Representation Forcing for Bottleneck-Free Unified Multimodal Models

Anonymous ECCV 2026 Submission

Paper ID #17

Abstract. Unified multimodal models (UMMs) aim to handle perception and generation in a single model. Yet existing UMMs still rely on a frozen, separately pretrained VAE for image generation, imposing a structural bottleneck. Naively removing it introduces a quality gap, as the model must learn both high-level structure and low-level details from raw pixels. In this paper, we propose Representation Forcing (RF), a technique that closes this gap by making representation prediction a native capability of the model. Concretely, RF forces the decoder to autoregressively predict visual representations as intermediate tokens before pixels; these tokens then stay in context to guide pixel diffusion within the same backbone. By turning representations from perception outputs into generation targets, RF eliminates the need for any external generative latent space. We find that RF benefits both understanding and generation. On image generation, our pixel-space model with RF matches state-of-the-art VAE-based unified models. On image understanding, pixel-space RF generally outperforms its VAE-based variant. Together, these results offer an effective step toward end-to-end, bottleneck-free UMMs.

Keywords: Unified Multimodal Models · Image Generation · Representation Learning · Pixel-Space Diffusion

1 Introduction

The ability to perform understanding (text output) and generation (pixel output) within a unified framework represents a fundamental step toward general-purpose multimodal intelligence [5, 11, 48, 52, 55]. Prevailing unified multimodal models (UMMs) pursue this by bringing language and image generation into a shared transformer backbone [11, 32, 64], with next-token prediction for language and diffusion for image generation. However, despite this unification, the image generation pathway still depends on a separately pretrained, frozen VAE [16, 27, 41]: images are compressed into latents before diffusion is applied, and pixels are recovered through a fixed decoder. This creates a structural bottleneck. The latent space is optimized for reconstruction rather than the objectives of the unified model, and its lossy compression imposes a hard upper bound on generation quality that further training of the UMM cannot overcome. Removing this bottleneck is an important step toward end-to-end UMMs.

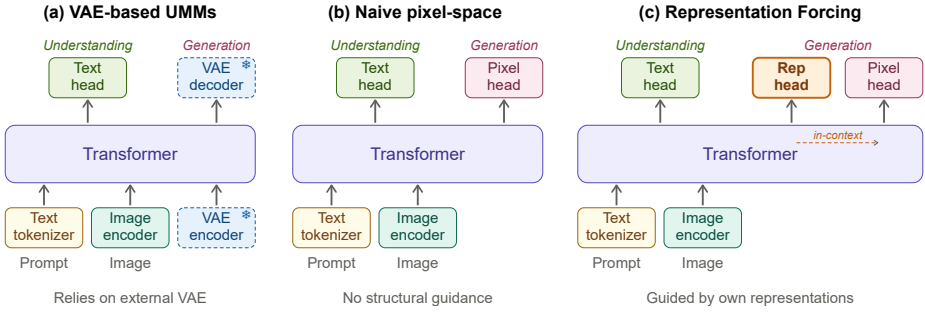


Fig. 1: Architectural comparison. (a) Prevailing UMMs rely on a frozen VAE encoder and decoder for image generation, creating a structural bottleneck. (b) Naively removing the VAE and generating directly in pixel space eliminates this bottleneck but loses structural guidance, leading to a quality gap. (c) Representation Forcing closes this gap by training the transformer decoder to autoregressively predict visual representations (Rep head) before pixel generation. These representations are trained to match features from the model's own understanding encoder and remain in context within the shared transformer, providing structural guidance for pixel-space diffusion without any external latent space.

A natural alternative is to generate directly in pixel space. Recent works have shown pixel-space diffusion to be feasible for standalone generation models [8, 28, 47]. However, we find that directly applying these methods in UMMs fails to match the quality of VAE-based counterparts. We attribute this to the broader image distribution and richer text conditioning in UMMs: the model must learn both the high-level semantic structure and fine-grained details of an image from the same raw signal. This motivates an intermediate representation that separates these two factors, so that the diffusion process can focus on low-level rendering without falling back to an external latent space.

The key question is where such a representation should come from. We observe that UMMs already provide one internally: their understanding pathway learns visual representations that capture high-level structure, such as object identity, spatial layout, and scene composition. In understanding, the encoder [36, 43, 45, 62] extracts these representations from observed images. In generation, however, no image is available, and the model must produce them from the input context alone. This means the model must learn to predict these representations on its own.

In this paper, we propose **Representation Forcing** (RF), an approach that closes this gap by making representation prediction a native capability of the model. Our key idea is to ground the high-level visual representation in the decoder itself. Concretely, we use visual representations extracted by the understanding encoder as targets, and train the model decoder to predict them autoregressively under the same next-token prediction objective used for language. These predicted representations provide an explicit structural scaffold

between text and pixels; they remain in the sequence as in-context conditioning, guiding pixel-space generation within the shared transformer. By turning representations from perception outputs into generation targets, RF grounds understanding and generation in a single representation space, without relying on a separately pretrained latent space (Figure 1).

