Images are Worth Variable Numbers of Tokens

Anonymous Author(s)
Affiliation
Address
email

Abstract

Most existing vision encoders map images into a fixed-length sequence of tokens, overlooking the fact that different images contain varying amounts of information. For example, a visually complex image (e.g., a cluttered room) inherently carries more information and thus deserves more tokens than a simple image (e.g., a blank wall). To address this inefficiency, we propose DOVE, a dynamic vision encoder that produces a variable number of tokens to reconstruct each image. Our results show that DOVE significantly reduces the average number of tokens while maintaining high reconstruction quality. In several linear probing and downstream multimodal tasks, it outperforms existing autoencoder-based tokenization methods when using far fewer tokens, capturing more expressive semantic features compared to fixed-length encoding. We further extend DOVE with query-conditioned tokenization. By guiding the model to focus on query-relevant regions, it achieves more efficient and targeted semantic extraction.

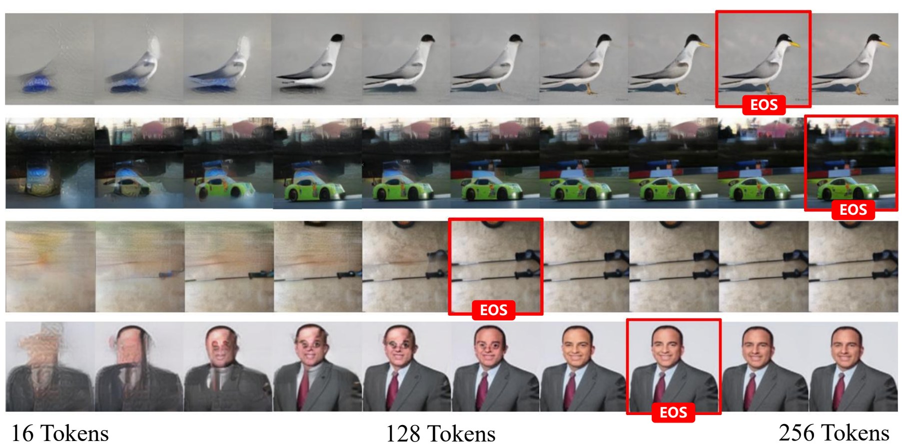


Figure 1: **Dynamic Visual Representations.** As the number of tokens used by DOVE increases, the reconstructed images shows finer and high frequency details.

1 Introduction

Image representation learning [56] is a fundamental component of computer vision; it plays a pivotal role in various visual tasks, including image classification [38, 12], object detection [62, 61], and semantic segmentation [26, 27]. Vision representation models are also widely used in multi-modal learning, where they serve as powerful vision encoders within vision-language models (VLMs), converting image information into discrete token sequences. Existing image representation learning methods generally fall into two categories: semantic feature learning (e.g., CLIP [46], DINO [10]) and autoencoder-based image tokenization (e.g., VQGAN [21], VAE [31]). All of which aim to generate

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

fixed length sequences. However, studies have shown that vision tokens suffer from information redundancy [11]. We conjecture that different images have different complexity such that they can be represented with different lengths of tokens for reconstruction.

To this end, we propose DOVE (Dynamic Output Vision Encoder), a visual tokenizer that adaptively generates variable-length vision token sequences for image reconstruction. Our method extends the standard visual autoencoder framework by incorporating a transformer-based dynamic token generator (Figure 2), which is capable of generating an end-of-sequence (EOS) token at any position to terminate the output sequence. We jointly optimize image reconstruction quality and EOS token prediction based on an MSE threshold, and truncate token sequences at the predicted EOS. Our method effectively shortens the token sequence length while maintaining high reconstruction quality (Figure 1). As token sequences progress, their reconstructions show more high-frequency details and additions of objects, and then saturate at (EOS) token.

By learning dynamic token lengths, we find that the tokenizer learns richer semantics and observe the emergence of zero-shot semantic segmentation by PCA on the hidden features. We perform extensive experiments on reconstruction, classification, and question answering by replacing vision backbones in vision language models. Our approach consistently and significantly outperforms other autoencoder-based tokenization methods while enjoying improved efficiency from dynamic length.

Considering that human vision is an active and task-driven process, and that humans tend to focus on task-relevant regions while ignoring irrelevant ones when answering questions [4, 35, 17], we additionally introduce a query-conditioned variant of DOVE. This model is able to read the user's query and reconstruct the input by focusing on semantically relevant regions, thereby further reducing the length of the generated token sequence. In practice, given a text query and a corresponding salient image region during training, we feed the text query to the token generator and apply higher weights to the reconstruction loss specifically corresponding to the salient region. We find that this approach further improves token efficiency, semantics, and vision language model performance.

We summarize our contributions as follows:

- We propose DOVE, a visual tokenizer that dynamically generates tokens based on image complexity. Unlike previous visual tokenization, our model supports arbitrary control over the token sequence length in a single parallel forward.
- We propose a variant of DOVE that grounds token generation on a text query and its corresponding salient visual regions. This query-conditioned model achieves a higher token compression rate (averaging 68%) and demonstrates stronger semantic representation.
- We observe a phenomenon of emergent semantics by probing the latent representation. Compared to other autoencoder-based tokenization methods with fixed-length token representations, our model achieves significantly better performance on classification, vision-language QA, and shows emerging semantic segmentation properties.

