a standard encoder-decoder architecture with a discrete bottleneck.

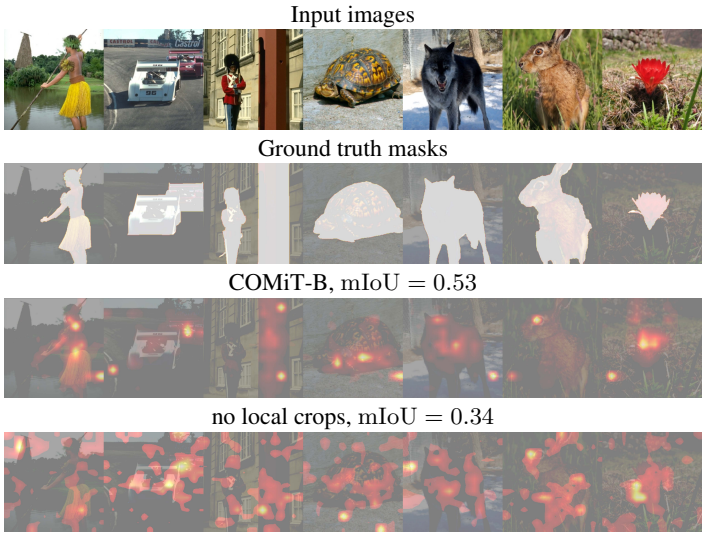


Figure 2: Effect of attentive tokenization on the visual grounding of tokens. Training with local crops yields much better token–object alignment than training with global crops only. Attention maps from the 10th layer of COMiT-B; both models embed only the global crop.

First, we evaluate both variants on a visual recognition benchmark using the same cropping policy at inference time, which encodes only the global crop. Although local crops are not used to construct the latent message at test time, the model trained with the attentive communication bias produces better structured latent messages as reflected with the probing accuracy shown in Table 2.

Next, we analyze token attention maps following Duggal et al. (2024). Using CSSD (Yan et al., 2013), we binarize each attention map by retaining the top $Q\%$ of values and select, per image, the token whose map attains the highest IoU (Everingham et al., 2010) with the ground-truth mask. All variants are frozen ImageNet-pretrained models evaluated on all 200 CSSD images with the same single-global-crop policy (10th-layer attention at $t = 0.1$, $Q = 30$); as the mask selects the best token, this is an oracle diagnostic rather than a segmentation method (see Appendix D.1 for details). Figure 2 compares these token attention maps for COMiT-B and a variant trained without local crops. COMiT produces attention maps tightly aligned with objects, achieving a mean mIoU of 0.53, whereas the ablated model exhibits diffuse attention across the image and reaches only 0.34 mIoU (95% image-bootstrap interval of the difference: $[0.175, 0.204]$). This alignment emerges from our attentive tokenization pipeline, indicating that while distillation enforces semantics, tokenization pipeline is key to structuring messages. Table 1 reports the same diagnostic for the SREPA ablation: removing SREPA costs 10.65 IN100 points but only 0.020 mIoU, whereas removing local crops costs 1.97 points but 0.189 mIoU. Tables 1 and 2 thus show complementary signatures: SREPA mostly governs the category content of the message, whereas recurrent local crops mostly govern where it is stored. Additional results are provided in Appendix C.

Cropping policy. Table 3 compares different cropping policies. A single global crop yields competitive average performance across tasks while incurring the lowest test-time cost (the number of crops). Based on this observation, we adopt global-only cropping as the default for the remainder of our probing experiments unless stated otherwise. Importantly, COMiT also naturally supports test-time scaling: adding local crops provides modest gains on compositional generalization and inter-object relations, suggesting that local processing can benefit certain benchmarks. Notably, all differences across cropping policies arise purely at test-time, without retraining the model.

4.2 Quantitative Results

We evaluate COMiT against relevant prior work in the 1D tokenization domain (see Table 4). For a fair comparison, we consider only unsupervised methods that do not rely on additional paired data from other modalities, such as text. We report results on our semantic probing test suite along with

Table 3: Effect of the cropping policy with COMiT-B. A single global crop gives the best trade-off between performance and test-time cost, while some tasks benefit from additional local crops.

Global crop	# Crops	Ordering	IN100 top-1 ↑	MSCOCO top-5 ↑	VG top-1 ↑
✓	1	–	**82.91**	**41.46**	52.11
✓	10	Random	**82.30**	40.21	52.12
-	9	Random	74.45	38.92	54.96
✓	10	Raster-scan	82.22	**41.83**	**55.78**
-	9	Raster-scan	80.62	38.61	51.10
✓	3	Adaptive	82.09	40.68	**55.00**
–	3	Adaptive	74.96	32.70	50.22

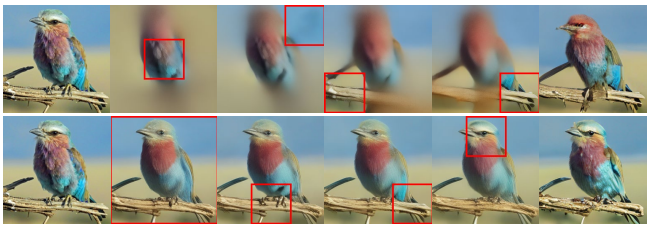


Figure 3: The difference between the *adaptive* (top) and *global+adaptive* (bottom) cropping policies. In both cases COMiT aggregates the crops of the input image (the leftmost column) into the latent message and decodes it to obtain the reconstructed image (the rightmost column, 10 NFE with CFG $= 7.5$). The columns in-between show which crops are selected together with immediate single step reconstructions (1 NFE with CFG $= 1.0$).

rFID Heusel et al. (2017) and PSNR Huynh-Thu & Ghanbari (2008) on ImageNet1k Deng et al. (2009) validation set. All methods, if not stated otherwise, were pretrained on ImageNet1k Deng et al. (2009). COMiT consistently outperforms prior work on semantic tasks, including FlexTok, which also distills DINOv2 features through REPA.

The semantic probes measure what the complete message encodes, but not whether individual tokens are tied to objects. To compare tokenizers on this axis, Appendix D.5 scores the attention map of every individual token against the object masks of multi-object COCO images, with the extraction layer calibrated separately for each tokenizer. COMiT-L localizes 36.52% of the large objects, 14.78 points more than FlexTok-d18-d18, and spreads its best matches over twice as many tokens. Importantly, token-level localization does not rely on SREPA: COMiT-B trained without SREPA, whose semantic supervision then matches that of FlexTok (DINOv2 REPA only), localizes objects as often as FlexTok-d18-d18 (23.48% versus 21.74%) with a fifth of its parameters, whereas removing the recurrent local crops lowers localization to 17.39%.

Similarly to prior work Bachmann et al. (2025), we observe that scaling the model from B to L improves both reconstruction and representation, and further scaling the model from L to XL allocates the additional capacity to aid reconstruction, while diminishing the semantics.

Among the tokenizers evaluated under our frozen-token protocol, COMiT attains the best semantic scores while trading off some reconstruction quality. This representation-reconstruction trade-off is intentional: our work argues that improving the semantics of token sequences is important for better performance on understanding tasks. The trade-off is not fixed: scaling from B to L improves rFID from 11.06 to 3.67 while all three semantic scores improve, and enlarging the bottleneck improves PSNR at unchanged probing accuracy (Figure 5c in the Appendix). At the same time, we believe the reconstruction fidelity of COMiT can be further improved with additional training. Many prior methods use multi-stage training pipelines, including decoder fine-tuning to improve reconstruction quality (Yu et al., 2024a; Chang et al., 2023). Importantly, weaker reconstruction does not imply weaker spatial understanding. To demonstrate this, we additionally ran segmentation probing over a subset of MSCOCO, where SelfTok-512 achieves 0.44 IoU, while COMiT-L achieves

Table 4: We evaluate COMiT and several baseline 1D tokenizers on our test suite comprising three representative benchmarks described in Section 4. COMiT outperforms prior work on semantic probing. †: trained on larger data mixtures.

