INVERSELY AUTOENCODING MULTIMODAL DATA INTO A SHARED LATENT SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning multimodal representations in a shared latent space is a fundamental challenge towards unified models. In this work, we introduce the Inverse Autoencoder (Inv-AE), a novel framework based on the insight that decoding a shared latent representation into disparate modalities is far more tractable than encoding multimodal data into an aligned space. Inv-AE first optimizes shared latent representations alongside multiple modality-specific decoders, followed by training encoders to map heterogeneous inputs into this established feature space. To accommodate the disparate information densities across modalities, we propose the Latent Representation Sequence (LRS), a multi-scale hierarchical sequence of tensors with progressively increasing sizes, allowing each decoder to selectively leverage an adaptive prefix length. Consequently, low-information-density to high-information-density autoencoding (e.g., images to labels) naturally maps as a classification task, whereas high information density to low-density autoencoding utilizes a Transformer to compensate for finer details via conditional autoregression. Experiments on standard benchmarks demonstrate that our approach delivers competitive performance across multiple tasks while remaining highly extensible to expert architectures with minimal modification.

1 INTRODUCTION

Multimodal models have attracted significant attention in recent years (Bachmann et al., 2024). A key challenge lies in aligning representations from different modalities into a unified latent space. Previous approaches, such as Contrastive Language-Image Pretraining (CLIP; Radford et al., 2021), typically employ contrastive loss functions to regularize the outputs of modality-specific encoders, thereby enforcing them into a shared latent space. In contrast, generative multimodal frameworks tokenize different modalities independently and utilize a unified Transformer architecture to bridge them. However, the shared latent space of CLIP primarily achieves a loose soft-alignment and lacks a decoder for reconstruction, and unified frameworks suffer from token elongation as modalities scale up due to their independent tokenization.

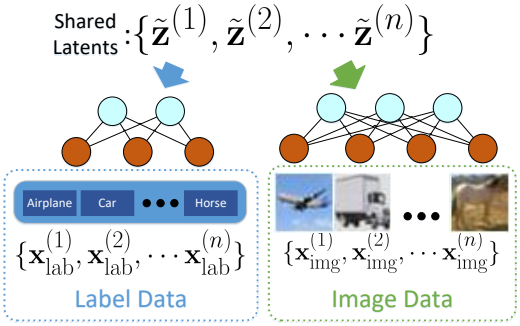

Figure 1: It is straightforward to decode a shared latent representation into different modalities, such as images and labels, using two distinct decoders.

In this work, we propose a novel framework to address this bottleneck, based on the insight that while encoding data from disparate modalities into a shared latent space is highly challenging, decoding a shared latent representation into different modalities is far more tractable. Consequently, we introduce the Inverse Autoencoder (Inv-AE), which maps diverse modalities into shared latent representations by reversing the conventional autoencoder pipeline. In particular, the shared latent representations are tokenized to feed the latent representations into the Transformer for learning. A Latent Representation Sequence (LRS), which is a sequence including tensors representing to-

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.

kens with different sizes, is devised to handle the disparity of information density across different modalities.

As illustrated in Fig. 1, a shared latent vector $\tilde{\mathbf{z}}$ can be easily decoded into modality-specific outputs, e.g., labels $\mathbf{x}_{\text{lab}}$ and images $\mathbf{x}_{\text{img}}$, via two dedicated decoders $d(\tilde{\mathbf{z}}; \phi_{\text{lab}})$ and $d(\tilde{\mathbf{z}}; \phi_{\text{img}})$, respectively. Therefore, rather than following the conventional autoencoder workflow of encoding data into latents followed by decoding the latent back into data, our Inv-AE first initializes shared latent representations to decode data across multiple modalities. In this stage of Inv-AE, the decoders and the latent representations are optimized jointly so that modality-specific outputs are reconstructed from a shared latent space. In particular, the latent representation sequence is devised to handle the disparity in information density across different modalities: it feeds token sequences of varying length into different decoders, which are also fed into an autoregressive model such as a Transformer for autoregression. After the shared latent representations and decoders have been learned, modality-specific encoders are subsequently trained to map heterogeneous inputs into the corresponding latent space. This inverse framework guarantees that the latents of the decoder input are in the same feature space, and the encoders are used to map the data to this shared latent space. It naturally evolves into a versatile any-to-any autoencoding framework, which can perform autoencoding across modalities. If a low-information-density modality like images is autoencoded to a high-density one like labels, it is a classification task and has no need to involve the autoregression. However, if it is a high-information-density modality to low-information-density one, like labels to images, the autoregressive model is used to compensate for the details, and the autoencoding for labels to images actually becomes a label-conditional generation process.

The pipeline of the proposed approach begins with a learnable shared Latent Representation Sequence (LRS), where an optimizable latent variable is assigned to each training instance. In particular, rather than employing a single monolithic tensor, the latent representation is structured as a sequence of tensors to accommodate the varying information densities across different modalities. The spatial dimensions of the nodes, which are latent tensors within this sequence, vary hierarchically. Inspired by visual intuition, we design a multi-scale hierarchy where the tensor sizes increase progressively (e.g., 1×1, 2×2, ..., 32×32). During decoding, each modality leverages a different prefix length of the sequence based on its inherent information density. For example, labels are reconstructed from the first node only, and images utilize all nodes in the sequence. Crucially, this selective conditioning preserves the shared nature of the latent space while providing finer-grained reconstruction as more nodes are appended. Once these latents are fully optimized, modality-specific encoders are trained to project different types of data onto their designated nodes. In particular, since the tokens can be made discrete using quantization methods such as VQ-VAE (van den Oord et al., 2017), cross-entropy loss can be applied to train models to predict these discrete tokens. After the training of encoders and decoders, the entire latent tokens sequence is input into a Transformer for learning autoregression, with the motivation to handle the autoencoding from a high information density modality to a low information density modality, with the ultimate goal of building an any-to-any autoencoding framework. Our main contributions are as follows:

- We propose the Inverse Autoencoder (Inv-AE), a novel framework for addressing the alignment of multimodal data. Shared latent representations and decoders across heterogeneous modalities are optimized first, followed by training encoders to map different modalities into the shared latents.
- The Latent Representation Sequence (LRS) is devised to handle the difference in information density across modalities. It can be input into a Transformer for learning autoregression, with the motivation to do compensation in the autoencoding from a high information density modality (e.g., labels) to a low information density modality (e.g., images).
- Experiments on CIFAR-10, COCO2017 and ImageNet-1K demonstrate the effectiveness and versatility of the proposed model across multiple tasks, establishing a solid foundation for its extension to broader applications and larger-scale studies.

2 Related Work

Research on multimodal modeling has advanced rapidly, driven by the goal of exploiting multiple types of data for better decision making, unified generation, or even a general model for any-to-any

tasks (Bachmann et al., 2024). A key challenge in multimodal modeling lies in how to align different types of data in the same latent space. Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021) aligns multimodal representations into a shared latent space by minimizing a contrastive loss between image and text pairs. However, it primarily achieves a loose soft-alignment and lacks a decoder for cross-modal reconstruction, which limits its flexibility in generative or any-to-any downstream tasks. Besides, unified generation architectures usually model different types of data through a sequential task. Early explorations like Chameleon (Team, 2024) and 4M (Mizrahi et al., 2023) showed that images, text, and other modalities can be modeled through shared token-based representations. Based on it, Transfusion (Zhou et al., 2024) trains a single transformer to predict discrete text tokens and generate continuous image patches using diffusion. OmniGen (Xiao et al., 2025; Wu et al., 2025) handles text-to-image generation and downstream tasks through a unified diffusion model, eliminating the need for task-specific modules. Janus-Pro (Chen et al., 2025b) advances the Janus framework through optimized training and scaling to improve multimodal understanding and instruction following. In addition, several recent methods, including Ming-Omni (AI et al., 2025), Mogao (Liao et al., 2025), BLIP3-o (Chen et al., 2025a), LLaVA-ST (Li et al., 2025), and ILLUME (Wang et al., 2025a), are proposed to further improve omni-modal perception, generation, editing, and spatial-temporal understanding. Conversely, UniCom (Zhao et al., 2026) uses compressed continuous semantic representations instead of discrete visual tokens. As a baseline for traditional perception, EfficientDet (Tan et al., 2020) proposes a weighted bi-directional feature pyramid network to scale resolution, depth, and width for efficient object detection. Shifting to sequential modeling, Bai et al. (2024) train Large Vision Models without linguistic data by representing diverse visual data as "visual sentences" for next-token prediction. LVM-Lite (Li et al., 2026b) decouples single-image and multi-image sequential modeling into a two-stage pipeline to substantially reduce LVM training costs. Diception (Zhao et al., 2025a) repurposes a pretrained text-to-image diffusion model as a generalist perception backbone by treating tasks as image-to-image conditional generation within the RGB space. Similarly, UniVid (Chen et al., 2026) fine-tunes a pretrained video diffusion transformer on diverse vision tasks using visual sentences to demonstrate cross-modal generalization.

Beyond multimodal representation learning, related studies have investigated multimodal tokenization and efficient generation. Recent tokenization methods explore discrete tokenizer design (Li et al., 2026a; Mentzer et al., 2023), vision-language alignment (Zhao et al., 2025b; Xie et al., 2025), and large-scale visual tokenizers (Xiong et al., 2025). Other studies improve multimodal generation and editing through autoregressive generation (Tian et al., 2024), efficient generation (Deng et al., 2026), in-context learning (Sun et al., 2024; Wang et al., 2025b), and multimodal editing (Jiang et al., 2025; Ye et al., 2025).

The proposed approach is different from these previous methods. It is an inverse autoencoder that encodes different modalities of data into a shared latent space. A Transformer compensates for the missing information when autoencoding from a high-information-density modality to a low-information-density modality, such as label-to-image autoencoding.

3 Inverse Autoencoding Framework

3.1 Latent Representation Sequence

The proposed approach starts with the latent representation sequence $\tilde{\mathbf{Z}}_{1:S}$, which consists of sequential nodes $\tilde{\mathbf{Z}}_s$. Each node contains optimizable tensors dedicated to reconstructing the multimodal data. Mathematically, the sequence is structured as follows:

$$\begin{aligned}\tilde{\mathbf{Z}}_{1:S} &= \left\{\tilde{\mathbf{Z}}_1, \tilde{\mathbf{Z}}_2, \cdots, \tilde{\mathbf{Z}}_S\right\} = \left\{\{\tilde{\mathbf{z}}_1^{(n)}\}_{n=1}^N, \{\tilde{\mathbf{z}}_2^{(n)}\}_{n=1}^N, \cdots, \{\tilde{\mathbf{z}}_S^{(n)}\}_{n=1}^N\right\} \\ &= \left\{\{\tilde{\mathbf{z}}_1^{(1)}, \tilde{\mathbf{z}}_1^{(2)}, \cdots, \tilde{\mathbf{z}}_1^{(N)}\}, \{\tilde{\mathbf{z}}_2^{(1)}, \tilde{\mathbf{z}}_2^{(2)}, \cdots, \tilde{\mathbf{z}}_2^{(N)}\}, \cdots, \{\tilde{\mathbf{z}}_S^{(1)}, \tilde{\mathbf{z}}_S^{(2)}, \cdots, \tilde{\mathbf{z}}_S^{(N)}\}\right\}, \\ &= \left\{\{\tilde{\mathbf{z}}_s^{(n)}\}_{n=1}^N\right\}_{s=1}^S\end{aligned} \tag{1}$$