2 Dynamic Vision Tokenizer

We introduce DOVE, a dynamic vision encoder that adaptively generates a variable number of tokens to reconstruct each image.

2.1 Model Architecture

An overview of our model is shown in Figure 2. Our model consists of four main components: VQGAN Encoder, VQGAN Decoder, transformer-based dynamic token generator, and transformer-based token decoder. We use 70M transformer [7] as the backbone for both the autoregressive token generator and a non-autoregressive version for token decoder.

For each image X_v, the VQGAN Encoder converts the visual information into a fixed-length token sequence H_v. Timestamp encodings $t_1, t_2, \ldots, t_n$, generated using periodic embeddings such as sinusoidal encodings [55], are then appended to H_v. This combined sequence is input into the dynamic token generator f_ϕ. To enable sequential token generation, we restrict each position to attend only to its current or preceding timestamps. The dynamic token generation process from timestamp t_0 to t_i is defined as:

$$D = f_\phi(H_v, t_1, t_2, \ldots, t_i) = (d_1, d_2, \ldots, d_i) \tag{1}$$

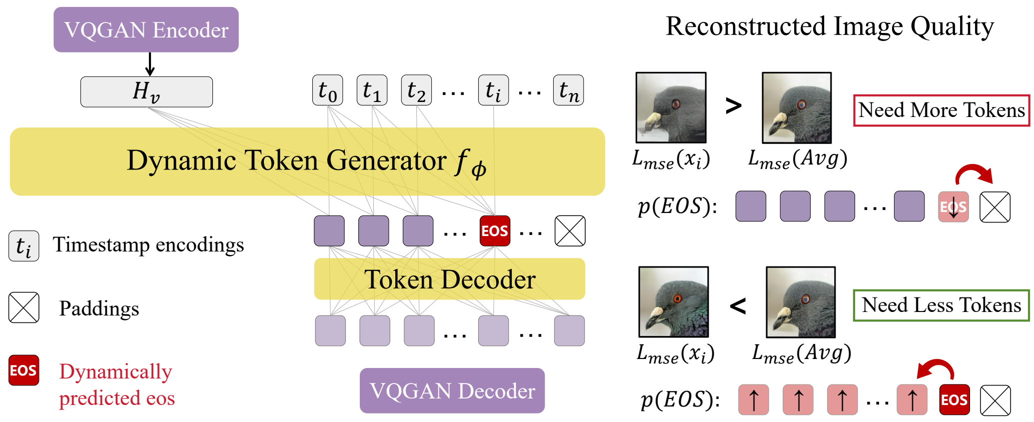


Figure 2: **Dynamic Tokenizer.**

where D denotes the generated token sequence, and d_i is the token produced by the model at t_i. We introduce dynamic length variation by detecting the EOS token from the model's discrete output and replacing all vision tokens (latent outputs) from that position onward with zero vectors. Since the EOS token can appear at any position, the length of the generated token sequence can vary based on the complexity of the image. We use an additional non-autoregressive token decoder g_ϕ to decode the padded dynamic vision token sequence and feed it to the final VQGAN decoder.

2.2 Dynamic Image Reconstruction

A more complex image, which contains richer and finer-grained details, will require more tokens to capture all its visual information compared to a simpler one. By learning when to generate EOS, the model can adaptively produce a token sequence that is just long enough to capture the image's essential visual content.

We jointly train all components of the model. Following the training strategy of VQGAN [21], we adopt a combination of mean squared error (MSE) loss and perceptual loss to supervise the image reconstruction process. A lightly weighted adversarial (GAN) loss is also applied to enhance the realism of reconstructed images. The final reconstruction loss L_{rec} between the input image X_v and the reconstructed image $\hat{X}_v$ is defined as:

$$L_{\text{rec}} = \lambda_{\text{mse}} \cdot L_{\text{mse}} + \lambda_{\text{perc}} \cdot L_{\text{perc}} + \lambda_{\text{gan}} \cdot L_{\text{gan}} \quad (2)$$

During training, we set the weighting factors to $\lambda_{\text{mse}} = 1$, $\lambda_{\text{perc}} = 0.1$, and $\lambda_{\text{gan}} = 5 \times 10^{-10}$ to prevent hallucination. In parallel with improving reconstruction quality, we guide the model to adaptively adjust the length of the generated token sequence through EOS prediction. Specifically, we use the average reconstruction loss L_{rec} over the previous 100 training steps as a dynamic threshold. For a given sample, if its current reconstruction loss is lower than the threshold, it indicates that fewer tokens are sufficient for satisfactory reconstruction, and we encourage earlier EOS prediction by maximizing the EOS probabilities at all preceding positions. Conversely, if the reconstruction loss exceeds the threshold, it suggests that more tokens are needed, and we minimize the EOS probability at the current position.