Method	#params	msg len.	voc.	IN1k rFID↓	IN1k PSNR↑	IN100 top-1↑	MSCOCO top-5↑	VG top-1↑
TiTok-L (Yu et al., 2024a)	614M	32	2^{12}	2.21	15.60	17.26	6.22	26.06
TiTok-B (Yu et al., 2024a)	172M	64	2^{12}	1.70	16.80	19.43	12.64	27.31
TiTok-S (Yu et al., 2024a)	44M	128	2^{12}	1.71	17.52	19.47	6.30	26.81
ALIT (Duggal et al., 2024)	229M	256	2^{10}	7.39	19.76	29.43	8.56	32.14
FlexTok$_{(d12\text{-}d12)}$ (Bachmann et al., 2025)	254M	256	64k	4.20	18.41	80.25	38.17	53.75
FlexTok$_{(d18\text{-}d18)}$ (Bachmann et al., 2025)	860M	256	64k	1.61	18.59	81.54	38.58	54.46
FlexTok$_{(d18\text{-}d28)}$ (Bachmann et al., 2025)	2.5B	256	64k	1.45	18.46	80.93	39.14	54.35
SelfTok Wang et al. (2025)†	2.17B	512	2^{15}	**0.70**	**24.14**	40.64	17.91	37.15
SelfTok Wang et al. (2025)†	2.17B	1024	2^{16}	**0.54**	**26.30**	35.90	17.32	36.69
COMiT-B (ours)	174M	256	64k	11.06	17.75	82.91	**41.46**	52.11
COMiT-L (ours)	610M	256	64k	3.67	17.81	**85.80**	**45.31**	**56.42**
COMiT-XL (ours)	900M	256	64k	3.50	17.83	**84.69**	39.45	**55.61**

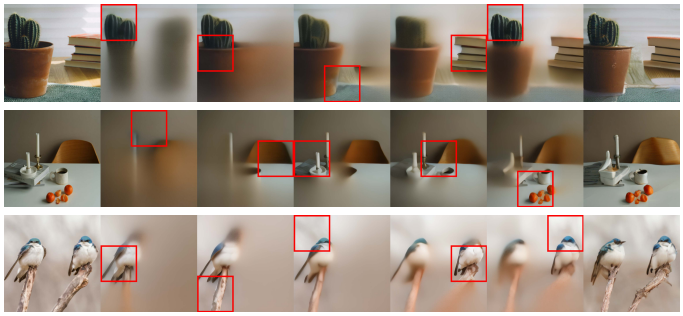


Figure 4: The way COMiT adds information to the latent message is inherently compositional.

0.53 IoU (identical frozen-token cross-attention decoder predicting the binary foreground mask; see Appendix D.6 for details). This shows that lower rFID, e.g., 0.70 for SelfTok-512, does not necessarily translate to better dense prediction performance.

4.3 Qualitative Results

Evolution of Uncertainty. One way to visually probe the information encoded in the latent messages is via reconstructing the images. However, in the case when only local crops are used to encode the message, 1–2 crops may not be enough to reliably reconstruct the image. Since COMiT is a generative model of images given the message, it will hallucinate the content in the unobserved regions. However, the flow matching framework guarantees that at $t = 0$ the optimal velocity field is the one that points to the mean of the data that shares the same conditioning. Therefore, decoding messages with a single denoising step results in images that have sharp details in the recently observed regions and blurry content in the regions where the model is not certain about the scene. Figures 3 (top) and 4 demonstrate how the information in the latent message is gradually refined with more crops aggregated. This process is also inherently compositional. Figure 4 shows how objects are added to the latent messages as soon as observed. More visual results can be found in Appendix F.

5 Conclusion and Discussion

COMiT opens a path toward more structured and semantically meaningful 1D visual tokenization. By treating tokenization as an iterative communication process, it encourages visual information to be organized across discrete token sequences in a way that better supports semantic grounding, compositionality, and relational reasoning. Future directions are discussed in Appendix G.

AI USE STATEMENT

Among the tasks for which ICLR requires disclosure, we used generative AI tools to implement methods and to interpret results. Concretely, AI assistance was used to write and run evaluation code for the additional analyses in Appendix D, including the token-localization study on COCO and the affine readout, to recompute confidence intervals from the stored per-sample results, and to propose readings of the resulting measurements. We have not used generative AI tools to develop the conceptual framework of this study or to propose or refine its hypotheses: the communication-inspired formulation, the COMiT architecture and training objective, the SREPA loss and the probing benchmarks are the authors' own. Generating synthetic data sets, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and translation are not applicable to this work, as all data are public real-image datasets and the paper contains no theoretical results. Among the tasks for which disclosure is recommended, we used generative AI tools to draft, restructure and edit parts of the paper for readability, to format tables, and to search for and verify bibliographic information on related work. We have reviewed all AI-assisted work. Every reference was checked against the original publication before being cited, and AI-suggested interpretations were retained only where the measurements supported them. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The architecture, training objective and loss weights are specified in Section 3, and all training hyperparameters are listed in Appendix A (Table 5). The cropping policies used at inference are defined in Section 3.5. The construction of the probing benchmarks, the probe architectures and their hyperparameters (Table 6), and the sampling settings used for reconstruction are given in Appendix B. The protocols of all additional analyses are given in Appendix D. All training and evaluation data are public (ImageNet-1k, MSCOCO, Visual Genome and CSSD). Each tokenizer is a single pretraining run; the affine readouts report five probe seeds, and the localization results report bootstrap intervals over images. The code and the pretrained models will be made publicly available upon publication.

ETHICS STATEMENT

This work studies image tokenizers trained and evaluated on public image datasets. It involves no human subjects, and no new data were collected. As with other generative image models, COMiT could be misused to produce misleading content and may reflect biases present in its training data; these considerations are discussed in Appendix H.

REFERENCES

Roman Bachmann, Jesse Allardice, David Mizrahi, Enrico Fini, Oğuzhan Fatih Kar, Elmira Amirloo, Alaaeldin El-Nouby, Amir Zamir, and Afshin Dehghan. Flextok: Resampling images into 1d token sequences of flexible length. In *Forty-second International Conference on Machine Learning*, 2025.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11315–11325, 2022.

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.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025a.

Yinbo Chen, Rohit Girdhar, Xiaolong Wang, Sai Saketh Rambhatla, and Ishan Misra. Diffusion autoencoders are scalable image tokenizers. *arXiv preprint arXiv:2501.18593*, 2025b.

Yitong Chen, Zuxuan Wu, Xipeng Qiu, and Yu-Gang Jiang. Catok: Taming mean flows for one-dimensional causal image tokenization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23161–23171, June 2026.

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*, pp. 248–255. Ieee, 2009.

Shivam Duggal, Phillip Isola, Antonio Torralba, and William T Freeman. Adaptive length image tokenization via recurrent allocation. In *First Workshop on Scalable Optimization for Efficient and Adaptive Foundation Models*, 2024.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018.

S. M. Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S. Morcos, Marta Garnelo, Avraham Ruderman, Andrei A. Rusu, Ivo Danihelka, Karol Gregor, David P. Reichert, Lars Buesing, Theophane Weber, Oriol Vinyals, Dan Rosenbaum, Neil Rabinowitz, Helen King, Chloe Hillier, Matt Botvinick, Daan Wierstra, Koray Kavukcuoglu, and Demis Hassabis. Neural scene representation and rendering. *Science*, 360(6394): 1204–1210, 2018. doi: 10.1126/science.aar6170.

SM Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Geoffrey E Hinton, et al. Attend, infer, repeat: Fast scene understanding with generative models. *Advances in neural information processing systems*, 29, 2016.

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*, pp. 12873–12883, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=FPnUhsQJ5B.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv preprint arXiv:2502.05171*, 2025.

Pengsheng Guo and Alexander G Schwing. Variational rectified flow matching. *arXiv preprint arXiv:2502.09616*, 2025.

Demis Hassabis and Eleanor A Maguire. Deconstructing episodic memory with construction. *Trends in cognitive sciences*, 11(7):299–306, 2007.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Yan Li, Ning Liao, Xiangyu Zhao, Shaofeng Zhang, Xiaoxing Wang, Yifan Yang, Junchi Yan, and Xue Yang. Evotok: A unified image tokenizer via residual latent evolution for visual understanding and generation. In *Computer Vision – ECCV 2026*, pp. 186–204. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-37553-7_11.

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 *European conference on computer vision*, pp. 740–755. Springer, 2014.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5(5):5, 2017.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. In *The Twelfth International Conference on Learning Representations*, 2024.

Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. *Advances in neural information processing systems*, 27, 2014.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Maitreya Patel, Jingtao Li, Weiming Zhuang, Yezhou Yang, and Lingjuan Lv. Vibetoken: Scaling 1d image tokenizers and autoregressive models for dynamic resolution generations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2058–2068, June 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Tushar Prakash, Onkar Susladkar, Inderjit S Dhillon, and Sparsh Mittal. Pyramidal spectrum: Frequency-based hierarchically vector quantized vae for videos. In *WACV*, pp. 2073–2082, 2026.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10619–10629, 2022.

Yunpeng Qu, Kaidong Zhang, Yukang Ding, Ying Chen, and Jian Wang. Semantic one-dimensional tokenizer for image reconstruction and generation, 2026. URL `https://arxiv.org/abs/2603.16373`.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Seyedmorteza Sadat, Otmar Hilliges, and Romann M. Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=e2ONKX6qzJ`.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Onkar Susladkar, Tushar Prakash, Adheesh Juvekar, Kiet A. Nguyen, Dong-Hwan Jang, Inderjit S Dhillon, and Ismini Lourentzou. Pyratok: Language-aligned pyramidal tokenizer for video understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 37906–37917, June 2026.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Bohan Wang, Zhongqi Yue, Fengda Zhang, Shuo Chen, Li'an Bi, Junzhe Zhang, Xue Song, Kennard Yanting Chan, Jiachun Pan, Weijia Wu, et al. Selftok: Discrete visual tokens of autoregression, by diffusion, and for reasoning. *arXiv preprint arXiv:2505.07538*, 2025.

Tianwei Xiong, Jun Hao Liew, Zilong Huang, Jiashi Feng, and Xihui Liu. Gigatok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18770–18780. IEEE, 2025.

Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1155–1162, 2013.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *Advances in Neural Information Processing Systems*, 37: 128940–128966, 2024a.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024b.

Anlin Zheng, Xin Wen, Xuanyang Zhang, Chuofan Ma, Tiancai Wang, Gang Yu, Xiangyu Zhang, and Xiaojuan Qi. Vision foundation models as effective visual tokenizers for autoregressive generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=PESrAH82Zh.

A Additional Implementation Details

In this section, we provide additional implementation details critical for reproducing the results in the paper. The code and the pretrained models will be made publicly available upon publication.

Model architecture. As noted in the main paper, the architecture of COMiT is based on DiT (Peebles & Xie, 2023). The only substantial difference is that we use different set of AdaLN (Perez et al., 2018) layers to modulate the image and the message tokens. This is similar to (Yang et al., 2024), where different modulations were applied to text and video tokens. Other architectural details, such as the number of layers and transformer dimensions can be found in Table 5.

Unified encoder and decoder. Our design choice was to avoid encoder-decoder separation and allow the model to autonomously decide which part of the network's capacity to dedicate to which task. This choice mimics the symmetry of the communication game, eases development by removing a degree of freedom in designing the architecture and potentially removes redundancies in the model's parameters. We leave further exploration of such fused architectures, including the analysis of how different tasks are balanced in a single network, to future work.

Training hyperparameters. All models are trained on full ImageNet1k (Deng et al., 2009) using Adam (Kingma & Ba, 2017) optimizer with a base learning rate of 3e-4 and square root decay schedule with 5000 iterations of warmup. The B and L variants were trained with batch size of 512 images, while for the XL variant we used batch size of 256 with 2 gradient accumulation steps, resulting in the same effective batch size of 512 images. Our models were trained for 200 epochs on 32 $\times$ GH200 GPUs. We use EMA with decay 0.999. For convenience, all hyperparameters are summarized in Table 5.

Training stability. Despite the recurrent encoding and the combined objective, optimization is straightforward: COMiT uses one shared transformer and one end-to-end loss (Equation 5) with a frozen DINOv2 teacher, and there is no discriminator, alternating optimization, or separate decoder fine-tuning stage. The stop-gradient through all but the last message update bounds activation memory, and gradient clipping and EMA are used as listed in Table 5. The same recipe converged for the B, L and XL variants without per-size tuning.

Cropping policy at training. During training the input images are of resolution 256×256. The local crops of resolution 96×96 are extracted at random locations within the image, such that the local crops do not go beyond the image domain. The number of crops K is randomly selected from $\{1, \ldots, K_{\max}\}$. We found that instead of uniformly selecting K it is better to slightly skew its distribution towards 1. More precisely, similarly to Geiping et al. (2025), we use the following sampling scheme:

$$K = \min(\max(1, \xi + 1), K_{\max}), \text{ where } \xi \sim \text{Poisson}(e^{\tau}), \text{ and } \tau = \ln(K_{\text{mode}}) + \varepsilon \cdot 0.5 - 0.125 \tag{6}$$

Flow matching timestamp. In the flow matching framework (Lipman et al., 2023), the default choice for the timestamp sampling distribution is uniform on the unit interval $U[0, 1]$. However, in prior work it has been discovered that sometimes it is better to skew the timestamp distribution towards the ends of the interval (Esser et al., 2024). We found that it is better to skew the distribution of the timestamps towards 0. This prioritizes high noise regimes and allows for better training of the encoder. We use the logitnormal distribution with $\sigma^2 = 1$ and $\mu = -1$.

Message handling. For initializing the message tokens we pick one fixed token from the vocabulary (namely with `id` 32000) and use it to initialize all tokens in m_0. Prior to feeding the message tokens to the DiT, we embed them using a 2 layer MLP with a SiLU (Elfwing et al., 2018) activation in between. As noted in the main paper, at each message update step, we append the current message with a new one that acts as a buffer to write the aggregated information from the message and the crop to. The tokens in the buffer message are copies of learned embedding.

Position encodings. Each modality in the input uses its own set of position encodings. The message and the offset tokens are augmented with learned position encodings. The image tokens use 2D sine–cosine position encodings. We define some maximum resolution of the image input (depends on the resolution of the images, the downsampling factor of the VAE and the patch size of the DiT). Each crop only uses the top-left corner of the 2D image position encodings.

	COMiT-B	COMiT-L	COMiT-XL
DiT parameters			
Depth	12	24	28
Hidden size	768	1024	1152
Heads	12	16	16
MLP ratio	4	4	4
REPA/SREPA parameters			
Representation layer	4	4	4
Number of MLP layers	3	3	3
MLP hidden size	768	768	768
Bottleneck parameters			
Number of message tokens	256		
Token dimensions	6		
Vocabulary size	64000		
FSQ Levels	[8, 8, 8, 5, 5, 5]		
Other training parameters			
Flow matching time distribution	logitnormal		
Cropping policy	randomized		
Mode of number of crops K_{mode}	1		
Maximum number of crops $K_{\max}$	9		
Number of updates to backpropagate through	1		
p_{CFG}	0.18		
p_{G}	0.55		
Optimizer	Adam Kingma & Ba (2017)		
Betas	[0.9, 0.999]		
ε	1e-8		
Weight decay	0.0		
Learning rate	0.0003		
Scheduler	square root		
Warmup steps	5000		
Gradient clipping	1.0		
Epochs	200		
Batch size	512	512	256
Gradient accumulation steps	1	1	2
EMA decay	0.999		

Table 5: Architecture and training details of COMiT variants.

Classifier-free guidance. To smoothen the dependence of the metrics on the CFG strength and allow for high values of the latter, similarly to Bachmann et al. (2025), we use the Adaptive Projected Guidance (Sadat et al., 2025) with the rescaling threshold as $r = 2.5$, parallel component as $\eta = 0$, and momentum as $\beta = -0.5$.

B Evaluation Details

B.1 Test Suite

We evaluate the ability of image tokenizers to encode semantic content using a probing protocol. Images are first tokenized using the evaluated tokenizers, and the resulting image-token sequences are kept frozen. For each benchmark, we train a lightweight probe with identical capacity across tokenizers to ensure a fair comparison. Across all tasks, we apply early stopping based on validation performance and report results from the best-performing checkpoint.

ImageNet100. We evaluate category-level semantic information using ImageNet100 as a single-label classification task. Given frozen image tokens, the probe is trained to predict one of 100 object

Benchmark	Model dim	Depth	Heads	Global Batch Size
ImageNet100	768	2	8	512
MSCOCO	768	2	8	256
Visual Genome	64	2	8	512

Table 6: Probe hyperparameters used for all benchmarks.

categories using a cross-entropy objective. Performance is measured using top-1 accuracy on the validation set.