where s and n denote the indices of the sequence nodes and the data instances, respectively. S represents the total number of nodes in the latent representation sequence, and N represents the total number of data instances (e.g., multimodal tuples). Each node contains N optimizable tensors, one for each data instance. To enforce discreteness, we adapt the concept of mutually exclusive random

variables (Zhao & Basu, 2026). Specifically, by injecting symmetric noise into a $\tanh$ activation function wrapped around the optimizable parameters, the continuous latent values are gradually driven towards binary boundaries (i.e., -1 or 1). Formally, the latent variable is formulated as:

$$\tilde{\mathbf{z}}_s^{(n)} = \tanh(\mathbf{z}_s(\theta^{(n)})) + \sigma \cdot \epsilon_s^{(n)}, \quad \epsilon_s^{(n)} \sim \mathcal{N}(0, \mathbf{I}) \tag{2}$$

where $\theta^{(n)}$ denotes the instance-specific optimizable parameters, $\epsilon_s^{(n)}$ is the injected symmetric noise, and σ controls the noise intensity.

3.2 Adaptive Information Density via Latent Representation Sequences

It is widely recognized that information density varies significantly across different modalities. For instance, the discrete and abstract composition of text labels is fundamentally different from the continuous, pixel-level representation of rich visual images. Consequently, when reconstructing multimodal data, the information capacity allocated to each modality should adaptively align with its inherent information density. Our Latent Representation Sequence (LRS) is devised to address this challenge by employing progressively organized nodes to reconstruct different modalities, as illustrated in Fig. 2. Specifically, our design establishes a multi-scale representation tailored to the complexity of each modality. On the one hand, low-density, high-entropy information, such as detailed visual images, requires the full latent representation sequence to capture intricate structures. On the other hand, high-density, low-entropy information, such as text labels, can be effectively reconstructed using only an initial segment of the sequence (e.g., the first node). Taking CIFAR-10 as a benchmark example, the LRS consists of two nodes, as detailed in Tab. 1. For the more complex COCO dataset, which encompasses richer modalities, a more sophisticated node configuration is employed (see Tab. 2). Ultimately, this mechanism yields a unified sequence of shared discrete representations that can adaptively reconstruct diverse multimodal data, as demonstrated in Fig. 2.

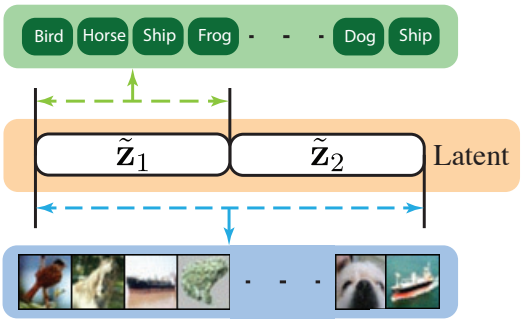

Figure 2: The LRS allocates an adaptive prefix length to each modality: labels are reconstructed from the first node alone, whereas images use the full sequence.

3.3 Inverse Autoencoder

The main framework of the proposed approach is shown in Fig. 3. This approach is fundamentally different from traditional Autoencoders. We strategically choose to decode first, as generating diverse modalities from a single latent representation is substantially easier than the reverse process. In this setting, the latent token sequence is devised and initialized for the different modalities of data. Specifically, every data instance is assigned its own optimizable latent, which serves as the input to the decoders. Mathematically:

$$\begin{aligned}
\mathcal{L}_{\text{dec}}(\{\mathbf{X}_{\text{m}}\}_{\text{m}}; \Phi, \tilde{\mathbf{Z}}_{1:S}) &= \sum_{\text{m}\in\{\text{img,lab}\}} \mathbb{E}_n\Big[\ell_{\text{m}}\big(d(\tilde{\mathbf{Z}}_{1:\beta(\text{m})}; \phi_{\text{m}}), \mathbf{X}_{\text{m}}\big)\Big] \\
&= \mathbb{E}_n\Big[\ell_{\text{img}}\big(d(\tilde{\mathbf{z}}^{(n)}_{1:\beta(\text{img})}; \phi_{\text{img}}), \mathbf{x}^{(n)}_{\text{img}}\big)\Big] + \mathbb{E}_n\Big[\ell_{\text{lab}}\big(d(\tilde{\mathbf{z}}^{(n)}_{1:\beta(\text{lab})}; \phi_{\text{lab}}), \mathbf{x}^{(n)}_{\text{lab}}\big)\Big] \\
&= \mathbb{E}_n\Big[\big\|d(\tilde{\mathbf{z}}^{(n)}_{1:S}; \phi_{\text{img}}) - \mathbf{x}^{(n)}_{\text{img}}\big\|_2^2\Big] + \mathbb{E}_n\Big[-\sum_{k=1}^{K} \mathbb{1}(\mathbf{x}^{(n)}_{\text{lab}} = k) \cdot \log\big[d(\tilde{\mathbf{z}}^{(n)}_1, \phi_{\text{lab}})\big]_k\Big], \\
\text{where } \{\mathbf{X}_{\text{m}}\}_{\text{m}} &= \{\{\mathbf{x}^{(n)}_{\text{m}}\}_{n=1}^{N}\}_{\text{m}\in\{\text{img,lab}\}}; \;\; \Phi = \{\phi_{\text{m}}\}_{\text{m}\in\{\text{img,lab}\}}; \;\; \beta(\text{img}) = S, \; \beta(\text{lab}) = 1.
\end{aligned} \tag{3}$$

