CoUPLe: Continuous Flow Unifying Pixels and Language

Anonymous authors
Paper under double-blind review

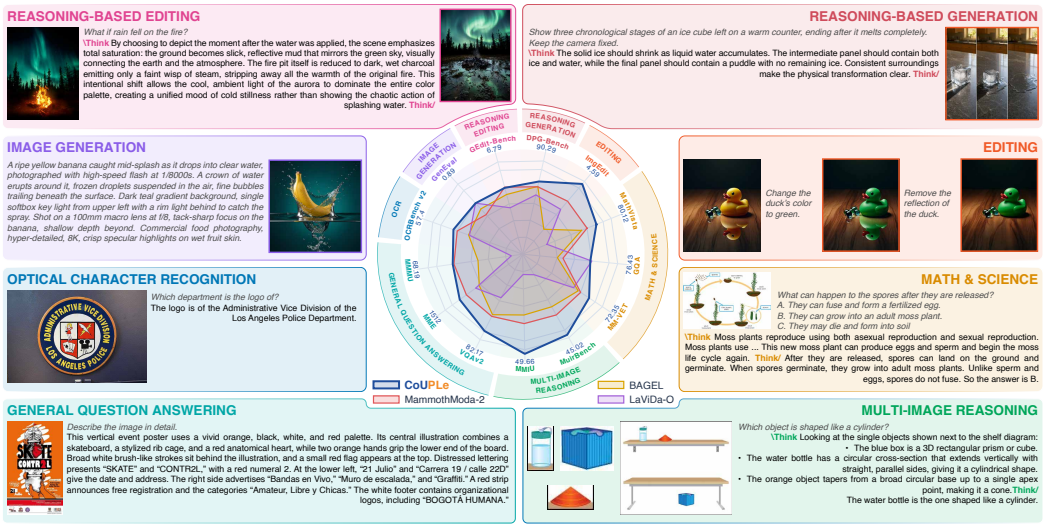


Figure 1: **CoUPLe unifies multimodal latents through continuous flow.** ,we extends pretrained T2I models to multimodal understanding, generation, editing and reasoning.

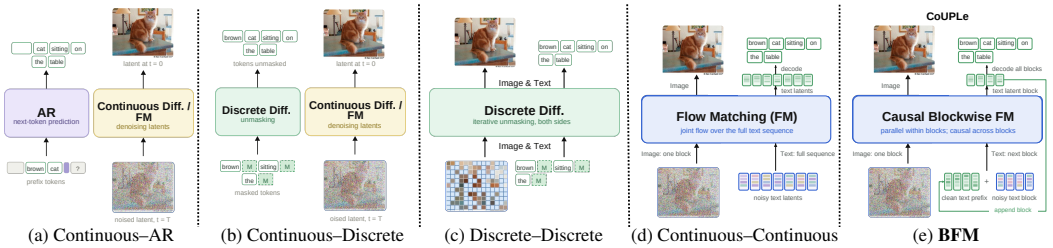
Abstract

We present **CoUPLe**, a framework that turns pretrained text-to-image flow models into unified multimodal generators by modeling text responses and images with a shared continuous velocity field. Our key design decouples text conditioning from text generation: a frozen reference encoder and learned connector align conditioning features with the pretrained backbone, while a Text VAE is jointly adapted with the flow model under reconstruction and reference-posterior KL regularization to maintain response decodability. Causal blockwise flow matching enables variable-length responses through autoregressive generation across latent blocks and parallel refinement within each block, with image synthesis treated as a single image-latent block. This formulation supports visual understanding, multimodal reasoning, and instruction-guided editing through interleaved reasoning–image generation. Instantiated on Flux2-9B and Z-Image, **CoUPLe** supports both MM-DiT and single-stream DiT architectures. Moreover, we find that training **CoUPLe** with both generation and understanding losses achieves better understanding performance than training with understanding loss alone, demonstrating that generation can help understanding in our settings. We will release code and model weights.

1 Introduction

Unified multimodal models aim to integrate visual understanding and generation within a single architecture, enabling shared modeling of language and vision. Existing hybrid approaches typically retain the generative process best suited to each modality. For instance, images are modeled in continuous latent spaces using diffusion or flow-based objectives, while language is generated autoregressively or through discrete diffusion (Zhou et al., 2025; Ma et al., 2025; Deng et al., 2025; Li et al., 2025d; Bill et al., 2026) (Figures 2a and 2b). While effective, this design relies on distinct generative process with different training objectives, state spaces, and sampling dynamics. Such

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.