MSCOCO. We assess compositional generalization using a multi-label classification task derived from MSCOCO. We retain only images containing exactly two object categories and construct pair-disjoint training and validation splits such that object category pairs observed at validation time are never seen during training. Each image is annotated with a 2-hot target vector over 90 categories. The probe is trained with a binary cross-entropy loss, and performance is reported using top-5 accuracy, requiring both ground-truth categories to appear among the top-5 predictions.

Visual Genome. We evaluate relational reasoning on Visual Genome, which provides supervision in the form of $(\text{subject}, \text{predicate}, \text{object})$ triplets with a large and open-ended vocabulary. Training and validation splits are constructed to be image-disjoint to avoid visual leakage. To ensure a meaningful probing signal, we filter out relations involving objects that occupy less than 5% of the image area and restrict the vocabulary to the 150 most frequent subject/object categories and predicates. This preprocessing avoids evaluation failures caused by out-of-vocabulary textual concepts at validation time, allowing the probe's performance to more faithfully reflect the semantic information encoded in the image tokens rather than limitations of the classifier's label space. For each image, the probe is trained to distinguish the correct triplet from a set of negative triplets sampled from other images using a contrastive objective. At evaluation time, accuracy is reported based on whether the correct triplet is identified among its associated negatives.

B.2 Probing Network Design

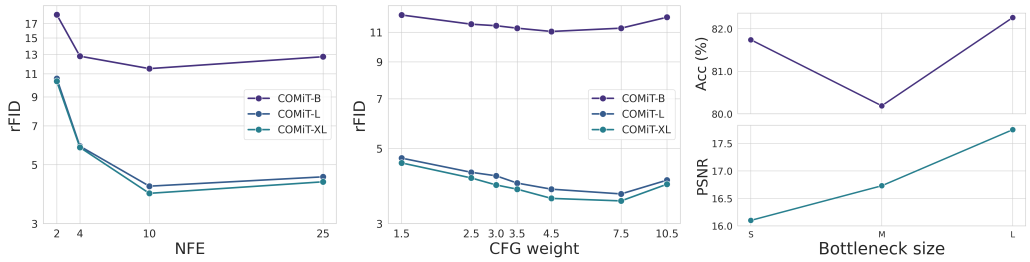
All probes operate on frozen image-token sequences and use a shallow Transformer encoder to aggregate information across tokens. Image tokens $\mathbf{m} \in \mathbb{R}^{L \times d}$ are first projected to a shared model dimension, and a learnable `[CLS]` token is prepended to the sequence. Learned positional embeddings and discrete type embeddings are added to all inputs before applying self-attention. Predictions are obtained from the final `[CLS]` representation.

For ImageNet100 and MSCOCO, the input sequence consists only of the `[CLS]` token followed by the projected image tokens. A two-way type embedding distinguishes image tokens from `[CLS]`. The Transformer output corresponding to `[CLS]` is passed to a linear classification head, producing either single-label logits (ImageNet100) or multi-label logits (MSCOCO).

For Visual Genome, the probe additionally conditions on a candidate $(\text{subject}, \text{predicate}, \text{object})$ triplet. Subject, object, and predicate text are encoded using a frozen T5 encoder (Raffel et al., 2020) and independently projected to the model dimension. These embeddings are appended to the image-token sequence, and a five-way type embedding is used to distinguish image tokens, textual roles, and `[CLS]`. The Transformer encoder processes the full sequence jointly, and a linear head maps the `[CLS]` representation to a single logit representing image–triplet compatibility.

B.3 Hyperparameters

All probes are trained using the AdamW (Loshchilov et al., 2017) optimizer with a fixed learning rate, which we keep constant across benchmarks to ensure comparability. For Visual Genome, we use a smaller model dimension than for ImageNet100 and MSCOCO due to the increased input length from concatenating textual triplet embeddings and the substantially larger effective batch size required for contrastive training with multiple negatives, which together impose higher memory and compute demands. An overview of the hyperparameters is given in Table 6.



(a) rFID vs. NFE with $\mathrm{CFG} = 3.0$ (b) rFID vs. CFG with $\mathrm{NFE} = 10$ (c) Ablation of the bottleneck size

Figure 5: (a) and (b): Ablation of the reconstruction fidelity under different sampling hyperparameters. (c): Ablation of the bottleneck size with COMiT-B.

B.4 Reconstruction.

To select the optimal parameters for the NFE and CFG we start by fixing $\mathrm{CFG} = 3.0$ and sweep the NFE in $\{2, 4, 10, 25\}$ (see Figure 5a). We find $\mathrm{NFE} = 10$ to be the optimal number of evaluation steps across all model sizes. We observed that increasing NFE typically leads to sharper details and more salient artifacts, which slightly increases rFID. We then fix $\mathrm{NFE} = 10$ and sweep CFG in $\{1.5, 2.5, 3.0, 3.5, 4.5, 7.5, 10.5\}$ (see Figure 5b). We observe the rFID curves behave smoothly under the changes in CFG, with $\mathrm{CFG} = 7.5$ yielding the lowest rFID with COMiT-L and -XL. Therefore, for all our experiments we fix $\mathrm{NFE} = 10$ and $\mathrm{CFG} = 7.5$. Notably, scaling COMiT from B to L leads to significant drop in rFID, while further scaling from L to XL only moderately improves the reconstruction quality.

B.5 Bottleneck size.

We additionally ablate the bottleneck size of COMiT. We train 3 variants of COMiT-B with small (S), medium (M) and large (L) bottleneck sizes, where S corresponds to the message length of 64 tokens from a vocabulary of 1k unique entries, M stay for the message length of 128 tokens and 16k vocabulary size, while L is the default configuration in the main paper with 256 tokens in the messages and 64k tokens in the vocabulary. We then report the classification probing accuracies of these variants along with the reconstruction PSNRs (see Figure 5). We observe that the accuracy stays within the same range, while PSNR keeps increasing. This may indicate the potential to further improve the reconstruction fidelity of the model without losing the probing capabilities.

C Analysis of the Attention Maps

In this section, we extend the preliminary analysis of the tokens' attention maps demonstrated in the main paper.

Emergence of Objectness. Similarly to (Duggal et al., 2024), we study the emergence of object discovery via visualizing the attention maps of message tokens to the image tokens in the 24th layer of COMiT-XL during the decoding process. First, we visually investigate the attention maps and observe that COMiT's tokens tend to correspond to some semantically meaningful regions in images, such as objects or object parts (see Figure 6).

Quantifying Objectness. To quantify these findings, we follow the same procedure described in the main paper. In Figure 8, we visualize the attention maps with 10%, 20%, 30% and 40% thresholding along with the mean IoUs for the whole CSSD dataset (Yan et al., 2013). We found that with 30% thresholding one can obtain 0.58 mIoU, which is considered remarkable for a model that has never seen segmentation maps or object categories during training. The attention maps are extracted from layer 24 at denoising time $t = 0.1$ using COMiT-XL.

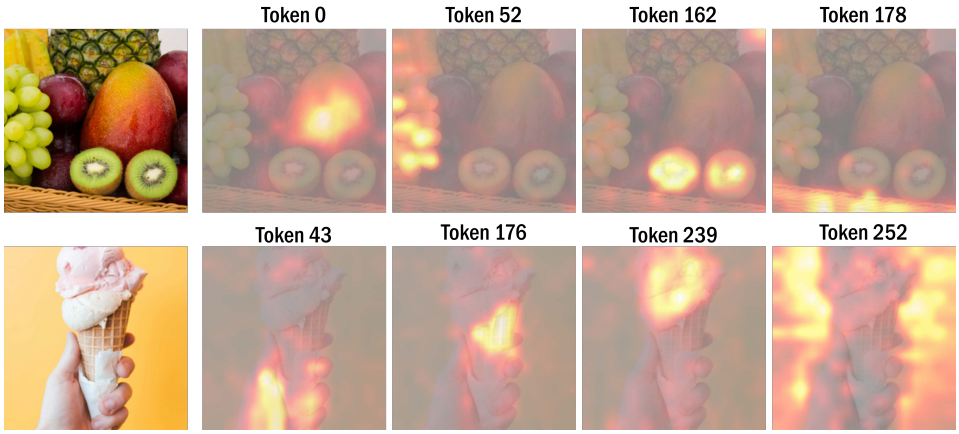


Figure 6: Emergence of object-centric tokens in COMiT. We visualize the attention maps of specific tokens in one of the deep layers of the network during the decoding stage. One can see that the tokens naturally attend to the objects and their parts.