where $\text{m} \in \{\text{img}, \text{lab}\}$ denotes the index of modalities. Here we take CIFAR-10 as an example, where only the label and image modalities are involved. $d(\cdot; \phi_{\text{m}})$ denotes the decoder parameterized by ϕ_{m}, and ℓ_{m} is the loss function used for modality m. Since CIFAR-10 is used here, the image loss is the mean squared error and the label loss is the cross-entropy. β_{m} specifies which nodes are used

by which modality. After the decoders and the latents are trained, the encoding process is applied as follows. Mathematically:

$$\begin{aligned}\mathcal{L}_{\text{enc}}(\{\mathbf{X}_\text{m}\}_\text{m}, \tilde{\mathbf{Z}}_{1:S}; \Psi) &= \sum_{\text{m}\in\{\text{img,lab}\}} \mathbb{E}_n \Big[\mathcal{D}_\text{m}\big(e(\mathbf{X}_\text{m}; \psi_\text{m}), \tilde{\mathbf{Z}}_{1:\beta(\text{m})}\big)\Big] \\ &= \sum_{\text{m}\in\{\text{img,lab}\}} \mathbb{E}_n \Big[-\sum_{k=1}^{K} \mathbb{1}(\tilde{\mathbf{z}}^{(n)}_{1:\beta(\text{m})} = k) \cdot \log\big[e(\mathbf{x}^{(n)}_\text{m}; \psi_\text{m})\big]_k\Big],\end{aligned} \tag{4}$$

where $\mathcal{D}(\cdot)$ denotes the distance function used for encoder. Since the latents are discrete tokens, a cross-entropy loss can be applied and driven close to zero. In our implementation, the token prediction accuracy of the encoders reaches nearly 99%. Finally, the overall objective of the proposed Inv-AE is formulated as:

$$\text{Inv-AE}(\{\mathbf{X}_\text{m}\}_\text{m}) = \underbrace{\min_{\tilde{\mathbf{Z}}_{1:S}, \Phi} \sum_{\text{m}\in\{\text{img,lab}\}} \mathcal{L}_{\text{dec}}(\{\mathbf{X}_\text{m}\}_\text{m}; \Phi, \tilde{\mathbf{Z}}_{1:S})}_{\text{decoding first}} + \underbrace{\min_{\Psi} \sum_{\text{m}\in\{\text{img,lab}\}} \mathcal{L}_{\text{enc}}(\{\mathbf{X}_\text{m}\}_\text{m}, \tilde{\mathbf{Z}}_{1:S}; \Psi)}_{\text{then encoding}}, \tag{5}$$

where Φ and Ψ denote the parameters of decoder and encoder respectively. $\tilde{\mathbf{Z}}_{1:S}$ represents our latent representation sequence (LRS). $\{\mathbf{X}_\text{m}\}_\text{m}$ denotes the multimodal data, with m as the modalities index.