To validate Representation Forcing, we apply it to both pixel-space and VAE-based UMMs under controlled settings with the same architecture, data, and training budget. On image generation, our pixel-space model with RF matches the VAE-based baseline across standard benchmarks while preserving rich textural details (Figure 2). On understanding, pixel-space RF outperforms its VAE-based variant, showing that pixel-space generation is more compatible with unified multimodal modeling than VAE-based generation. Ablations further reveal that RF is critical for pixel-space generation, while also bringing improvements to VAE-based settings. Together, these findings demonstrate that Representation Forcing benefits both directions of unified multimodal modeling, offering an effective step toward pixel-space, bottleneck-free UMMs.

The main contributions of this work are summarized as follows:

- We propose Representation Forcing (RF), a simple approach that closes the quality gap of pixel-space image generation in unified multimodal models, eliminating the need for any pretrained VAE. RF trains the decoder to autoregressively predict visual representations as intermediate tokens, which then serve as in-context structural guidance for pixel-space diffusion within the same backbone.
- RF benefits both image generation and understanding across pixel-space and VAE-based UMMs. In particular, our pixel-space model with RF matches the VAE-based counterpart on generation and outperforms it on understanding, suggesting that pixel-space generation is more compatible with unified multimodal modeling than VAE-based generation.
- Our work advocates for unified multimodal models where perception and generation share a single, end-to-end-learned representation space, rather than coordinate across separately pretrained components such as external VAEs. Representation Forcing is a step toward fully unified multimodal models, where all capabilities are learned end-to-end within the model itself rather than inherited from independently trained components.

2 Related Work

Unified Multimodal Models. Existing UMMs broadly fall into two families. The first generates within a *single backbone*, either by modeling images as discrete tokens (Chameleon [5], Emu3 [48]) or by attaching diffusion in a VAE latent space (Transfusion [64], Show-o [55], JanusFlow [32]). These approaches further address the interference between understanding and generation through decoupled visual encoders in Janus series [9, 32, 52] or modality-specific experts in BAGEL [11], but all depend on a separately pretrained visual tokenizer—a



Fig. 2: Text-to-image generation results at 1024×1024 resolution from our pixel-space unified model with Representation Forcing.

VQVAE or a continuous VAE—to define the latent space for generation. The second family *stitches an LLM with an external diffusion model*: the LLM predicts visual representations such as CLIP features, which then condition a separately trained diffusion decoder to render images, as in Emu2 [44], SEED-X [20], BLIP3-o [6], and MetaQueries [37]. More recently, Omni [58] expands the set of natively trained modalities to text, images, video, and 3D geometry, and rolls out intermediate multimodal context within the backbone before decoding. A growing line of work further seeks to remove the pretrained VAE that bottlenecks these models. Concurrent with our work, SenseNova-U1 [14] and Tuna-2 [31] build pixel-space UMMs that discard the visual encoder and generate directly from raw patches. Representation Forcing shares this goal of a bottleneck-free, end-to-end UMM, but takes a different path. It builds on the same MoT backbone design [11] and retains a jointly trained understanding encoder, turning its representations into intermediate generation targets that serve as in-context conditioning for pixel-space diffusion within the same transformer, yielding a fully end-to-end UMM.

Pixel-Space Generation. State-of-the-art diffusion models [3, 41] typically operate in the latent space of a pretrained VAE [27], which reduces compute and enables high-resolution synthesis but prevents end-to-end training. A recent line of work has explored generating directly in pixel space [13, 23], enabling end-to-end learning from raw data: JiT [28] demonstrates that plain Vision Transformers with x-prediction can generate on raw pixels, and other approaches explore alternative pixel-space architectures [8, 24, 47]. These methods mostly focus on

standalone generation on the ImageNet [12] dataset. In UMMs, the broader image distribution and richer text conditioning leave naive pixel-space diffusion with a clear quality gap; Representation Forcing closes this gap by providing a structural scaffold from the model's own understanding encoder.

Representation Learning for Generation. Several works explore richer visual representations to improve generation. REPA [59] aligns intermediate diffusion features with frozen pretrained representations to accelerate training convergence; RAE and related works [42, 49, 63] goes further by replacing the VAE entirely with frozen pretrained encoders such as DINOv2 [36] and SigLIP [62], providing a semantically richer latent space. Closely related, Latent Forcing [1] produces high-level structure before pixels by jointly diffusing pixels and frozen DINOv2 latents under separate noise schedules. Despite their differences, all these methods still rely on a frozen, separately pretrained representation space. Representation Forcing instead unfreezes this space: the understanding encoder is trained jointly and end-to-end with the rest of the model, and its features are discretized online into representation tokens that the decoder learns to predict autoregressively. High-level visual structure thus becomes a native output learned within the model rather than inherited from a fixed external space, consistent with our goal of a bottleneck-free, end-to-end UMM.

3 Representation Forcing

The design principle behind Representation Forcing is simple: in understanding, the encoder maps images to representations that capture high-level structure; in generation, the decoder mirrors this by predicting representations from text alone, before rendering them into pixels. In Sec. 3.1, we describe where these representations come from and how they are formulated. In Sec. 3.2, we show how the decoder predicts them to guide pixel-space generation. In Sec. 3.3, we present the overall architecture and training objective. The full training pipeline is illustrated in Figure 3.