Define: Image X_v, max tokens K, window W, weights λ_{rec}, λ_{eos}, time encodings T

$H_v \leftarrow$ VQGAN_Encoder(X)
Initialize $\text{EMA}_{\text{rec}} \leftarrow 0$
for each training iteration **do**
  $D \leftarrow [\,]$, $i \leftarrow 1$
  while $i \leq K$ **do**
    $d_i \leftarrow f_\phi(H_v, T_{1:i})$ *(generating token)*
    append d_i to D, $i \leftarrow i + 1$
  Find the first index j such that $D[j] = \text{EOS}$
  if such j exists **then**
    for $k = j + 1$ to K **do**
      $D[k] \leftarrow 0$
  $\hat{X} \leftarrow$ VQGAN_Decoder$(g_\phi(D))$
  Compute L_{rec} via Eq. (2)
  Update EMA_{rec} over the last W losses
  if $L_{\text{rec}} > \text{EMA}_{\text{rec}}$ **then**
    $L_{\text{eos}} \leftarrow p_{\text{eos}}(i)$
  else
    $L_{\text{eos}} \leftarrow -\frac{1}{i-1} \sum_{j=1}^{i-1} p_{\text{eos}}(j)$
  $L_{\text{total}} \leftarrow \lambda_{\text{rec}} L_{\text{rec}} + \lambda_{\text{eos}} L_{\text{eos}}$
  Update parameters ϕ using $\nabla_\phi L_{\text{total}}$

Table 1: Training Pseudocode

We denote the predicted EOS probability at position i as $p_{\text{eos}}(i)$, where m indicates the current EOS position. The token length control loss is defined as:

$$L_{\text{eos}} = \begin{cases} p_{\text{eos}}(m), & \text{if } L_{\text{rec}} > \text{Threshold} \\ -\dfrac{1}{m-1}\sum_{i=1}^{m-1} p_{\text{eos}}(i), & \text{if } L_{\text{rec}} \leq \text{Threshold} \end{cases} \tag{3}$$

Finally, we jointly optimize L_{rec} and L_{eos} to guide the model in dynamically reconstructing the image. The overall training loss is defined as:

$$L_{\text{total}} = \lambda_{\text{rec}} L_{\text{rec}} + \lambda_{\text{eos}} L_{\text{eos}} \tag{4}$$

where λ_{rec} and λ_{eos} are the corresponding weighting coefficients. To facilitate faster convergence, we initially set λ_{eos} to a small value and gradually increase it during training, allowing the model to first focus on accurate reconstruction before learning to adaptively control the token sequence length.

2.3 Q-DOVE: Query-conditioned Tokenization

We extend DOVE to Q-DOVE for use in text-conditioned vision and language domains (Figure 3), allowing it to dynamically adapt image representations in a query-dependent manner. Q-DOVE is trained to focus image representation resources on image regions relevant to a given query.

Given a supervised dataset of images paired with text queries and bounding boxes encapsulating their answers, we modify the reconstruction loss to focus over image regions within each example's set of bounding boxes S_{bb}. Specifically, we upsample each image region contained by a bounding box $b^i \in S_{bb}$ to an image I^i_{bb} and compute the reconstruction loss over it as in Eq. 2:

$$L^i_{\text{rel}} = L_{\text{rec}}(I^i_{bb}) \tag{5}$$

In order to encourage the model to maintain some fidelity over the region outside of the bounding boxes, we also compute the MSE loss over I_o, the complement of S_{bb}:

$$L_{\text{irr}} = L_{\text{mse}}(I_o) \tag{6}$$

The final loss averages over relevant regions and weighs loss over the irrelevant region down by λ_o:

$$L_{\text{qry}} = \frac{\sum_{b^i \in S_{bb}} L^i_{\text{rel}}}{|S_{bb}|} + \lambda_o \cdot L_{\text{irr}} \tag{7}$$

In our experiments, we set λ_o to 1e-10. To compute L_{eos}, we employ the same procedure as in Eq. 3, comparing L_{rel} to a threshold determined by its average loss over previous training steps. If L_{irr} falls below the threshold, we introduce an additional penalty L_{pen} to explicitly encourage the model to generate the EOS token earlier. $L_{\text{pen}} = -\dfrac{1}{m-1}\sum_{i=1}^{m-1} p_{\text{eos}}(i)$

Our supervised masking strategy yields a dual benefit, allowing the model to learn both where to look and how much information to encode from image regions relevant to inputted queries. Bounding boxes are only used during training.

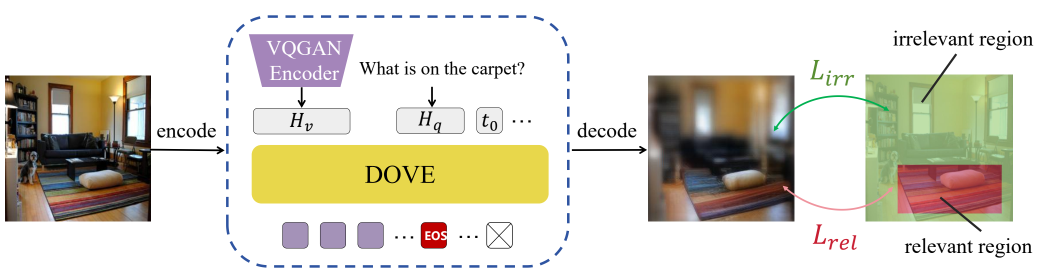


Figure 3: **Query Conditioning.** DOVE is trained with a bounding-box based loss, learning to focus its dynamic token resources on representing query-relevant image regions.