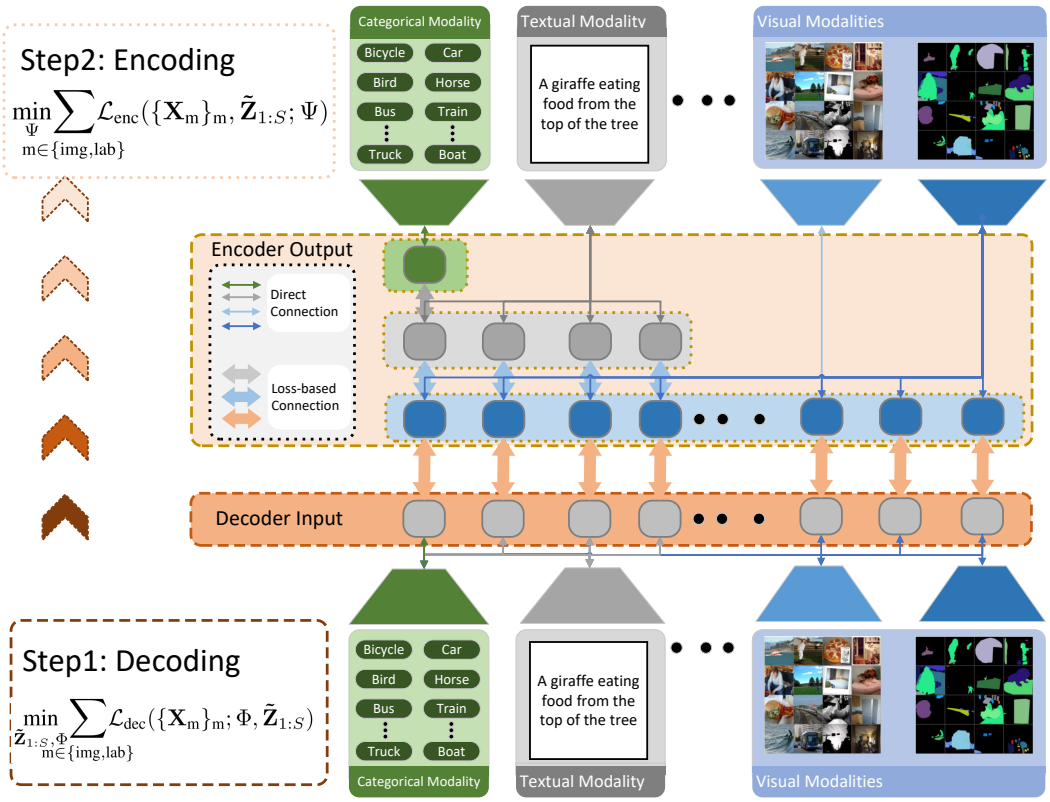

Figure 3: Overview of the proposed Inv-AE. In Step 1, the shared latent representation sequence and the modality-specific decoders are optimized jointly by reconstructing the categorical, textual and visual modalities. In Step 2, the decoders and latents are frozen and modality-specific encoders are trained to map heterogeneous inputs onto their designated nodes of the sequence.

3.4 Autoregression for Any-to-Any Autoencoding

After training, our Inv-AE is able to perform autoencoding across modalities, where the input and output modalities may differ. This property naturally enables any-to-any autoencoding. For exam-

ple, one can encode images into latents and feed those latents to the label decoder; this workflow performs classical image recognition. However, autoencoding from images to labels goes from a low-information-density modality to a high-information-density one, for which no extra information is needed. If autoencoding is performed from a high-information-density modality such as labels to a low-information-density one such as images, a probabilistic model is needed to learn the global distribution of the LRS in order to perform conditional generation. Mathematically:

$$P(\tilde{\mathbf{Z}}_{1:S} \mid e(\mathbf{x}_{\text{src}}; \psi_{\text{src}})) = \underbrace{P(\tilde{\mathbf{Z}}_{1:\beta(\text{src})} | e(\mathbf{x}_{\text{src}}; \psi_{\text{src}}))}_{\text{autoregression anchor}} \underbrace{\prod_{i=\beta(\text{src})+1}^{S} P(\tilde{\mathbf{Z}}_i \mid \tilde{\mathbf{Z}}_{\beta(\text{src}):(i-1)}, e(\mathbf{x}_{\text{src}}; \psi_{\text{src}}))}_{\text{autoregression inference}}, \tag{6}$$

where $\mathbf{x}_{\text{src}}$ denotes the source modalities data. $\beta(\text{src})$ is the index of the entry in the latent representation sequence corresponding to the source modality data. Based on the structural formulation of LRS, multiple algorithmic implementations are technically viable. Mechanistically, a mask-based dual-stream Transformer would likely be the most effective configuration. However, constrained by our available computational budget, we adopt a streamlined causal Transformer that propagates from high-scale to low-scale dimensions. Rather than a mere compromise due to hardware limitations, this asymmetric unidirectional design inherently aligns with the hierarchical topology of our LRS framework. The any-to-any autoencoding can be expressed as:

$$\hat{\mathbf{x}}_{\text{tar}} = \widehat{\text{Inv-AE}}(\mathbf{x}_{\text{src}}; \text{src} \to \text{tar}) = d(\mathbf{z}^*_{\text{tar}}; \phi_{\text{tar}}), \quad \text{where } \mathbf{z}^*_{\text{tar}} \sim P_\omega(\tilde{\mathbf{Z}}_{1:\beta(\text{tar})} \mid e(\mathbf{x}_{\text{src}}; \psi_{\text{src}})), \tag{7}$$

where $\mathbf{z}^*_{\text{tar}}$ denotes the latent representation predicted by the autoregressive model parameterized by ω.

3.5 Transfer Learning

Due to the shared latent space in the proposed framework, a wide range of well-studied network architectures can be seamlessly incorporated into our workflow. As an example, the latent representation sequence (LRS) used for the CIFAR-10 experiments is illustrated in Tab. 1. Specifically, the latent representation sequence consists of two nodes with sizes of $N \times 8 \times 1 \times 1$ and $N \times 8 \times 16 \times 16$, respectively. The first node is decoded into image class labels, while both nodes are decoded into images. After the latent representations and their corresponding decoders are trained, the mapping from images to the first latent node can directly adopt well-established image classification architectures together with their pretrained weights. Specifically, a feature extractor from a pretrained image classification network is adopted to extract features first, and a lightweight Multi-Layer Perceptron (MLP) is attached to project these features into the first LRS node. This architectural compatibility eliminates the need to train the entire framework from scratch; only a lightweight MLP has to be trained. Consequently, the computational cost is significantly reduced while preserving the performance of the pretrained backbone. More importantly, the shared latent representation enables existing task-specific architectures and their pretrained weights to be seamlessly integrated into the proposed any-to-any framework. As a result, the proposed framework naturally inherits the strong performance of well-established backbones and achieves competitive results against not only multimodal models but also task-specific specialist models. This strategy can be readily extended to more complex downstream tasks, such as semantic segmentation and object detection. All comparisons with state-of-the-art methods presented in this paper reuse existing task-specific architectures in this manner. For COCO2017 and ImageNet-1K, the pretrained weights of these architectures are additionally reused, whereas the CIFAR-10 model in Tab. 3 is trained entirely from scratch. Implementation details are provided in the supplementary material.

Table 1: Data modalities of latent representation sequence on CIFAR-10 dataset.

LRS Size	Type	Label	Image
$N \times 8 \times 1 \times 1$	Discrete	✓	✓
$N \times 8 \times 16 \times 16$	Discrete		✓

4 Experiments

All experiments are run on a single RTX 4080 GPU, with experimental details discussed in the following sections.