3.1 Representations from Understanding

Rather than relying on an external latent space, we seek an intermediate representation that captures high-level structure *from within the model itself*, so that pixel-space diffusion can focus on low-level rendering. The understanding encoder provides a natural source: its features, trained for visual comprehension, already encode the structural content we need. The encoder is jointly trained with the rest of the model. To make its features predictable by the decoder under the same next-token objective as text, and easy to sample at inference, we discretize them into a sequence of *visual representation tokens* via vector quantization. Beyond enabling unified training, discretization encourages the representations to retain high-level structure while discarding fine-grained details, which achieves the factorization we seek between representation prediction and pixel rendering. We validate this choice against continuous regression in Sec. 4.4.

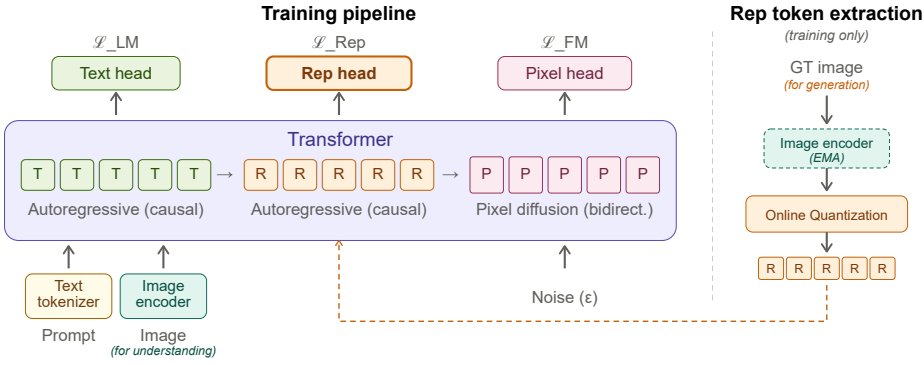


Fig. 3: Training pipeline of Representation Forcing. *Left:* The decoder processes a unified sequence of text tokens (T), representation tokens (R), and pixel patches (P) within a shared transformer. Text and representation tokens are predicted autoregressively under next-token prediction ($\mathcal{L}_{\text{LM}}$ and $\mathcal{L}_{\text{Rep}}$), while pixel patches are generated via bidirectional diffusion from noise ($\mathcal{L}_{\text{FM}}$). The image encoder provides continuous visual features to the transformer for understanding tasks. *Right:* For generation training, an EMA copy of the image encoder extracts features from the ground-truth image, which are discretized via online quantization into representation tokens. These tokens provide both the training targets for $\mathcal{L}_{\text{Rep}}$ and the teacher-forcing inputs at R positions. At inference, the right panel is bypassed entirely: the decoder predicts representation tokens from the text prompt alone, and these tokens remain in context to guide pixel-space diffusion.

We perform this discretization through online vector quantization, which requires no separate pretrained tokenizer. Since the encoder is jointly trained, its features evolve throughout training; we therefore extract features from an exponential moving average (EMA) copy of the encoder, providing slow-moving targets that keep the discrete assignments stable. Specifically, we use patch-level features from the last layer of the EMA encoder, before the final norm.

We maintain a codebook of K prototype embeddings, and tokenization proceeds in three steps. First, we compute the cosine similarity between each patch-level feature and all K prototypes, yielding a score matrix over the batch. Second, rather than greedily assigning each feature to its most similar prototype—which would let a few popular prototypes absorb most features and collapse the codebook—we balance the score matrix with a Sinkhorn–Knopp normalization following SwAV [4], which softly constrains every prototype to receive an equal share of the batch. Third, each feature is assigned to its highest-scoring prototype under the balanced scores, producing a discrete token index: intuitively, a feature keeps its most similar prototype unless that prototype is over-subscribed, in which case it falls back to its next-best match. Finally, the codebook is updated online with EMA updates following VQ-VAE [35], where each prototype moves toward the mean of the features assigned to it. This yields one representa-

tion token per spatial patch, forming a sequence that mirrors the spatial layout of the image and can be predicted in raster-scan order.

3.2 Generating Pixels via Predicted Representations

With the representation tokens defined, we now describe how the decoder uses them to generate images. During training, the EMA encoder provides the representation tokens as ground-truth targets, and the decoder learns to predict them autoregressively under cross-entropy loss, within the same next-token prediction stream as text. At inference, the encoder is no longer involved: the decoder produces the representation tokens from the text prompt alone. We call this process **Representation Forcing**: on one hand, the understanding encoder's representations force the decoder to learn the same high-level visual structure; on the other hand, the decoder's predicted representations force the pixel generation process to follow the intended semantic layout.

Once predicted, the representation tokens remain in the sequence and serve as in-context conditioning for pixel-space generation; the pixel patches form the final image, while the representation tokens themselves are not part of the visible output. The same backbone performs this generation through flow matching. Following JiT [28], we adopt x-prediction with velocity loss. Given clean patches $\mathbf{x}$ and Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, noisy patches at time $t \in [0, 1]$ are

$$\mathbf{z}_t = t\,\mathbf{x} + (1 - t)\,\boldsymbol{\epsilon}. \tag{1}$$

The decoder predicts $\mathbf{x}_\theta$, and the flow-matching loss is

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\big\|\mathbf{v}_\theta - \mathbf{v}\big\|^2, \tag{2}$$

where $\mathbf{v} = \mathbf{x} - \boldsymbol{\epsilon}$ and $\mathbf{v}_\theta = (\mathbf{x}_\theta - \mathbf{z}_t)/(1 - t)$.