3 Experiments

In this section, We evaluate our approach at multiple levels, including the quality of the generated vision tokens (e.g., image reconstruction and token length distribution), as well as their effectiveness in downstream vision-language tasks. The results demonstrate that our model achieves high reconstruction quality with significantly fewer tokens, while capturing richer semantic information compared to static autoencoder-based tokenization methods. We further investigate the phenomenon of emergent semantics in Section 3.4.

3.1 Experimental Setup

Training Details. We use a pretrained VQGAN [21] with a codebook size of 8192 and a lightweight Pythia-70M [7] language model as the backbone of our framework. The model is fine-tuned on ImageNet-1K [16] for 20 epochs using two NVIDIA RTX 4090 GPUs. For the query-conditioned variant, we conduct an additional 5 epochs of training on the Visual Genome [32] and Open Images [34] datasets. We directly use the provided questions and region-level captions in Visual Genome as textual queries to guide the model in reconstructing content within specified bounding boxes, while ignoring irrelevant regions. Since Open Images does not offer region-level descriptions or questions, we instead construct text queries from relation graph annotations—for example, "a cup on a table"—and define the target region by concatenating the bounding boxes of the associated objects. To improve the model's generalization ability, we randomly replace 50% of the training text queries with the string "null", and train the model to reconstruct the entire image when this placeholder is provided as input.

Baselines. We compare our model against several state-of-the-art encoder-decoder frameworks, including TiTok[60] and VQGAN. We choose VQGAN with an output length of 256 tokens. For TiTok, we consider three variants with token lengths of 32, 64, and 128. We also include ALIT [20], a dynamic vision encoder trained via recurrent distillation from VQGAN. Unlike our method, however, ALIT only supports token lengths that are multiples of a fixed stride (e.g., 32). All models are trained on ImageNet-1K under the same configuration to ensure a fair comparison.

3.2 Token-Level Evaluation

Image Reconstruction Quality. We report FID scores of the reconstructed images across varying token lengths. Our results show that as the token length increases, the reconstruction quality of our model consistently improves. At all evaluated token lengths, our method outperforms ALIT. This advantage becomes especially clear at lower token counts. ALIT often generates hallucinated content, including severe object distortions. For example, when the token length is limited to 32, the reconstructed chameleon and beetle exhibit noticeable deformations (Figure 4). In contrast, our model produces slightly blurry but structurally and semantically faithful reconstructions. When using the full token length of 256, our method surpasses VQGAN on the COCO and WIT datasets. Detailed results are provided in Table 2.

Approach	ImageNet100								COCO			Wikipedia (WIT)		
	32	64	96	128	160	192	224	256	32$^{\#}$ / 64	128	256	32$^{\#}$ / 64	128	256
TiTok-L-32	11.60	-	-	-	-	-	-	-	14.18$^{\#}$	-	-	53.57$^{\#}$	-	-
TiTok-B-64	-	8.22	-	-	-	-	-	-	9.15	-	-	42.86	-	-
TiTok-S-128	-	-	8.22	-	-	-	-	-	-	9.15	-	-	38.16	-
VQGAN	-	-	-	-	-	-	-	7.04	-	-	7.77	-	-	31.27
ALIT	22.31	15.92	13.08	11.45	10.01	9.12	8.37	8.06	22.01	13.98	9.51	61.32	47.52	38.10
DOVE	18.91	11.46	10.84	9.28	8.61	8.25	7.96	7.73	15.50	9.83	**7.54**	14.83	8.56	7.84

Table 2: FID scores (↓) across the ImageNet100, COCO, and WIT datasets. Our method consistently outperforms ALIT across all token lengths, and achieves comparable or even better results than VQGAN and TiTok at several lengths.

Classification. We evaluate the representation quality of DOVE as an off-the-shelf, frozen backbone across three standard recognition benchmarks, including CIFAR-100 [33], ImageNet-100 [18], and STL-10 [45]. Specifically, we train a lightweight MLP classifier on top of the frozen features, using both mean and max pooling over the final layer representations. As the number of tokens increases, the classification accuracy of both DOVE and ALIT steadily improves. Our approach consistently