Table 2: Data modalities of latent representation sequence on COCO2017 dataset and ImageNet-1K dataset.

LRS Size	Type	Bounding Box	Mask	Keypoint	Image	Label
$N \times 12 \times 1 \times 1$	Discrete				✓	✓
$N \times 12 \times 8 \times 8$	Continuous	✓			✓	
$N \times 12 \times 8 \times 8$	Continuous	✓			✓	
$N \times 12 \times 16 \times 16$	Continuous	✓			✓	
$N \times 12 \times 16 \times 16$	Continuous	✓			✓	
$N \times 12 \times 32 \times 32$	Continuous	✓	✓	✓		
$N \times 12 \times 32 \times 32$	Continuous	✓	✓	✓		
$N \times 12 \times 64 \times 64$	Discrete	✓		✓		

4.1 Comparison with State-of-the-Art Methods

Table 3: Comparison on CIFAR-10. Our model is trained entirely from scratch without using pre-trained models.

Method	Reconstruction		Classification	Generation	
	PSNR ↑	rFID ↓	Acc. ↑	IS ↑	FID ↓
VQ-VAE (van den Oord et al., 2017)	23.32	39.67	-	-	-
VQ-WAE (Vuong et al., 2023)	25.93	54.30	-	-	-
SQ-VAE (Takida et al., 2022)	24.07	37.92	-	-	-
CVQ-VAE (Zheng, 2023)	24.72	24.73	-	-	-
TADA+SGD (Nguyen et al., 2025)	-	-	95.60	-	-
DCQ (CHENYUE et al.)	-	-	94.89	-	-
MWT-L (Gielisse, 2025)	-	-	60.75	-	-
ResNet-110 (He et al., 2016)	-	-	93.57	-	-
DenseNet-BC (Huang et al., 2017)	-	-	95.49	-	-
ECM (Geng et al., 2025)	-	-	-	-	3.60
EDM (Zheng et al., 2023)	-	-	-	-	2.51
CD (Song et al.)	-	-	-	9.75	2.93
iCT (Song & Dhariwal, 2024)	-	-	-	9.54	2.83
GET (Geng et al., 2023)	-	-	-	9.16	6.91
Ours	**34.16**	**1.71**	**96.63**	**9.73 ± 0.15**	**2.72**

Table 4: Comparison on COCO2017 and ImageNet-1K. Our model utilizes pretrained specialist models for feature extraction rather than training all components from scratch.

	Method	COCO Det. mAP ↑	COCO Seg. mIoU ↑	COCO KP. mAP ↑	IN-1K Cla. Acc ↑
Special.	YOLO26x (Jocher et al., 2026)	57.50	-	-	-
	SegMAN-B (Fu et al., 2025)	-	48.4	-	-
	RTMPose-t (Jiang et al., 2023)	-	-	66.00	-
	SwinV2-B (Liu et al., 2022)	-	-	-	85.99
Multi.	UnifiedIO-B (Lu et al., 2022)	-	32.9	-	63.3
	UnifiedIO2-B (Lu et al., 2024)	47.20	-	-	-
	Florence2-B (Xiao et al., 2024)	34.7	-	-	-
	DICEPTION (Zhao et al., 2025a)	-	-	57.8	-
	4M7-B (Mizrahi et al., 2023)	49.7	41.3	-	84.5
	4M21-B (Bachmann et al., 2024)	-	42.5	-	84.5
	Ours	**46.32**	**46.4**	**62.29**	**85.80**

Special.: Specialist Model. Multi.: Multimodal Model



Figure 4: Qualitative results of the proposed Inv-AE on COCO2017. Each column shows one input image together with the object detection, keypoint estimation and semantic segmentation outputs decoded from the shared latent representation sequence.

The proposed approach is evaluated on CIFAR-10, COCO2017 and ImageNet-1K. The comparison results on CIFAR-10 and COCO2017/ImageNet are reported in Tab. 3 and Tab. 4, respectively.

For CIFAR-10, the proposed approach is trained from scratch with two modalities, including images and class labels. The Latent Representation Sequence (LRS) is designed as shown in Tab. 1. In particular, the latent tokens are learned as discrete representations using the tokenization method proposed by Zhao & Basu (2026). After the latent sequences are learned, a standard Transformer (Vaswani et al., 2017) is utilized to learn the token distribution for autoregressive modeling. We directly adopt a popular implementation available on GitHub in this work. For Image Classification, the network architecture in TADA+SGD (Nguyen et al., 2025) is utilized. Three tasks are evaluated, including image reconstruction, image classification, and class-conditional image generation. For image reconstruction and classification, the performance is evaluated on the CIFAR-10 test set. For conditional image generation, we follow the standard evaluation protocol and compute FID between the generated images and the training images. As shown in Tab. 3, the proposed approach achieves a PSNR of 34.16 and an rFID of 1.71 for image reconstruction, and an accuracy of 96.63% for image classification. For conditional image generation, it achieves an IS of 9.73 ± 0.15 and an FID of 2.72, showing competitive performance compared with specialist models.