At the sequence level, as shown in Figure 3, the model processes a unified token sequence structured as [text tokens, representation tokens, pixel patches]. Within the shared self-attention of the backbone, text and representation tokens follow causal attention as in standard autoregressive modeling, while the noisy pixel patches attend bidirectionally to each other and causally to all preceding text and representation tokens. This attention pattern is what makes the representation tokens act as in-context conditioning: the high-level structure they encode flows into pixel generation through standard self-attention, without any additional cross-attention or injection module.

3.3 Training and Inference

Our model adopts the Mixture-of-Transformers (MoT) architecture following BAGEL [11]. All tokens share the same multi-head self-attention layers, but are routed to modality-specific feed-forward experts based on their type. We maintain three groups of experts: one for multimodal understanding, one for representation token prediction, and one for pixel generation. Each representation token is embedded as the sum of two learnable embeddings: a 2D spatial

position embedding, and a token identity embedding indexed by the discrete token ID. The latter is stored in a separate K-entry table.

The model is trained end-to-end with a combined objective:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \mathcal{L}_{\text{FM}} + \mathcal{L}_{\text{Rep}}, \tag{3}$$

where $\mathcal{L}_{\text{LM}}$ is the cross-entropy loss for text next-token prediction, $\mathcal{L}_{\text{Rep}}$ is the cross-entropy loss for representation token prediction, and $\mathcal{L}_{\text{FM}}$ is the flow-matching loss for pixel generation following the x-prediction formulation in Sec. 3.2. To support classifier-free guidance at inference, we independently drop the text conditioning and the representation token sequence with probability 0.1 during training.

At inference, generation proceeds in two stages. The decoder first produces the full representation token sequence autoregressively from the text prompt. Conditioned on both the text and the predicted representation tokens, the decoder then performs iterative denoising from Gaussian noise to synthesize the final image directly in pixel space. Classifier-free guidance is applied to both conditions.

4 Experiments

4.1 Experimental Setup

Architecture. Our model is initialized from Qwen3-A3B [57], a pretrained Mixture-of-Experts language model with 3B active parameters per token, and follows the Mixture-of-Transformers (MoT) architecture [11]: all tokens share self-attention layers but are routed to one of three modality-specific feed-forward expert pools—understanding, representation prediction, and pixel generation. The image encoder is DINOv3 ViT-H+/16 [43] with NaViT-style variable-resolution support [10], jointly trained with the rest of the model. We use a codebook of K=16,384 prototypes for the online vector quantization. For pixel-space generation we use a 16×16 patch size and adopt x-prediction with velocity loss [28]. The pooling factor is a hyperparameter that trades off representation granularity against sequence length; we use 2×2 pooling throughout, yielding N representation tokens for every $4N$ pixel patches over a shared spatial layout.

Data. We follow the data construction and filtering pipeline of BAGEL [11], training on a mixture of (1) text-only data for language modeling and (2) large-scale text–image pairs covering both image-to-text understanding (general VQA, document comprehension, spatial reasoning) and text-to-image generation.

Training. We adopt a three-stage training strategy following [11]: (i) *alignment*: with the backbone and encoder frozen, we train only the MLP connector for 10K iterations; (ii) *joint pre-training*: we unfreeze all components and jointly optimize on text and text–image pairs at resolutions up to 256 for 50K iterations; (iii) *continued training*: we extend resolutions up to 1024 for 20K iterations. Throughout training, image resolutions are sampled dynamically within

Table 1: Evaluation of text-to-image generation. † refers to methods using LLM rewriter. Our pixel-space model with Representation Forcing (RF-Pixel) matches state-of-the-art VAE-based unified models on both GenEval and DPG-Bench, without relying on any pretrained VAE.