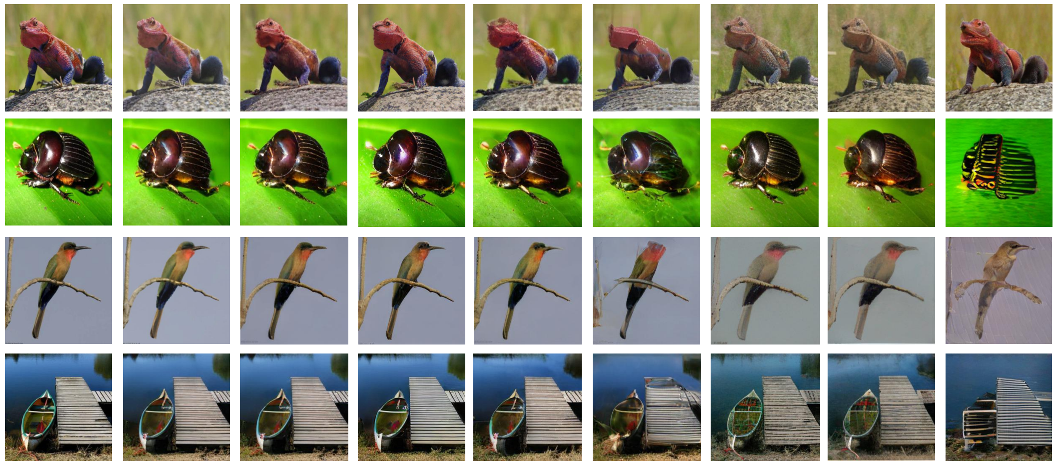
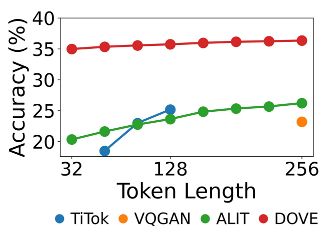
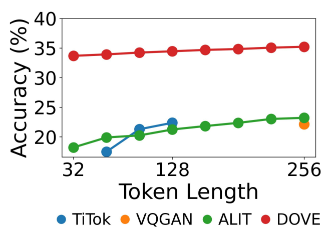


Figure 4: Reconstructed images on ImageNet-1K using different methods. As the token length increases, our method produces progressively clearer reconstructions with more visual details.

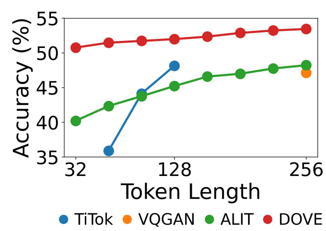
outperforms all other vision tokenizers by a substantial margin. Even when using as few as 32 tokens, it achieves higher classification accuracy than all competing methods. We attribute this advantage to our dynamic reconstruction training objective, which enables the model to capture additional semantic information during representation learning. This is further evidenced by the linear probing and PCA-based zero-shot segmentation results presented in Section 3.4.



(a) CIFAR100



(b) ImageNet100



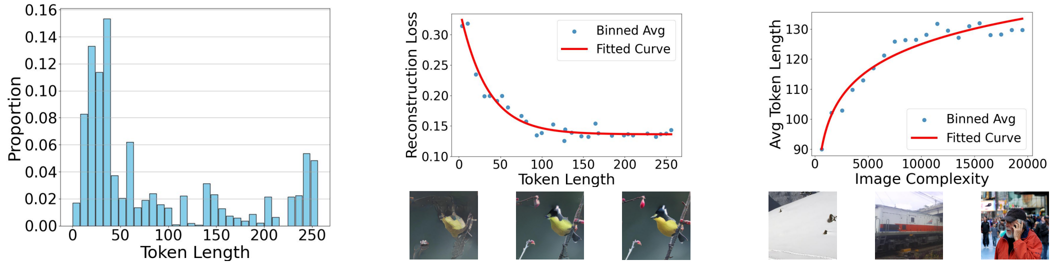
(c) STL-10

Figure 5: Classification accuracy with different visual tokenizers under varying token lengths. DOVE consistently outperforms all baselines across all lengths.

Token Length Distribution. Unlike ALIT, our model explicitly supports a mechanism for generating arbitrary-length token sequences at inference time. We analyze the distribution of token sequence lengths (i.e., EOS positions) generated by DOVE. As shown in Figure 6a, most sequences are shorter than 100 tokens, with smaller peaks around 150 and 250. We randomly sample 5,000 images from the MS COCO 2017 validation set [36] and compute the reconstruction loss across different token lengths. Figure 6b shows that reconstruction loss decreases as token length increases. This decline is steepest between 0 and 100 tokens, and becomes more gradual beyond that. To further investigate the relationship between token length and image content, we calculate the complexity of input images using Laplacian variance [5] and analyze the correlation between image complexity and the length of the generated token sequences. As shown in Figure 6c, by encouraging samples with lower reconstruction quality to delay the EOS position and those with higher quality to emit EOS earlier during training, DOVE naturally learns to allocate longer token sequences to more complex images, while assigning shorter sequences to simpler ones. The Pearson correlation coefficient between image complexity and token sequence length is 0.742.

3.3 Downstream Vision-Language Task Evaluation

Query-conditioned Tokenization. We visualize the behavior of our query-conditioned DOVE (Q-DOVE) on the Visual Genome dataset. Figure 7 presents several examples. The results show that when the input query is “null”, the model clearly reconstructs the entire image. In contrast, when a relevant question or description is provided, the reconstruction focuses on the semantically



(a) Distribution of token sequence lengths (i.e.,EOS positions) generated by DOVE.

(b) The relation between token length and reconstruction loss across different input samples.

(c) The relation between token sequence lengths (i.e.,EOS positions) and image complexity.

Figure 6: Token length analysis

related regions and produces lower frequency outputs for background. This task-driven compression even further reduces the average token sequence length. We then evaluate Q-DOVE and the original DOVE model as vision encoders in downstream vision-language tasks.