The latent sequence used for COCO2017 and ImageNet-1K is shown in Tab. 2. Specifically, the first latent node is designed as a discrete representation and is mainly used to represent image-level semantic information such as class labels. The six intermediate latent nodes are designed as continuous representations to preserve the fine-grained spatial information required by dense prediction tasks, including object detection, semantic segmentation, and keypoint estimation. The final latent node is also designed as a discrete representation for high-capacity visual information. In our experiments, directly discretizing the intermediate spatial representations results in substantial information loss. Given our limited computational resources, we therefore adopt continuous representations for these intermediate nodes as a practical compromise. Consequently, the current COCO2017 setting does not support fully token-based autoregressive generation across the entire latent sequence. Nevertheless, this compromise does not affect the central formulation of Inv-AE, which focuses on mapping different modalities into a shared latent representation rather than requiring all latent nodes to be discrete. As shown in Tab. 4, the proposed approach achieves 46.32 detection mAP,

46.4 segmentation mIoU, and 62.29 keypoint mAP on COCO2017, with qualitative results shown in Fig. 4. The proposed approach outperforms the compared multimodal models on segmentation and keypoint estimation, and remains competitive with them on detection as well as with the specialist models. These results demonstrate that the proposed LRS can effectively represent different modalities within a unified framework.

Finally, we evaluate the proposed approach on ImageNet-1K with the image classification task. Similar to the experiments on COCO2017, we utilize the pretrained SwinV2-B (Liu et al., 2022) as the feature extractor and map the extracted features into the shared LRS through a lightweight mapping network. For image classification, the first discrete latent node is used to represent the image-level semantic information. As shown in Tab. 4, the proposed approach achieves a classification accuracy of 85.80%, compared with 85.99% achieved by the original SwinV2-B. This result demonstrates that the proposed LRS can preserve the representational capability of the pretrained model while mapping its features into the shared latent representation.

5 Limitations and Discussion

One limitation of this work is the restricted computational budget. Due to the high computational cost of training a unified multimodal model from scratch, we reuse features extracted from pretrained specialist models for some tasks and learn lightweight mappings from these features to the LRS. Although this strategy enables the proposed framework to benefit from strong pretrained representations with limited additional training, directly combining features from independently trained specialist models may not provide the optimal representation for unified multimodal learning. Training these architectures jointly within our framework could potentially provide better performance and stronger interactions across different modalities. Another limitation is that the current LRS does not fully share latent representations across all modalities. Due to the same computational constraint, some modalities are represented by relatively independent latent nodes rather than being jointly learned within a fully shared LRS. For example, in the COCO experiments, the latent representations for classification labels and segmentation are not fully shared. This is a limitation of the current implementation rather than the proposed framework itself, since the LRS naturally supports shared latent representations across modalities. With sufficient computational resources, an important future direction is therefore to jointly train the complete LRS across all modalities, allowing stronger cross-modal interactions and potentially further improving multimodal performance.

6 Conclusion

In this work, we propose the Inverse Autoencoder (Inv-AE), a novel framework for learning shared latent representations across different modalities. Instead of directly aligning the outputs of different encoders, Inv-AE first optimizes shared latent representations with modality-specific decoders, and subsequently trains encoders to map heterogeneous inputs into the established latent space. To handle the difference in information density across modalities, we further introduce the Latent Representation Sequence (LRS), which organizes latent representations as a multi-scale sequence and allows different modalities to utilize different lengths of latent tokens. Combined with autoregressive modeling, the proposed framework naturally supports any-to-any autoencoding across modalities. Experiments on CIFAR-10, COCO2017, and ImageNet-1K demonstrate competitive performance across reconstruction, classification, generation, detection, segmentation, and keypoint estimation. These results demonstrate the effectiveness and flexibility of the proposed framework for unified multimodal representation learning, and provide a foundation for extending Inv-AE to more modalities and larger-scale models.

AI Use Statement

We used a large language model to improve the language and readability of the manuscript. All edits were reviewed by the authors.

Ethics Statement

This work studies multimodal representation learning using existing benchmark datasets, including CIFAR-10, COCO2017, and ImageNet-1K, without collecting new personal data or conducting human-subject studies. The proposed framework may inherit biases present in these datasets and the pretrained models it incorporates. Its generative and visual perception capabilities also carry potential risks of misleading content generation and privacy-sensitive applications. Any deployment should therefore consider dataset and model usage restrictions, fairness, and privacy.

Reproducibility Statement

Section 3 describes the Inv-AE framework, learning objectives, and autoregressive formulation. Tables 1 and 2 specify the latent representation configurations, while Section 4 describes the benchmarks, evaluation metrics, and task-specific architectures. Sections 3.5 and 4 clarify which experiments are trained from scratch and which reuse pretrained specialist models. All experiments are conducted on a single NVIDIA RTX 4080 GPU. The source code will be made publicly available upon acceptance of this paper.

References

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, et al. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*, 2025.

Roman Bachmann, Oğuzhan Fatih Kar, David Mizrahi, Ali Garjani, Mingfei Gao, David Griffiths, Jiaming Hu, Afshin Dehghan, and Amir Zamir. 4m-21: An any-to-any vision model for tens of tasks and modalities. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 61872–61911. Curran Associates, Inc., 2024. doi: 10.52202/079017-1977. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/71883294314045d60c900113a359934b-Paper-Conference.pdf.

Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22861–22872, 2024.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Lan Chen, Yuchao Gu, and Qi Mao. Univid: Unifying vision tasks with pre-trained video generation models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6754–6763, 2026.

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*, 2025b.

YU CHENYUE, Lingao Xiao, Jinhong Deng, Ivor Tsang, and Yang He. Dataset color quantization: A training-oriented framework for dataset-level compression. In *The Fourteenth International Conference on Learning Representations*.

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*, 2026.

Yunxiang Fu, Meng Lou, and Yizhou Yu. Segman: Omni-scale context modeling with state space models and local attention for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19077–19087, June 2025.