	GenEval							DPG
Model	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑	Overall↑
Generation Only								
PixArt-α [7]	0.98	0.50	0.44	0.80	0.08	0.07	0.48	71.11
SDv2.1 [41]	0.98	0.51	0.44	0.85	0.07	0.17	0.50	68.09
DALL-E 2 [40]	0.94	0.66	0.49	0.77	0.10	0.19	0.52	–
SDXL [38]	0.98	0.74	0.39	0.85	0.15	0.23	0.55	74.65
DALL-E 3 [2]	0.96	0.87	0.47	0.83	0.43	0.45	0.67	83.50
SD3-Medium [15]	0.99	0.94	0.72	0.89	0.33	0.60	0.74	84.08
FLUX.1-dev† [3]	0.98	0.93	0.75	0.93	0.68	0.65	0.82	84.00
Seedream 3.0 [19]	0.99	0.96	0.91	0.93	0.47	0.80	0.84	88.27
Z-Image-Turbo [61]	1.00	0.95	0.77	0.89	0.65	0.68	0.82	84.86
Qwen-Image [51]	0.99	0.92	0.89	0.88	0.76	0.77	0.87	88.32
Unified Models								
Chameleon [5]	–	–	–	–	–	–	0.39	–
LWM [30]	0.93	0.41	0.46	0.79	0.09	0.15	0.47	–
SEED-X [20]	0.97	0.58	0.26	0.80	0.19	0.14	0.49	–
TokenFlow-XL [39]	0.95	0.60	0.41	0.81	0.16	0.24	0.55	73.38
ILLUME [46]	0.99	0.86	0.45	0.71	0.39	0.28	0.61	–
Janus [52]	0.97	0.68	0.30	0.84	0.46	0.42	0.61	–
Transfusion [64]	–	–	–	–	–	–	0.63	–
Emu3† [48]	0.99	0.81	0.42	0.80	0.49	0.45	0.66	81.60
Show-o [55]	0.98	0.80	0.66	0.84	0.31	0.50	0.68	–
Show-o2 [56]	1.00	0.87	0.58	0.92	0.52	0.62	0.76	86.14
Janus-Pro-7B [9]	0.99	0.89	0.59	0.90	0.79	0.66	0.80	84.19
MetaQuery-XL† [37]	–	–	–	–	–	–	0.80	82.05
BLIP3-o [6]	–	–	–	–	–	–	0.84	81.60
UniWorld-V1† [29]	0.98	0.93	0.81	0.89	0.74	0.71	0.84	81.38
OmniGen2† [53]	0.99	0.96	0.74	0.98	0.71	0.75	0.86	83.57
BAGEL [11]	0.99	0.94	0.81	0.88	0.64	0.63	0.82	85.07
BAGEL† [11]	0.98	0.95	0.84	0.95	0.78	0.77	0.88	–
RF-Pixel (ours)	0.99	0.93	0.84	0.89	0.74	0.66	0.84	84.15
RF-Pixel (ours)†	0.98	0.95	0.88	0.87	0.92	0.70	0.88	-

the per-stage maximum and packed via NaViT-style variable-resolution batching. More details are in the appendix.

Baselines. For controlled comparison, we train VAE-based variants using the WanX-2.1 VAE [50], replacing the pixel input/output with VAE latents while keeping the rest of the architecture, training data, and optimization identical. The four variants in our controlled study (Pixel, Pixel+RF, VAE, VAE+RF) differ *only* in (a) the generation pathway and (b) whether RF is enabled. Ablations are conducted at 256 resolution; main results are reported at 1024 resolution.

Table 2: Impact of RF on understanding. We compare understanding performance with and without RF under both VAE-based and pixel-space generation. MME* reports the average accuracy across all perception and cognition questions. +x/−x denotes the change from adding RF. RF improves both settings, with large gains on general visual understanding and only slight reductions on document-oriented tasks. Pixel+RF benefits more from RF than VAE+RF (6/8 vs. 5/8 benchmarks improved), and outperforms VAE+RF on 6 out of 8 benchmarks.

	General Visual Understanding					Document & Diagram		
	MMMU	HalluBench	MME*	BLINK	RealWorldQA	AI2D	DocVQA	ChartQA
VLM-only	56.2	65.0	79.7	56.2	65.8	90.3	89.3	86.0
VAE	51.0	55.7	71.3	52.2	65.2	90.7	90.0	78.8
VAE+RF	49.6 −1.4	61.3 +5.6	79.3 +8.0	52.9 +0.7	66.6 +1.4	87.8 −2.9	88.3 −1.7	80.5 +1.7
Pixel	49.9	63.7	76.6	49.4	63.1	85.8	90.0	81.7
Pixel+RF	54.2 +4.3	64.8 +1.1	80.2 +3.6	53.0 +3.6	65.8 +2.7	90.3 +4.5	88.0 −2.0	81.3 −0.4

4.2 Image Generation

We evaluate our pixel-space model with RF (denoted RF-Pixel in Table 1) on two standard text-to-image benchmarks: GenEval [21] for compositional generation and DPG-Bench [25] for dense prompt following. We compare against both generation-only models and unified multimodal models. Existing unified models all rely on separately pretrained generative components—either a frozen VAE/VQVAE within a single backbone (e.g., BAGEL, Show-o, Janus-Pro), or an external diffusion module conditioned on LLM-predicted features (e.g., BLIP3-o, MetaQuery, SEED-X). In contrast, our model operates entirely in pixel space without any separately pretrained generative module.

As shown in Table 1, without LLM rewriter, RF-Pixel achieves a GenEval overall score of 0.84, slightly outperforming the BAGEL baseline (0.82) and matching unified models such as BLIP3-o (0.84). With LLM rewriter, RF-Pixel reaches 0.88, matching the state-of-the-art among unified models. On DPG-Bench, RF-Pixel scores 84.15 without rewriter, comparable to state-of-the-art VAE-based unified models. These results demonstrate that Representation Forcing effectively closes the quality gap between pixel-space and VAE-based generation, enabling an end-to-end pixel-space unified model to match VAE-based counterparts.

4.3 Image Understanding

Beyond generation quality, we compare how different generation pathways affect understanding performance. As shown in Table 2, we train four generation variants—VAE-based and pixel-space, each with and without RF—on top of the same first-stage VLM baseline, using identical architecture and training data. Since we focus on pretraining-stage comparison, no post-training is applied. We include the VLM-only baseline as a reference.

We evaluate on 8 benchmarks spanning two categories: (1) general visual understanding, including MMMU [60], HallusionBench [22], MME [17], BLINK [18], and RealWorldQA [54], which test general visual understanding, hallucination robustness, and real-world perception; and (2) document and diagram understanding, including AI2D [26], DocVQA [34], and ChartQA [33], which test structured visual comprehension.