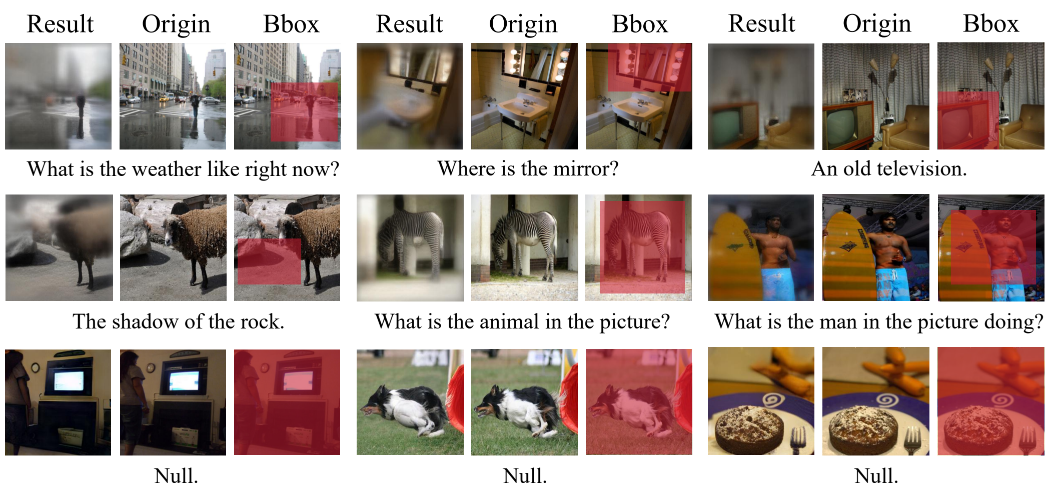


Figure 7: Reconstructed images from the Q-DOVE. When the text query is set to "null", the model reconstructs the entire image. When a query is provided, the model focuses on query-relevant regions.

Visual Question Answering Evaluation. To evaluate the quality of our model's token representations, we replace the vision encoder in a vision-language model with different visual representation methods and evaluate them on downstream vision-language tasks. We adopt Vicuna-7B-v1.5 [37] as the language model, interfacing it with a two-layer MLP that maps the vision encoder outputs to the language model input space. Following the training strategy of AIM V2 [22], we set the learning rate of the language model to 2e-5 and that of the adapter layers to 2e-4. This setup enables joint fine-tuning in a single-stage training process. We fine-tune the model with different vision encoders for one epoch on the 665K mixed VQA dataset used in LLaVA [37]. The model is evaluated on a broad set of benchmarks, including VQAv2 [23], GQA [2], OK-VQA [41], TextVQA [51], DocVQA [44], InfoVQA [43], ChartQA [42], and ScienceQA [39].

Results show that the VLM equipped with DOVE significantly outperforms other models across all datasets. Moreover, integrating Q-DOVE further improves the accuracy. By leveraging DOVE's EOS token as a truncation point, we achieve a substantial reduction in token count with performance comparable to the full set of 256 tokens. For Q-DOVE, we include two input strategies for the vision encoder: providing the actual question or directly inputting a "null". While the "null" setting yields slightly better performance than using the question—which filters out task-irrelevant regions—the question-guided strategy achieves comparable accuracy while further reducing the token length.

We also measure the inference time and floating-point operations (FLOPs) of each model, as shown in Table 3. Both our method and ALIT can effectively reduce FLOPs by shortening the length of the visual token sequence. However, due to ALIT's use of recurrent distillation, where dynamic tokens are generated through multiple passes over VQGAN tokens, its inference speed is adversely affected despite the reduced sequence length. In contrast, our method relies on a single forward pass, resulting in much faster inference.

Model	# Token Count	VQAv2	GQA	OKVQA	TextVQA	DocVQA	InfoVQA	ChartQA	ScienceQA
Titok	128 (S)	43.3	38.8	38.6	14.3	8.1	17.0	11.8	67.1
VQGAN	256	40.2	38.1	37.7	14.3	8.2	16.3	11.1	66.3
ALIT	32	38.4	37.6	35.6	14.2	7.8	16.0	11.4	66.0
	64	39.7	38.0	36.4	14.3	8.1	16.2	11.6	66.2
	128	41.0	38.0	37.2	14.3	8.2	16.3	11.7	66.5
	256	43.8	38.3	37.8	14.3	8.2	16.5	12.0	66.8
DOVE	32	50.3	47.2	42.2	14.6	7.9	18.4	11.2	69.6
	64	51.8	50.2	43.5	14.9	8.2	18.8	12.1	71.7
	128	52.0	50.7	44.8	15.0	8.2	19.1	12.4	72.5
	256	52.4	51.8	46.2	15.0	8.4	19.4	12.6	72.8
	121.6 (Avg)	52.2	51.4	46.0	15.0	8.2	19.2	12.6	72.6
Q-DOVE	256#	**55.0**	**53.2**	**46.7**	**15.3**	**8.6**	**19.7**	**12.8**	**74.8**
	256	53.9	52.6	46.2	15.2	8.2	19.4	12.5	74.0
	82.4 (Avg)	52.8	52.1	46.0	15.2	8.2	19.2	12.4	73.1

Table 3: Performance comparison of VLMs equipped with different vision encoders. DOVE/Q-DOVE consistently achieves the best performance on most tasks. For Q-DOVE, "#" indicates that the input query is set to "null"; otherwise, the original question is used.