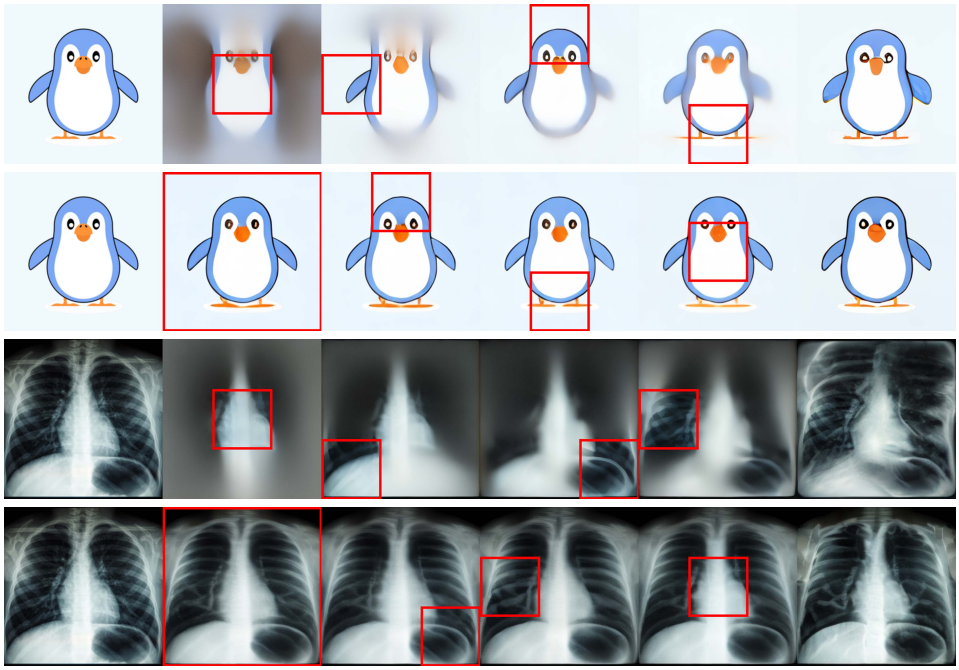


Figure 7: Generalization of COMiT to other domains, such as rendered animations or medical images. It can be seen that the model has certain symmetry bias that allows it to reduce the uncertainty about the right hand side of the image when the content on the left hand side is observed. Also notice how the adaptive cropping policy selects the most critical regions for reconstruction.

D Additional Ablations and Controls

This section documents the protocols and additional controls behind the ablations of Section 4.1 and the experiments in Section 4.2: the exact CSSD localization protocol, a non-attention readout, the robustness to truncated backpropagation, the effect of randomizing the number of crops during training, a cross-tokenizer token-localization study on multi-object COCO images, and the protocol of the segmentation probe.

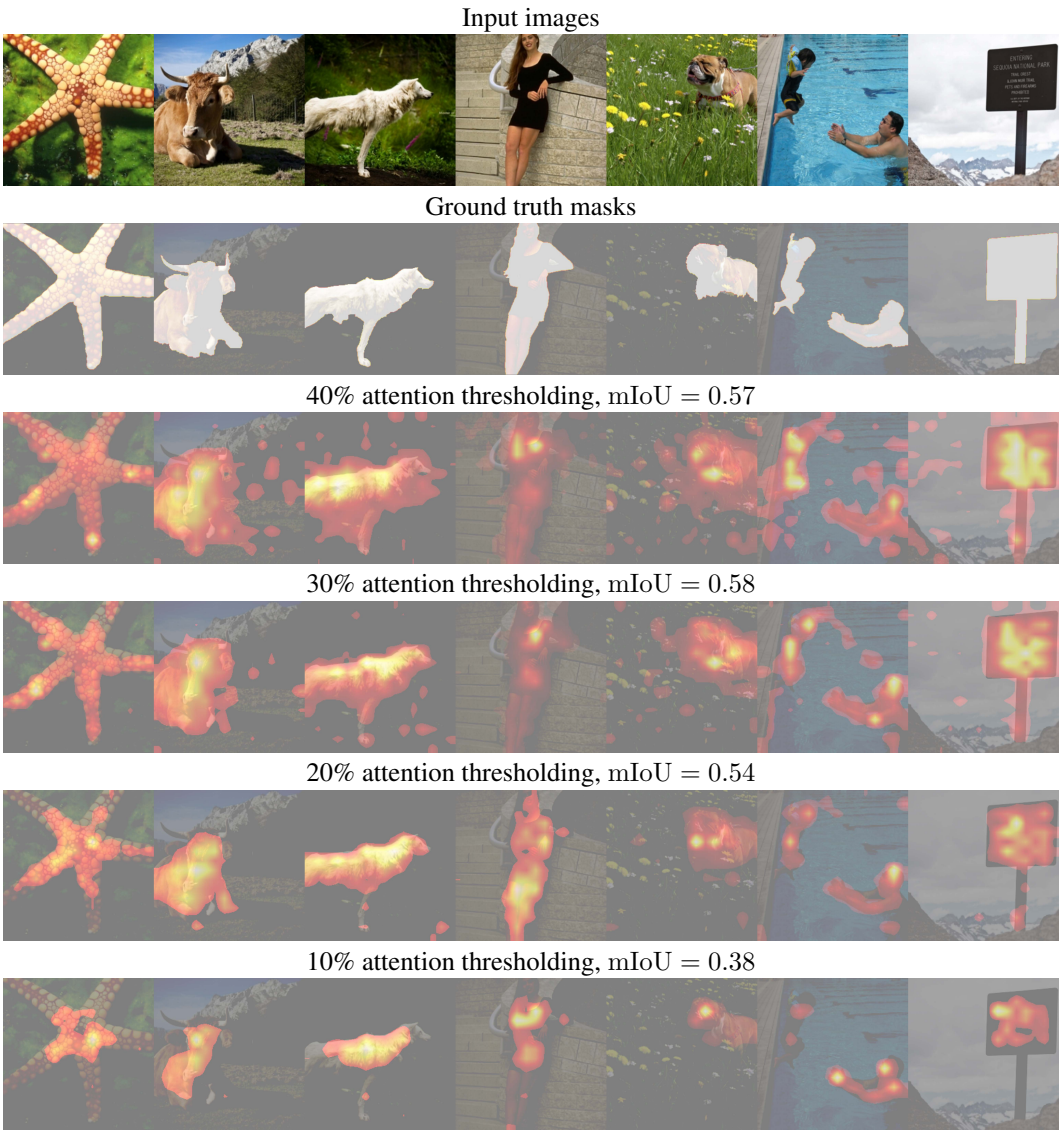


Figure 8: Visualization of COMiT-XL's attention maps. The attention maps of the tokens that yield the largest IoUs per sample are shown. With adjusting the thresholding percentage, a remarkable mIoU of 0.58 can be achieved. Note that the model has never seen any segmentation maps or class labels during training.

D.1 CSSD Token-Localization Protocol

All COMiT variants are pretrained on ImageNet1k only and kept frozen. CSSD (Yan et al., 2013) is used exclusively for evaluation: there is no training, fine-tuning or learned segmentation head, and all 200 images are used. Every variant encodes the image with the same single-global-crop policy. We then decode the message at denoising time $t = 0.1$ and extract the message-to-image attention maps of the 10th layer of COMiT-B (the 24th layer of COMiT-XL in Appendix C), averaged over heads. Each token's map is binarized by retaining its top $Q = 30\%$ locations, the IoU with the ground-truth mask is computed for every token separately, and the best token is selected per image. The mean of these best IoUs over the dataset is the reported mIoU. Because the ground-truth mask is used to select the token, the score is an oracle localization diagnostic of individual attention maps and not the output of a segmentation system, and it does not combine information across tokens through any learned readout. Table 7 extends Tables 1 and 2 from Section 4.1 with the corresponding 95%

image-bootstrap intervals, which resample the 200 images and hence quantify dataset uncertainty for these checkpoints.

Table 7: Three-way ablation of COMiT-B on IN100 (attention probe) and CSSD token localization. All variants report results for a single global crop at test time. The last column is the paired full-minus-variant difference in mIoU with its 95% image-bootstrap interval.