RF consistently improves understanding under both generation pathways. For pixel-space generation, RF improves 6 of 8 benchmarks, with substantial gains on MMMU (+4.3), MME (+3.6), BLINK (+3.6), AI2D (+4.5), and RealWorldQA (+2.7), with small reductions on DocVQA (−2.0) and ChartQA (−0.4). For VAE-based generation, RF improves 5 of 8 benchmarks, notably HalluBench (+5.6) and MME (+8.0). The improvements are concentrated on benchmarks that test high-level visual understanding—object recognition, spatial comprehension, and scene-level perception—which aligns with the nature of the representation tokens: they are derived from the understanding encoder and encode semantic structure rather than fine-grained details. The reductions on DocVQA and ChartQA, which rely heavily on precise text recognition and layout parsing, suggest that these capabilities are less directly supported by representation-level guidance.

Pixel+RF outperforms VAE+RF on 6 out of 8 benchmarks. We attribute this to the removal of the external VAE latent space, which allows understanding and generation to share a single representation space more tightly. This result aligns with our broader motivation of moving toward bottleneck-free unified multimodal models.

4.4 Ablation Studies

We conduct ablation studies to analyze the key design choices of Representation Forcing in Table 3.

Representations are critical for pixel-space generation. We evaluate RF under both pixel-space and VAE-based generation, with results shown in Table 3a. All four variants share the same architecture and training setup, differing only in the generation pathway and the presence of RF. Without RF, pixel-space generation scores only 0.25 on GenEval, while VAE-based generation reaches 0.52. As illustrated in Figure 4, the pixel-space model without RF tends to produce images with poor structure, such as distorted objects and incoherent compositions, suggesting that the model struggles to jointly learn high-level layout and low-level detail from raw pixels alone. With RF, pixel-space generation jumps to 0.76, matching the VAE-based counterpart at 0.77, and the generated images become structurally coherent. RF also improves VAE-based generation from 0.52 to 0.77, confirming that explicit representation guidance benefits both settings, though the effect is most pronounced in pixel space.

Decoder prediction outperforms auxiliary alignment. RF is not the only way to incorporate visual representations into generation. REPA [59] takes an alternative approach by adding an auxiliary loss on the diffusion side that

Table 3: Ablation studies. All experiments are conducted under pixel-space generation at 256 resolution unless otherwise noted. (a) Effect of RF under both pixel-space and VAE-based generation. (b) Comparison between RF (decoder prediction) and REPA (auxiliary alignment) as representation guidance strategies. (c) Discrete vs. continuous representation token formulation. (d) Effect of codebook size K. (e) Choice of understanding encoder evaluated on VLM-only benchmarks.

(a) RF on pixel-space and VAE-based generation.

	Pixel	VAE
w/o	0.25	0.52
w/	**0.76**	**0.77**

(b) Prediction vs. alignment.

	GenEval↑
w/o	0.25
REPA	0.43
RF (ours)	**0.76**

(c) Rep. token formulation.

	GenEval↑
w/o	0.25
Continuous	0.26
Discrete	**0.76**

(d) Rep. token codebook size.

	GenEval↑
K=16384	0.76
K=32768	**0.77**

(e) Understanding encoder.

	MMMU	HalluBench	MME*	BLINK	RealWorldQA
SigLIP2	**56.3**	64.4	76.8	53.4	62.5
DINOv3	56.2	**65.0**	**79.7**	**56.2**	**65.8**

encourages the model's intermediate features to align with encoder representations. We apply both methods to the same pixel-space UMM under identical training conditions, using the same visual encoder (DINOv3 [43]) as the representation source, with results shown in Table 3b. Without any guidance, pixel-space generation scores only 0.25. REPA improves this to 0.43, confirming that representation guidance helps, but the gain remains limited. RF achieves 0.76, substantially outperforming REPA. We attribute this difference to where the representation enters the generation process. REPA encourages feature similarity as a side objective, but the aligned features do not explicitly condition the generation output during inference. RF places predicted representations directly in the decoder's sequence, where pixel patches attend to them through shared self-attention. In high-dimensional pixel space, this direct conditioning proves more effective than implicit feature alignment.

Discrete prediction outperforms continuous regression. The representation tokens can be formulated in different ways. Beyond our discrete approach with autoregressive cross-entropy prediction, an alternative is to predict continuous features directly, for example by adding a diffusion head at each representation position to causally regress the encoder features. We compare both formulations under the same setting (Table 3c). Continuous regression scores only 0.26, providing no improvement over the baseline without RF, while discrete tokens achieve 0.76. We attribute this to two factors. First, causally predicting high-dimensional continuous vectors is prone to error accumulation, where small prediction errors at early positions compound and degrade later



Fig. 4: Qualitative comparison of pixel-space generation with and without RF. Without RF, the model tends to produce images with poor structure, such as distorted object shapes and incoherent compositions. With RF, the model generates more coherent structures by first predicting high-level visual representations before pixel rendering, which provides explicit structural guidance for the diffusion process.

predictions. Discrete tokens avoid this by reducing each position to a categorical choice, which is more robust under sequential prediction. Second, discretization naturally encourages the representations to retain high-level structure while discarding fine-grained detail—precisely the factorization that RF is designed to provide. Continuous targets preserve more low-level information, undermining this factorization.