Model	VQGAN-256	ALIT-256	ALIT-128	ALIT-64	ALIT-32	DOVE-256	DOVE-128	DOVE-64	DOVE-32
Speed (↑)	1.00×	0.63×	0.82×	0.88×	0.92×	0.96×	1.14×	1.19×	1.26×
FLOPs (T, ↓)	2.62	2.73	1.74	1.31	0.98	2.66	1.70	1.29	0.96

Table 4: Inference speed and FLOPs (in teraflops) of different models. Inference speed is reported as the ratio relative to VQGAN, based on actual inference time measured on the VQAv2 test set.

3.4 Emerging Semantics

From previous experiments, we observe that the visual representations generated by DOVE significantly outperform those produced by fixed-length, autoencoder-based tokenization methods in both classification and downstream multimodal tasks. In this section, we further investigate this emergent semantic property through a series of analyses. Specifically, we evaluate the quality of the learned representations via linear probing on model's hidden layers instead of generated visual tokens and PCA-based image segmentation. We compare DOVE, Q-DOVE, and other fixed-length autoencoder-based tokenizers by conducting linear probing on seven benchmark datasets: CIFAR-10 [33], CIFAR-100 [33], DTD [14], FGVC [40], Food101 [9], STL-10 [15], and SUN397 [57]. For Q-DOVE, we set all text queries to "null" to simulate the unconditional setting. Table 5 shows that DOVE consistently outperforms other methods by a large margin across all datasets, and Q-DOVE further improves upon DOVE's performance. To gain deeper insight into the structure of the learned representations, we apply PCA for dimensionality reduction and visualize the results in image space. As shown in Figure 8, DOVE yields more semantically coherent segmentations compared to VQGAN, while Q-DOVE exhibits even stronger semantic alignment and clarity.

Method	CIFAR-10	CIFAR-100	DTD	FGVC	Food101	STL-10	SUN397
TiTok-32	24.87	6.11	9.46	1.95	3.81	23.23	4.44
TiTok-64	25.95	7.34	10.74	2.61	4.53	28.06	5.23
TiTok-128	18.33	3.10	6.80	2.34	3.05	20.25	3.02
ALIT	41.08	16.87	26.96	4.47	14.47	42.15	20.94
VQGAN	41.23	19.37	24.47	4.38	13.28	40.46	15.20
DOVE	54.31	31.13	26.70	5.85	21.18	48.38	30.62
Q-DOVE	**56.44**	**33.70**	**30.48**	**6.03**	**25.32**	**54.86**	**38.18**

Table 5: Linear probing performance (%) of various models across benchmark datasets.

4 Related Works

Image Tokenization. Image tokenization methods represent images as discrete sets of patch embeddings. In ViT formulations [19], patch representations allow for efficient feature extraction with a

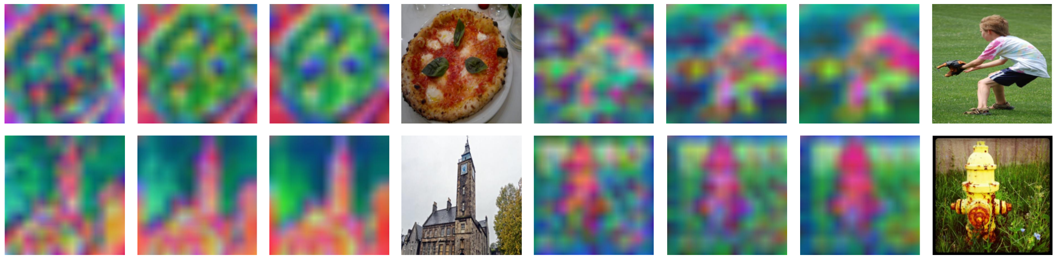


Figure 8: Semantics Visualization with PCA on latent features.

transformer [55] in addition to direct compatibility with tokenized representations in other modALITies, such as text, through the use of projection layers [46, 37]. Through vector quantization [54, 49], patch embeddings from both CNN and transformer encoders can be represented with a finite token codebook, allowing for autoregressive image generation both unimodally [21] and multimodally by conditioning on queries such as text descriptions of images [50, 59, 47]. Whether continuous or quantized, these formulations all encode images into standardized numbers of tokens, independent of image complexity or downstream task demands. In contrast, DOVE represents images using variable numbers of tokens, dynamically adapting to the complexity of images in unimodal settings and to the information demands of downstream tasks in text-conditioned ones.

Token Pruning and Compression. Token pruning methods reduce computation costs by iteratively reducing the set of tokens to be processed across transformer layers, either by dynamically omitting them [58, 48] or by aggregating them in between layers of the transformer [8]. Because these methods iteratively modify the number of tokens across transformer layers, they require modification of the internal structure of models they are applied to. In contrast, DOVE produces variable numbers of tokens, allowing for it to be directly integrated into model pre-training and fine-tuning pipelines. Another branch of work reduces computational costs by compressing token sets at the input level. The Perceiver architecture uses a transformer to compress a set of input tokens into a smaller, fixed set of latent tokens [30, 29], allowing for greater computational tractability in multimodal settings [3]. Similarly, TiTok [60] compresses image patches into a small set of latent tokens, which are then quantized for image reconstruction or other downstream tasks.