Training variant	IN100 top-1↑	CSSD mIoU↑	95% interval	Paired difference
Full	**82.91**	**0.527**	$[0.509, 0.546]$	–
No SREPA	72.26	0.507	$[0.490, 0.525]$	$+0.020$ $[+0.012, +0.028]$
No local crops	80.94	0.338	$[0.324, 0.352]$	$+0.189$ $[+0.175, +0.204]$

D.2 Affine Readout Controls

The attention probe of Appendix B is shared by all tokenizers in Table 4, so it does not give COMiT a method-specific advantage. As a complementary check that does not rely on an attention-based head, we also use the simplest position-sensitive readout: the frozen message is flattened in token order and mapped by a single affine layer to the class logits (153,700 parameters for a 256×6 message on IN100), with no attention and no nonlinear hidden layer. All affine probes use AdamW with a fixed learning rate of 0.01, no weight decay, batch size 1,024, at most 30 epochs with early stopping (patience 5) on the validation split, and five probe seeds. Each tokenizer is a single pretrained checkpoint, so the reported standard deviations cover probe initialization only. The readout is deliberately position-sensitive: mean pooling the six-dimensional token coordinates before the affine map removes nearly all accessible category information (6.42% IN100 top-1 for COMiT-B in a single run) and is therefore not an informative comparison.

We apply this protocol to COMiT and to prior tokenizers whose flattened message has the same dimensionality of 1,536 (Table 8). Such a position-specific readout has a different bias than the attention probe: it favors tokenizers that concentrate semantic information at fixed positions, as FlexTok does by construction. Indeed, all FlexTok variants exceed COMiT-B under this readout, whereas COMiT-L exceeds every evaluated baseline. The cross-model result at L scale therefore does not depend on the attention probe.

Table 8: Single affine readout over the flattened frozen message of tokenizers with a flattened message dimensionality of 1,536 (IN100 top-1, mean $\pm$ sample SD over five probe seeds; COMiT encodes a single global crop).

Tokenizer	Message shape	IN100 top-1↑
TiTok-S (128 tokens) (Yu et al., 2024a)	128×12	16.60 ± 0.14
FlexTok$_{\text{(d12-d12)}}$ (Bachmann et al., 2025)	256×6	72.07 ± 0.22
FlexTok$_{\text{(d18-d18)}}$ (Bachmann et al., 2025)	256×6	74.32 ± 0.14
FlexTok$_{\text{(d18-d28)}}$ (Bachmann et al., 2025)	256×6	72.76 ± 0.38
COMiT-B (ours)	256×6	71.66 ± 0.25
COMiT-L (ours)	256×6	$\mathbf{76.45 \pm 0.28}$

D.3 Robustness to Truncated Backpropagation

Section 3.1 stops gradients through all but the last message update. Table 9 evaluates an additional COMiT-B checkpoint under the CSSD protocol of Appendix D.1 that propagates gradients through up to the last four local-crop updates (depending on how many crops were sampled). Token–object localization persists, so the effect reported in Section 4.1 does not depend on the stop-gradient depth; the default is primarily a memory-saving choice. Since the two checkpoints are separate training runs with different seeds, the comparison serves as an indicator of robustness.

Table 9: CSSD token-localization mIoU of COMiT-B trained with the default stop-gradient heuristic and with gradients through up to four message updates (single global crop at test time; 95% image-bootstrap intervals).

Training setting	CSSD mIoU↑	95% interval
One backpropagated update (default)	0.527	$[0.509, 0.546]$
Up to four backpropagated updates	0.566	$[0.545, 0.587]$

D.4 Randomized Number of Crops

Section 3.1 argues that with a fixed number of crops during training the model would pre-allocate parts of the message to the crops it expects to follow, whereas a randomized number of crops encourages it to use the available tokens greedily. We test this with three COMiT-B checkpoints: the default model (randomized number of crops, one backpropagated update), the checkpoint of Appendix D.3 (randomized number of crops, up to four backpropagated updates), and a checkpoint trained with a fixed number of four local crops (one backpropagated update).

Token usage. Each checkpoint tokenizes the ImageNet100 validation set and a class-balanced subset of the training set (200 images per class), once with the global crop alone and once with the nine local crops of Section 3.5 in raster-scan order, and the message is recorded after every crop. To measure how much image-specific information a message position i carries, we compute

$$H_i = -\sum_{j=1}^{6}\sum_{l=1}^{L_j} p_{i,j}(l)\log_2 p_{i,j}(l), \tag{7}$$

where $L_j \in \{8, 5\}$ is the number of FSQ levels of coordinate j (Table 5) and $p_{i,j}(l)$ is the fraction of the 5,000 validation images in which coordinate j of token i takes level l. H_i is 0 if the position holds the same code in every image and at most $\sum_j \log_2 L_j = 15.97$ bits. Since the joint entropy of a code cannot exceed the sum of its coordinate entropies, a position with $H_i < 8$ bits behaves as if it used fewer than $2^8 = 256$ equally frequent codes out of the 64,000 codes of the vocabulary; we call such positions under-used. Table 10 shows the pre-allocation described in Section 3.1. After its first local crop, the model trained with a fixed number of crops leaves 36.7% of the message positions under-used (37.1% with a random crop order) and fills them during the next two crops (3.5% after the third), whereas the randomized models leave at most 5.9% of the positions under-used after any number of crops. After the global crop, none of the models leaves more than 1% of the positions under-used. These differences do not depend on the 8-bit threshold: after the first local crop, the fixed-K model has 19.1% and 50.0% of its positions below 4 and 12 bits, respectively, compared with at most 3.5% and 12.1% for the randomized models.

Table 10: Fraction of under-used message positions (%, $H_i < 8$ bits) after the global crop or after the first local crops (raster-scan order unless stated otherwise).

Message after	Randomized number of crops: one update	Randomized number of crops: up to four updates	Fixed number of crops: one update
Global crop	0.0	0.4	0.8
1 local crop	4.7	5.5	36.7
2 local crops	4.7	3.5	9.4
3 local crops	5.1	3.5	3.5
4 local crops	3.5	4.7	3.5
1 local crop, random order	4.7	5.9	37.1

Semantic probing. The affine readout of Appendix D.2 (five seeds) and a two-layer attention probe with the architecture of Appendix B (three seeds) are trained on the training subset and evaluated on the full validation set. Because of the smaller training set, absolute accuracies are lower than, and not directly comparable to, those of Table 1. Table 11 shows that the randomized models also make class information more accessible. Both exceed the fixed-K model with the global crop and after every number of local crops. The gap persists at four crops, the training length of the fixed-K model, and after the global crop, where the fixed-K model leaves fewer than 1% of the positions

under-used, so the benefit of randomization for probing goes beyond the greedy use of the first crops. The randomized number of crops is the only setting in which both randomized checkpoints differ from the fixed-K one, which makes the randomization critical for both token usage and probing.

Table 11: ImageNet100 top-1 accuracy (%) of probes trained on 200 images per class, mean ± std over probe seeds. R1 and R4: randomized number of crops with one and up to four backpropagated updates; F1: fixed number of four crops with one backpropagated update.

	Affine readout (5 seeds)			Attention probe (3 seeds)		
Message after	R1	R4	F1	R1	R4	F1
Global crop	56.08 ±0.28	63.41 ±0.35	46.78 ±0.46	74.71 ±0.54	77.09 ±0.50	65.40 ±1.68
1 local crop	26.44 ±0.43	29.49 ±0.37	16.90 ±0.15	27.55 ±1.50	30.56 ±0.34	18.53 ±0.75
2 local crops	41.54 ±0.42	46.12 ±0.32	29.64 ±0.36	46.93 ±1.15	49.27 ±0.48	36.54 ±1.52
4 local crops	52.69 ±0.36	58.91 ±0.36	41.97 ±0.49	59.93 ±0.69	63.43 ±0.21	51.81 ±0.39
9 local crops	64.66 ±0.31	71.08 ±0.23	51.90 ±0.41	72.07 ±1.55	74.08 ±0.25	61.79 ±0.78

D.5 Cross-Tokenizer Token Localization on COCO

CSSD contains single salient objects and the analysis above compares COMiT variants only. To test whether individual tokens localize objects in multi-object images, and how this compares to prior tokenizers, we evaluate on a fixed random subset of 150 COCO 2017 validation images (Lin et al., 2014). We compare only tokenizers that produce 256 tokens per image: since the score below takes the maximum over individual tokens, it grows with the number of candidate tokens and would favor tokenizers with longer sequences. We restrict the evaluation to large objects, i.e., instances whose original COCO annotation area is at least 96^2 pixels; crowd annotations are excluded, and an instance is retained only if its mask remains non-empty after square center-cropping and resizing the image to 256×256. This yields 230 objects in 119 images. The restriction follows from how a single token map is scored. A map that retains the top $Q\%$ of the pixels always covers $Q\%$ of the image, so its IoU with an object that covers a fraction f of the image is at most $\min(Q, f)/\max(Q, f)$, and an IoU above 0.5 is only possible if $Q/2 < f < 2Q$. Medium objects ($32^2 \leq$ area $< 96^2$) cover a median of 1.33% of the resized image, and 90% of them less than 3.54%, so only maps that keep a few cells of the 16×16 attention grid could localize them. Accordingly, every evaluated model localizes fewer than 10% of the medium objects at the IoU threshold of 0.5, and this setting does not discriminate between tokenizers. Large objects cover a median of 12.73% of the image.