Codebook size. The codebook size K controls the expressiveness of the discrete representation space. We compare K=16,384 and K=32,768 in Table 3d. The two settings perform comparably (0.76 vs. 0.77), suggesting that the model is not sensitive to codebook size within this range. We use K=16,384 in all other experiments, which is consistent with common practice in vector quantization.

Understanding encoder. The choice of understanding encoder determines the quality of the representations that RF learns to predict. We compare SigLIP2 [45], a contrastive vision-language model, and DINOv3 [43], a self-supervised vision model, as the image encoder backbone (Table 3e). We find DINOv3 outperforms SigLIP2 on 4 out of 5 understanding benchmarks in our setting. We attribute DINOv3's advantage to its self-supervised training objective, which produces features with richer spatial and structural information compared to SigLIP2's language-aligned features. This aligns with RF's design: the representation tokens are meant to capture visual structure, which benefits from an encoder that prioritizes spatial fidelity over text alignment. We use DINOv3 as the default encoder.

5 Discussion

Limitations. Due to computational constraints, our model is initialized from a pretrained large language model rather than trained from scratch on multimodal

data. While this provides a strong language-grounded starting point, fully from-scratch multimodal pretraining may yield richer joint representations and is an important direction for future work. We also focus on still-image generation and do not extend RF to video or other temporal modalities.

Conclusion. In this paper, we present Representation Forcing, a simple method for pixel-space image generation in unified multimodal models. The key idea is to let the decoder predict its own understanding representations as autoregressive targets before rendering pixels, providing structural guidance to pixel-space diffusion from within the same sequence. Our experiments show that this single mechanism is enough to close the quality gap with VAE-based generation while also improving multimodal understanding. The same representation serves both directions—interpreting visual inputs and guiding their generation—pointing to a closer integration of perception and generation. We see RF as a step toward fully end-to-end native multimodal learning, where all multimodal capabilities are acquired directly from raw inputs within a single model. We hope this work inspires further research in this direction.

References

1. Baade, A., Chan, E.R., Sargent, K., Chen, C., Johnson, J., Adeli, E., Fei-Fei, L.: Latent forcing: Reordering the diffusion trajectory for pixel-space image generation. arXiv preprint arXiv:2602.11401 (2026)
2. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., Dhariwal, P., Chu, C., Jiao, Y., Ramesh, A.: Improving image generation with better captions. OpenAI Technical Report, `https://cdn.openai.com/papers/dall-e-3.pdf` (2023)
3. Black Forest Labs: FLUX. `https://github.com/black-forest-labs/flux` (2024)
4. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: NeurIPS (2020)
5. Chameleon Team: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
6. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025)
7. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In: ICLR (2024)
8. Chen, S., Ge, C., Zhang, S., Sun, P., Luo, P.: PixelFlow: Pixel-space generative models with flow. arXiv preprint arXiv:2504.07963 (2025)
9. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-Pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
10. Dehghani, M., Mustafa, B., Djolonga, J., Heek, J., Minderer, M., Caron, M., Steiner, A., Puigcerver, J., Geirhos, R., Alabdulmohsin, I.M., et al.: Patch n' Pack: NaViT, a vision transformer for any aspect ratio and resolution. In: NeurIPS (2023)
11. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., Fan, H.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)

12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
13. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS. vol. 34, pp. 8780–8794 (2021)
14. Diao, H., Wu, P., Deng, H., Wang, J., Bai, S., Wu, S., Fan, W., Ye, W., Tong, W., Fan, X., et al.: Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture. arXiv preprint arXiv:2605.12500 (2026)
15. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)
16. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: CVPR (2021)
17. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)
18. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: BLINK: Multimodal large language models can see but not perceive. In: ECCV (2024)
19. Gao, Y., Gong, L., Guo, Q., Hou, X., Lai, Z., Li, F., Li, L., Lian, X., Liao, C., Liu, L., Liu, W., Shi, Y., Sun, S., Tian, Y., Tian, Z., Wang, P., Wang, R., Wang, X., Wang, X., Wang, Y., Wu, G., Wu, J., Xia, X., Xiao, X., Zhai, Z., Zhang, X., Zhang, Q., Zhang, Y., Zhao, S., Yang, J., Huang, W.: Seedream 3.0 technical report. arXiv preprint arXiv:2504.11346 (2025)
20. Ge, Y., Zhao, S., Zhu, J., Ge, Y., Yi, K., Song, L., Li, C., Ding, X., Shan, Y.: SEED-X: Multimodal models with unified multi-granularity comprehension and generation. arXiv preprint arXiv:2404.14396 (2024)
21. Ghosh, D., Hajishirzi, H., Schmidt, L.: GenEval: An object-focused framework for evaluating text-to-image alignment. In: NeurIPS (2023)
22. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., Zhou, T.: HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: CVPR (2024)
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. vol. 33 (2020)
24. Hoogeboom, E., Mensink, T., Heek, J., Lamerigts, K., Gao, R., Salimans, T.: Simpler diffusion (SiD2): 1.5 FID on ImageNet512 with pixel-space diffusion. In: CVPR (2025)
25. Hu, X., Wang, R., Fang, Y., Fu, B., Cheng, P., Yu, G.: ELLA: Equip diffusion models with LLM for enhanced semantic alignment. arXiv preprint arXiv:2403.05135 (2024)
26. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: ECCV (2016)
27. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
28. Li, T., He, K.: Back to basics: Let denoising generative models denoise. arXiv preprint arXiv:2511.13720 (2025)
29. Lin, B., Li, Z., Cheng, X., Niu, Y., Ye, Y., He, X., Yuan, S., Yu, W., Wang, S., Ge, Y., et al.: UniWorld-V1: High-resolution semantic encoders for unified visual understanding and generation. arXiv preprint arXiv:2506.03147 (2025)