Closest to our work is ALIT [20], which uses a recurrent process to distill 2D tokens into a set of 1D latent tokens. Although this iterative process allows for images to be represented by variable numbers of tokens, this is only evidenced through post-hoc analyses, and ALIT does not propose an automated method for dynamically determining the number of tokens to represent an image with at inference time. One of the key innovations of DOVE is the use of a dynamic EOS prediction mechanism, which is employed at inference time to produce per-image variable length token sequences based on image and downstream task complexity. DOVE uses a parallel transformer forward pass to generate variable number of tokens, which is more efficient ALIT's recurrent formulation.

Dynamic Sequence Termination. In the context of transformers, dynamic sequence termination is most commonly associated with the <EOS> token in LLMs [24, 53, 1], although the concept has been applied in language modeling since N-gram models [13]. This concept has also been generalized for generating variable length subsequences of specialized text, such as chain-of-thought chains generated between thinking tokens in LLMs [25]. In sequential decision making, dynamic termination has been operationalized through the use of terminal states in Hidden Markov Models [6], termination conditions in the options reinforcement learning framework [52], as well as by using specialized stop actions within the low-level components of hierarchical policies [28].

5 Conclusion

We have introduced DOVE, a dynamic vision encoder that adaptively generates variable-length token sequences based on image complexity. DOVE predicts an end-of-sequence (EOS) token to dynamically determine the number of tokens needed for image reconstruction, resulting in significantly improved efficiency and semantic representation. We further extended our model with a query-conditioned variant, enabling task-specific focus on relevant image regions. Q-DOVE further improves the representations and token compression achieving stronger efficiency and performance.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints, 2023.

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[4] Ruzena Bajcsy, Yiannis Aloimonos, and John K Tsotsos. Revisiting active perception. *Autonomous Robots*, 42:177–196, 2018.

[5] Raghav Bansal, Gaurav Raj, and Tanupriya Choudhury. Blur image detection using laplacian operator and open-cv. In *2016 International Conference System Modeling & Advancement in Research Trends (SMART)*, pages 63–67. IEEE, 2016.

[6] Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563, 1966.

[7] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR, 2023.

[8] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *The Eleventh International Conference on Learning Representations*, 2023.

[9] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[11] Jieneng Chen, Luoxin Ye, Ju He, Zhao-Yang Wang, Daniel Khashabi, and Alan Yuille. Efficient large multi-modal models via visual context compression. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[12] Leiyu Chen, Shaobo Li, Qiang Bai, Jing Yang, Sanlong Jiang, and Yanming Miao. Review of image classification algorithms based on convolutional neural networks. *Remote Sensing*, 13(22):4712, 2021.

[13] Stanley F Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394, 1999.

[14] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[15] Adam Coates, Honglak Lee, and AY Ng. An analysis of single layer networks in unsupervised feature learning aistats. 2011.

[16] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Scaling up dataset distillation to imagenet-1k with constant memory. In *International Conference on Machine Learning*, pages 6565–6590. PMLR, 2023.

[17] Marianne DeAngelus and Jeff B Pelz. Top-down control of eye movements: Yarbus revisited. *Visual Cognition*, 17(6-7):790–811, 2009.

[18] 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.

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[20] 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.

[21] 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.

[22] Enrico Fini*, Mustafa Shukor*, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Louis Béthune, Zhe Gan, Victor Turrisi, Alexander Toshev, Marcin Eichner, Yinfei Yang, Moin Nabi, Josh Susskind, and Alaaeldin El-Nouby*. Multimodal autoregressive pre-training of large vision encoders, 2024.

[23] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017.

[24] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[25] 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.

[26] Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S Lew. A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval*, 7:87–93, 2018.

[27] Shijie Hao, Yuan Zhou, and Yanrong Guo. A brief survey on semantic segmentation with deep learning. *Neurocomputing*, 406:302–321, 2020.

[28] Muhammad Zubair Irshad, Chih-Yao Ma, and Zsolt Kira. Hierarchical cross-modal agent for robotics vision-and-language navigation. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 13238–13246. IEEE, 2021.

[29] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

[30] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.

[31] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

[32] 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*, 123:32–73, 2017.

[33] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[34] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

[35] Michael Land, Neil Mennie, and Jennifer Rusted. The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28(11):1311–1328, 1999.

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[38] Dengsheng Lu and Qihao Weng. A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, 28(5):823–870, 2007.

[39] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022.

[40] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft, 2013.

[41] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge, 2019.

[42] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning, 2022.

[43] Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. Infographicvqa, 2021.

[44] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images, 2021.

[45] N/A. Stl-10, nov 2024.

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

[47] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[48] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

[49] 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.

[50] 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*, pages 10684–10695, 2022.

[51] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read, 2019.

[52] Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

[53] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[56] Rongkai Xia, Yan Pan, Hanjiang Lai, Cong Liu, and Shuicheng Yan. Supervised hashing for image retrieval via image representation learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28, 2014.

[57] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010.

[58] Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-vit: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10809–10818, 2022.

[59] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

[60] 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, 2024.

[61] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.

[62] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023.