Protocol. Every tokenizer completes its full 20-step decoding trajectory from its token sequence, and attention maps are captured at the same clean-progress value of 0.1 for all models (native $t = 0.1$ for COMiT and native noise time 0.9 for the rectified-flow decoder of FlexTok) rather than at unmatched native timesteps; the decoding noise is fixed per image. As for CSSD, the maps are the raw decoder spatial-query/token-key attention logits on the 16×16 spatial grid, softmaxed over spatial positions separately for every token, averaged over heads and bilinearly resized to 256×256. A token map is binarized by retaining its top $Q\%$ of pixels. For each ground-truth object o with mask S_o we compute

$$\mathrm{IoU}_o = \max_i \mathrm{IoU}(M_i, S_o), \tag{8}$$

where M_i is the binary attention map of the individual token i, and count the object as localized when $\mathrm{IoU}_o > p$. This evaluation uses no learned readout and never combines several token maps. Based on the size analysis above, we fix $Q = 10\%$ for all tokenizers: a map that retains 10% of the pixels can exceed an IoU of 0.5 for objects that cover between 5% and 20% of the image, a range that contains the median large object, and a common Q makes every tokenizer target the same object sizes. The extraction layer is calibrated independently for each tokenizer, on a disjoint set of 13 images containing 25 large objects, by sweeping every decoder layer and maximizing the localization rate at the primary threshold $p = 0.5$ (ties are broken by the mean best-token IoU, then by earlier layers). This selected the zero-based layers 1, 4, 4, 4, 12 and 15 for COMiT-B, its variants without SREPA and without local crops, COMiT-L, COMiT-XL and FlexTok, respectively.

Results. Table 12 summarizes the evaluation. COMiT-L localizes 14.78 points more large objects than FlexTok-d18-d18 (Bachmann et al., 2025) (paired bootstrap 95% interval $[+9.33, +20.38]$), has a higher mean best-token IoU (0.433 versus 0.386), and remains ahead at all three thresholds. COMiT

Table 12: Individual-token localization of the 230 large objects in the COCO subset. An object is localized if the best of the individual binarized token attention maps has IoU $> p$ with its mask. Intervals are 95% bootstrap intervals of the rate at $p = 0.5$. The last two columns count how many distinct token indices supply the best match over all 230 objects, and the normalized entropy of this distribution. All tokenizers produce 256 tokens per image.

Tokenizer	Localized at $p = 0.5$ rate↑	95% interval	Mean best IoU↑	Localized at $p = 0.25$	$p = 0.75$	Distinct best tokens	Norm. entropy
FlexTok$_{\text{(d18-d18)}}$	21.74	$[16.39, 27.44]$	0.386	79.13	**2.17**	59 / 256	0.662
COMiT-B, no local crops	17.39	$[12.86, 22.22]$	0.358	74.78	1.30	127 / 256	0.843
COMiT-B, no SREPA	23.48	$[18.14, 29.17]$	0.386	80.87	0.87	117 / 256	0.818
COMiT-B (ours)	25.65	$[19.72, 31.69]$	0.406	**83.91**	0.43	134 / 256	0.854
COMiT-L (ours)	**36.52**	$[30.13, 43.22]$	**0.433**	**84.35**	**3.48**	126 / 256	0.836
COMiT-XL (ours)	**34.78**	$[28.71, 41.07]$	**0.423**	**83.91**	1.30	125 / 256	0.836

also distributes the best matches over many more token indices than FlexTok (126 versus 59 of 256 tokens; normalized entropy 0.836 versus 0.662), i.e., localization is not carried by a few dedicated slots. The distribution statistics alone do not establish semantic disentanglement, but together with the higher per-object localization rates they support the intended effect of recurrent aggregation: spatial information is organized into more broadly utilized, object-aligned tokens. Within the B scale, recurrent local-crop training raises the localization rate from 17.39% to 25.65% (paired interval $[+1.42, +14.83]$) and the mean best-token IoU from 0.358 to 0.406 under the common mask size, so that COMiT-B numerically surpasses the larger FlexTok-d18-d18 (25.65% versus 21.74%), whereas its no-local variant does not. The no-SREPA variant of COMiT-B (Table 1) localizes 23.48% of the objects with a mean best-token IoU of 0.386. Its paired difference to the full model (-2.17 points, interval $[-8.30, +3.78]$; mean best-token IoU -0.020, interval $[-0.032, -0.008]$) is much smaller than that of the no-local variant, which it exceeds by 6.09 points (interval $[+0.47, +11.62]$). This variant is also the one whose semantic supervision matches that of FlexTok: both are trained with standard REPA on DINOv2 features and without aggregate message-level alignment. Under this matched supervision, COMiT-B without SREPA localizes large objects as often as FlexTok-d18-d18 (23.48% versus 21.74%; paired difference $+1.74$ points, interval $[-3.90, +7.23]$) with the same mean best-token IoU (0.386), although it has about a fifth of FlexTok's parameters (174M versus 860M, Table 4), and it spreads the best matches over twice as many token indices (117 versus 59 of 256; normalized entropy 0.818 versus 0.662). Token-level localization in COMiT thus does not rely on SREPA. The multi-object diagnostic thus reproduces the CSSD results of Section 4.1: removing SREPA barely changes token–object localization, whereas removing recurrent local crops degrades it substantially.

Taken together, the probes of this paper address distinct levels of the claim: the shared sequence probes of Table 4 and Appendix D.2 measure the accessibility of semantic information from the complete message, the segmentation probe below measures the dense spatial information available to a common decoder, and CSSD together with this analysis measure the alignment of individual tokens with objects.

D.6 Segmentation Probe

The segmentation probe of Section 4.2 evaluates COMiT-L and SelfTok-512 on the pair-disjoint MSCOCO subset of Appendix B. The target for each image is a binary foreground mask formed as the union of all COCO instance masks; images and masks are resized and center-cropped to 256×256, and the targets are downsampled to the output resolution of the probe. COMiT-L encodes only the global crop, without test-time aggregation over local crops. Both tokenizers are frozen and read out by the same lightweight decoder: the tokens are normalized and linearly projected to 128 dimensions, 256 learned spatial queries attend to them through two cross-attention blocks with four heads each, and the queries are arranged into a 16×16 feature map that is decoded to mask logits by an upsampling convolutional stem. Both probes use the same optimization protocol: AdamW for at most 20 epochs, batch size 256, learning rate 1e-3, weight decay 1e-4, early stopping with a patience of five epochs, and a binary cross-entropy plus Dice loss. At a prediction threshold of 0.5, COMiT-L obtains a foreground IoU of 0.532, compared with 0.436 for SelfTok-512, despite SelfTok's much

stronger reconstruction. Because the spatial queries can attend across all tokens and the target merges instances, this probe measures the spatial information retained by the complete token sequence; it is complementary to the individual-token analyses of Appendices D.1 and D.5 and is not evidence for a one-token-one-instance decomposition. However, because the probe is trained and evaluated on the pair-disjoint MSCOCO split, it also tests compositional generalization.