30. Liu, H., Yan, W., Zaharia, M., Abbeel, P.: World model on million-length video and language with blockwise RingAttention. arXiv preprint arXiv:2402.08268 (2024)
31. Liu, Z., Ren, W., Huang, X., Chen, S., Li, T., Chen, M., Ji, Y., He, S., Schult, J., Zeng, B., Xiang, T., Chen, W., Luo, P., Zettlemoyer, L., Cong, Y.: Tuna-2: Pixel embeddings beat vision encoders for multimodal understanding and generation. arXiv preprint arXiv:2604.24763 (2026)
32. Ma, Y., Liu, X., Chen, X., Liu, W., Wu, C., Wu, Z., Pan, Z., Xie, Z., Zhang, H., Yu, X., Zhao, L., Wang, Y., Liu, J., Ruan, C.: JanusFlow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In: CVPR (2025)
33. Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E.: ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In: Findings of ACL (2022)
34. Mathew, M., Karatzas, D., Jawahar, C.V.: DocVQA: A dataset for VQA on document images. arXiv preprint arXiv:2007.00398 (2020)
35. van den Oord, A., Vinyals, O., kavukcuoglu, k.: Neural discrete representation learning. In: NeurIPS (2017)
36. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (TMLR) (2024)
37. Pan, X., Shukla, S.N., Singh, A., Zhao, Z., Mishra, S.K., Wang, J., Xu, Z., Chen, J., Li, K., Juefei-Xu, F., Hou, J., Xie, S.: Transfer between modalities with MetaQueries. arXiv preprint arXiv:2504.06256 (2025)
38. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. In: ICLR (2024)
39. Qu, L., Zhang, H., Liu, Y., Wang, X., Jiang, Y., Gao, Y., Ye, H., Du, D.K., Yuan, Z., Wu, X.: TokenFlow: Unified image tokenizer for multimodal understanding and generation. In: CVPR (2025)
40. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125 (2022)
41. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
42. Shi, M., Wang, H., Zheng, W., Yuan, Z., Wu, X., Wang, X., Wan, P., Zhou, J., Lu, J.: Latent diffusion model without variational autoencoder. arxiv: 2510.15301 (2025)
43. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3. arXiv preprint arXiv:2508.10104 (2025)
44. Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Luo, Z., Wang, Y., Rao, Y., Liu, J., Huang, T., Wang, X.: Generative multimodal models are in-context learners (2023)
45. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2:

Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
46. Wang, C., Lu, G., Yang, J., Huang, R., Han, J., Hou, L., Zhang, W., Xu, H.: ILLUME: Illuminating your LLMs to see, draw, and self-enhance. In: ICCV (2025)
47. Wang, S., Gao, Z., Zhu, C., Huang, W., Wang, L.: PixNerd: Pixel neural field diffusion. arXiv preprint arXiv:2507.23268 (2025)
48. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
49. Wang, Y., Ma, C., Lin, Z., Teng, Y., Yu, L., Wang, S., Han, J., Feng, J., Jiang, Y., Liu, X.: Cubic discrete diffusion: Discrete visual generation on high-dimensional representation tokens. In: CVPR (2026)
50. WanTeam, Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
51. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
52. Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., Luo, P.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. In: CVPR (2025)
53. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., Liu, Z., Xia, Z., Li, C., Deng, H., Wang, J., Luo, K., Zhang, B., Lian, D., Wang, X., Wang, Z., Huang, T., Liu, Z.: OmniGen2: Towards instruction-aligned multimodal generation. arXiv preprint arXiv:2506.18871 (2025)
54. xAI: RealWorldQA. `https://huggingface.co/datasets/xai-org/RealworldQA` (2024)
55. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. In: ICLR (2025)
56. Xie, J., Yang, Z., Shou, M.Z.: Show-o2: Improved native unified multimodal models. In: NeurIPS (2025)
57. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
58. Yang, C., Lin, Z., Zhao, Y., Xiao, F., He, H., Zhao, Q., Deng, C., Li, K., Ding, Z., Guo, Y., Wang, F., Zhu, F., Nie, X., Zhu, S., Lin, S., Li, H., Huang, W., Shi, G., Fan, H.: Context unrolling in omni models. arXiv preprint arXiv:2604.21921 (2026)
59. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. In: ICLR (2025)

60. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., Chen, W.: MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In: CVPR (2024)
61. Z-Image Team: Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. arXiv preprint arXiv:2511.22699 (2025)
62. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV (2023)
63. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders. arXiv preprint arXiv:2510.11690 (2025)
64. Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model. In: ICLR (2025)