Zhengyang Geng, Ashwini Pokle, and J Zico Kolter. One-step diffusion distillation via deep equilibrium models. *Advances in Neural Information Processing Systems*, 36:41914–41931, 2023.

Zhengyang Geng, Ashwini Pokle, Weijian Luo, Justin Lin, and Zico Kolter. Consistency models made easy. In *International Conference on Learning Representations*, volume 2025, pp. 75158–75185, 2025.

Alexander et al Gielisse. End-to-end implicit neural representations for classification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18728–18737, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Houcheng Jiang, Junfeng Fang, Ningyu Zhang, Guojun Ma, Mingyang Wan, Xiang Wang, Xiangnan He, and Tat-seng Chua. Anyedit: Edit any knowledge encoded in language models. *arXiv preprint arXiv:2502.05628*, 2025.

Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023.

Glenn Jocher, Jing Qiu, Mengyu Liu, Shuai Lyu, Fatih Cagatay Akyon, and Muhammet Esat Kalfaoglu. Ultralytics yolo26: Unified real-time end-to-end vision models. *arXiv preprint arXiv:2606.03748*, 2026.

Hongyu Li, Jinyu Chen, Ziyu Wei, Shaofei Huang, Tianrui Hui, Jialin Gao, Xiaoming Wei, and Si Liu. Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8592–8603, 2025.

Jindong Li, Yali Fu, Jiahong Liu, Linxiao Cao, Wei Ji, Menglin Yang, Irwin King, and Ming-Hsuan Yang. Discrete tokenization for multimodal llms: A comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026a.

Xianhang Li, Hongru Zhu, Sucheng Ren, Linjie Yang, Peng Wang, Heng Wang, Xiaohui Shen, Qing Liu, and Cihang Xie. Lvm-lite: Training large vision models with efficient sequential modeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4566–4576, 2026b.

Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An omni foundation model for interleaved multi-modal generation. *arXiv preprint arXiv:2505.05472*, 2025.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12009–12019, 2022.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26439–26455, 2024.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*, 2023.

David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36:58363–58408, 2023.

Dang Nguyen, Jiping Li, Jinghao Zheng, and Baharan Mirzasoleiman. Do we need all the synthetic data? targeted image augmentation via diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Amanda Sastry, Girish Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *International Conference on Learning Representations*, volume 2024, pp. 15078–15097, 2024.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models.

Zeyi Sun, Ziyang Chu, Pan Zhang, Tong Wu, Xiaoyi Dong, Yuhang Zang, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. X-prompt: Towards universal in-context image generation in auto-regressive vision language foundation models. *arXiv preprint arXiv:2412.01824*, 2024.

Yuhta Takida, Takashi Shibuya, WeiHsiang Liao, Chieh-Hsin Lai, Junki Ohmura, Toshimitsu Uesaka, Naoki Murata, Shusuke Takahashi, Toshiyuki Kumakura, and Yuki Mitsufuji. Sq-vae: Variational bayes on discrete representation with self-annealed stochastic quantization. *arXiv preprint arXiv:2205.07547*, 2022.

Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10781–10790, 2020.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

van den Oord et al. Neural discrete representation learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Tung-Long Vuong, Trung Le, He Zhao, Chuanxia Zheng, Mehrtash Harandi, Jianfei Cai, and Dinh Phung. Vector quantized wasserstein auto-encoder. *arXiv preprint arXiv:2302.05917*, 2023.

Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. Illume: Illuminating your llms to see, draw, and self-enhance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21612–21622, 2025a.

Zhaoqing Wang, Xiaobo Xia, Runnan Chen, Dongdong Yu, Changhu Wang, Mingming Gong, and Tongliang Liu. Lavin-dit: Large vision diffusion transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20060–20070, 2025b.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4818–4829, 2024.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13294–13304, 2025.

Rongchang Xie, Chen Du, Ping Song, and Chang Liu. Muse-vl: Modeling unified vlm through semantic discrete encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24135–24146, 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18770–18780, 2025.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025.

Canyu Zhao, Yanlong Sun, Mingyu Liu, Huanyi Zheng, Muzhi Zhu, Zhiyue Zhao, Hao Chen, Tong He, and Chunhua Shen. Diception: A generalist diffusion model for visual perceptual tasks. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 40108–40155. Curran Associates, Inc., 2025a. doi: 10.52202/085713-1339. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/395371f778ebd4854b88521100af30ad-Paper-Conference.pdf.

Chenqiu Zhao and Anup Basu. Local autoregression with finite-support random variables for image generation. In *International Conference on Pattern Recognition*, pp. 162–176. Springer, 2026.

Yaqi Zhao, Wang Lin, Zijian Zhang, Miles Yang, Jingyuan Chen, Wentao Zhang, Zhao Zhong, and Liefeng Bo. Unicom: Unified multimodal modeling via compressed continuous semantic representations. *arXiv preprint arXiv:2603.10702*, 2026.

Yue Zhao, Fuzhao Xue, Scott Reed, Linxi Fan, Yuke Zhu, Jan Kautz, Zhiding Yu, Philipp Krähenbühl, and De-An Huang. Qlip: Text-aligned visual tokenization unifies auto-regressive multimodal understanding and generation. *arXiv preprint arXiv:2502.05178*, 2025b.

Chuanxia et al Zheng. Online clustered codebook. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22798–22807, 2023.

Kaiwen Zheng, Cheng Lu, Jianfei Chen, and Jun Zhu. Dpm-solver-v3: Improved diffusion ode solver with empirical model statistics. *Advances in Neural Information Processing Systems*, 36: 55502–55542, 2023.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.