E Extended Related Work

Here we expand the related work of Section 2 with recent and concurrent tokenizers. GigaTok (Xiong et al., 2025) scales 1D tokenizers to billions of parameters and regularizes their latents with DINOv2 features to resolve the reconstruction–generation dilemma of large tokenizers; its contribution is scale rather than the organization of the token sequence. VFMTok (Zheng et al., 2025) starts from the frozen spatial features of a vision foundation model and adaptively resamples them into tokens, whereas COMiT learns its discrete message end-to-end from pixels. SemTok (Qu et al., 2026) learns 1D tokens with a single-pass encoder, SigLIP-based semantic alignment and a large pretrained decoder that is trained in two stages, namely generation pretraining followed by reconstruction refinement. CaTok (Chen et al., 2026) learns causal register tokens whose intervals are tied to MeanFlow time intervals from a single global pass over the image, and its main representation is continuous. EvoTok (Li et al., 2026) evolves residual latents over a 2D token grid, with deeper residual levels becoming more semantic, and targets a unified understanding-and-generation model trained on much larger data. VibeToken (Patel et al., 2026) scales 1D tokenizers and autoregressive models to dynamic resolutions and aspect ratios without a semantic teacher. PyraTok (Susladkar et al., 2026) and PVQ-VAE (Prakash et al., 2026) are video tokenizers built on pyramidal representations.

These works span 1D and 2D codes, discrete and continuous representations, image and video modalities, and ImageNet-scale versus web-scale training. What distinguishes COMiT is the mechanism it studies: recurrent partial observations that repeatedly rewrite a fixed-length discrete message, together with the frozen-token evaluation of the resulting token organization.

F More Visual Results

Here we provide more qualitative results with COMiT that could not be included in the main paper due to the pages number limit. If not stated otherwise, the input images in the main paper and in this appendix are random images from `unsplash.com`. We observed that despite being trained only on ImageNet1k Deng et al. (2009), COMiT demonstrates remarkable generalization capabilities and remains robust under domain shifts (see Figure 7).

Nearest Neighbors. To study the learned space of latent messages, we visualize some nearest neighbors of sample images from the validation set of ImageNet100 (Deng et al., 2009). For simplicity, we concatenate the tokens in latent messages to obtain a single 1536-dimensional vector per image. For each query image we then find 4 closest samples among the whole validation set using the cosine distance as our similarity measure (see Figure 9). Despite the simplicity of this probing mechanism, we find that the latent space is well–structured and the neighboring images typically share semantics.

Cropping policies. Figure 3 demonstrates the difference between adaptive cropping policies in the absence and the presence of the global crop. When the global crop is leveraged, it provides a way to quickly embed the whole scene into the latent message with the subsequent local crops allowing to focus on the finer details (such as the shape and the color of the feathers near the bird's head). Figure 10 contains more demonstrations on the difference between the presence and absence of the global crop in the adaptive cropping policy. Notice that the content that the model does not observe in the sequence of crops misses from the final reconstruction. However, 3-4 local crops in the adaptive policy are typically enough to grasp the overall idea behind the scene without including too much details. This explains, why the adaptive policy with only 3 crops performs quite well on our semantic probing test suite.

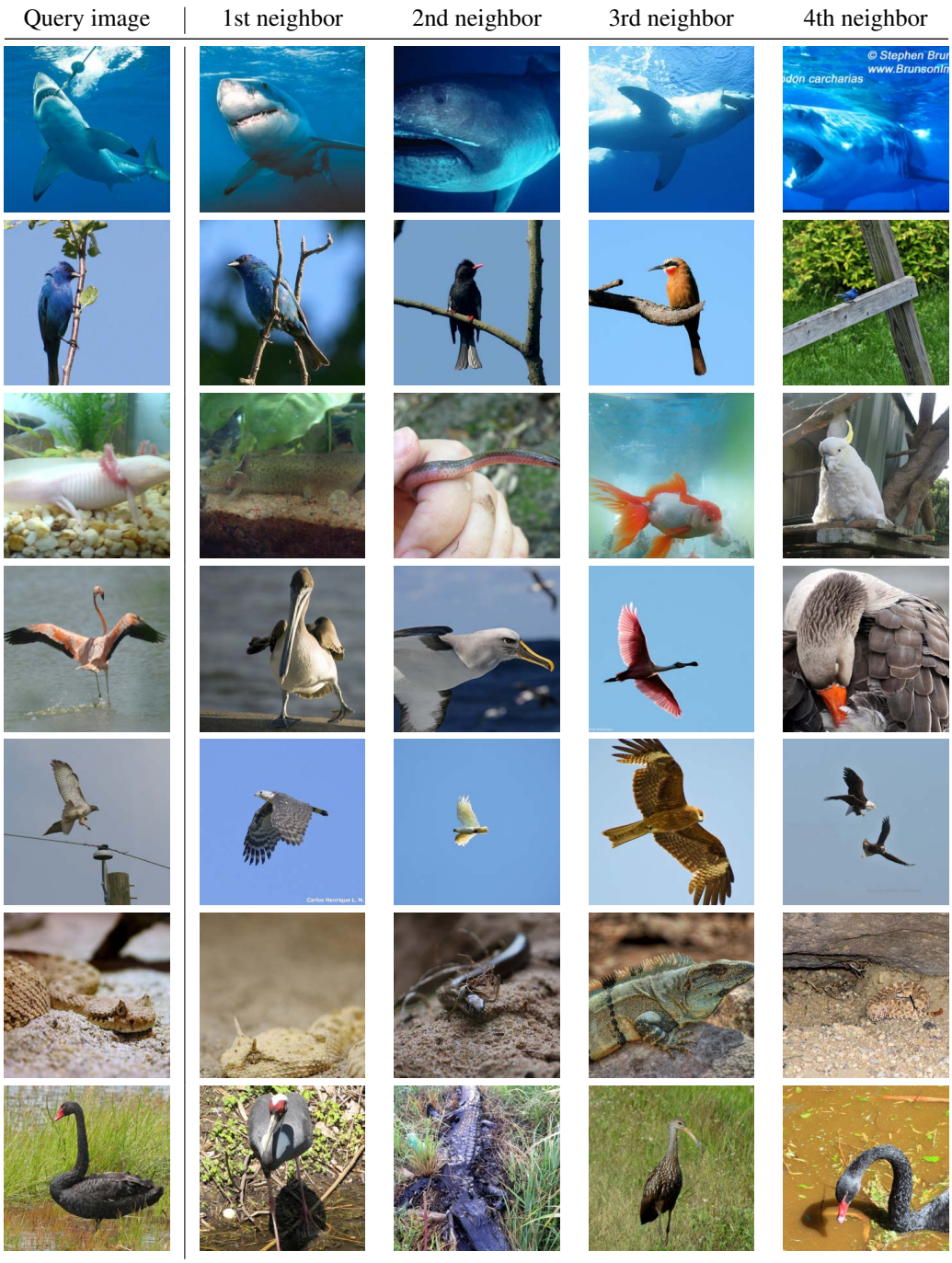


Figure 9: Visualization of the nearest neighbors of a sample from the ImageNet100 validation set in the space of COMiT's latent messages. One can see that despite unsupervised training and simple probing the learned messages tend to cluster into semantically meaningful groups.



Figure 10: Additional results on the difference between the *adaptive* (odd rows) and *global+adaptive* (even rows) cropping policies. The first column depicts the ground truth input image. The last column corresponds to the final reconstruction with 10 NFE. The columns in-between demonstrate how COMiT refines its latent message with incoming information (1 NFE reconstructions are shown).

G Future Work

Going forward, an important direction is to improve reconstruction fidelity without sacrificing semantic structure, making COMiT more suitable for generation-focused applications. More broadly,

structured token sequences could provide a stronger interface for multimodal models, especially in settings that require object-centric reasoning, compositional understanding, and dense visual grounding. The test-time flexibility of COMiT also suggests opportunities for adaptive tokenization, where cropping policies are selected dynamically for the task.

Future work will extend COMiT to video, where temporal redundancy makes compact and structured representations especially valuable. By accumulating spatiotemporal observations over time, COMiT could learn token sequences that capture motion, actions, and long-range temporal structure, enabling more efficient and semantically grounded video understanding.

H Impact Statement

This paper advances research in autoencoders, representation learning and generative modeling. As with other generative image models and tokenizers, the methods studied here may carry standard societal and ethical risks. These include the potential misuse of generated content for creating misleading or deceptive media, the amplification of biases present in training data, and the environmental costs associated with training and deploying large-scale models. While this work is primarily methodological, we acknowledge these well-documented considerations and emphasize that responsible dataset curation, evaluation, and deployment practices are important to mitigate such risks.