(a) Continuous–AR (b) Continuous–Discrete (c) Discrete–Discrete (d) Continuous–Continuous (e) **BFM**

Figure 2: **Generative paradigms for unified multimodal models. (a)** Unquantized visual latents, sequential text decoding, and separate objectives. **(b)** Parallel text refinement via discrete diffusion with modality-specific objectives. **(c)** Shared discrete diffusion requiring visual quantization. **(d)** Shared flow in continuous text and image latents without causal text ordering. **(e)** **CoUPLe** generates text causally across blocks with parallel refinement within blocks, and images as single latent blocks, using one flow backbone and objective.

modality-specific generation limits the extent to which understanding and generation can be expressed through a common generative prior, making cross-modal knowledge sharing less direct and requiring separate modeling choices for each modality. Recent efforts toward a common generative process have represented both language and images as discrete tokens, enabling their joint modeling through autoregressive prediction (Chameleon Team, 2024; Wang et al., 2024), masked diffusion (Swerdlow et al., 2025; Shi et al., 2025; Yang et al., 2025; Inclusion AI et al., 2026), or discrete flow matching (Gat et al.; Susladkar et al., 2026) (Figure 2c). While this provides a shared generative process across modalities, it requires continuous visual signals to first be compressed into a finite token vocabulary. This quantization introduces an additional representational bottleneck, where generation is constrained by the fidelity and capacity of the visual tokenizer and fine-grained information lost before generative modeling begins. Moreover, it drifts from the continuous latent spaces of pretrained image generators, making their representations and generative priors difficult to preserve.

Continuous modeling offers an alternative route to a common generative process for language and images. Flow matching learns transport between noise and data through regression of a time-dependent velocity field (Lipman et al., 2023; Liu et al., 2023). With continuous representations, both modalities can share the same training and sampling process, avoiding visual quantization while preserving the latent space and generative prior of pretrained image generators (Esser et al., 2024; Black Forest Labs, 2025; Cai et al., 2025). This provides a natural path to extend visual flow backbones to language under a single generative formulation.

Despite these advantages, fully continuous multimodal generation remains underexplored. OmniFlow (Li et al., 2025b), FlowInOne (Yi et al., 2026) and FlowBind (Cha et al., 2026) (Figure 2d) shows that text, audio and images can share a rectified-flow framework, but key challenges remain for general multimodal understanding and generation: **(1)** Language requires a continuous representation that is not only suitable for flow-based prior learning but also preserves sufficient information for reliable decoding back to discrete text. **(2)** Existing approaches compress an entire text sequence into a small fixed set of latent representations and jointly denoise the full sequence. While effective for short-form generation, this formulation does not explicitly impose the causal factorization desirable for longer responses and multi-step reasoning, where later generations should build on previously generated context. **(3)** Existing continuous multimodal models have primarily been demonstrated on cross-modal generation tasks, leaving more general capabilities such as question-conditioned visual understanding, multi-step reasoning, instruction following, and image editing largely unexplored.

We introduce **CoUPLe** (Figure 2e), a unified continuous-flow framework that adapts pretrained text-to-image backbones to visual understanding, image synthesis, and instruction-guided editing. **CoUPLe** equips the pretrained visual backbone with the ability to generate decodable text latents, bringing response generation and image synthesis under a shared flow-matching objective. Text responses are generated causally across latent blocks, with parallel refinement within each block, while images are generated as single latent blocks. Experiments with two pretrained backbones demonstrate strong performance across visual understanding, image generation, and instruction-guided editing, with ablations supporting its text representation and causal blockwise design.

Whether generative training can help a model's understanding capability is an open question in the unified model research. To shed some light, we compare **CoUPLe** trained with both generation and

understanding losses with the model trained with understanding loss alone, where the combined losses achieves better understanding performance evaluated with MMMU, and MathsVista benchmarks. This demonstrates that the generative training can help understanding capability of **CoUPLe**.

Our main contributions are summarized as follows:

- **A unified continuous multimodal framework.** **CoUPLe** models text and image latents under a shared flow-matching objective, without discrete token prediction or additional latent quantization. Its causal blockwise formulation supports variable-length language generation alongside image synthesis, unifying visual understanding, generation, and editing.
- **Adapting pretrained visual priors for multimodal learning.** We develop a training strategy that integrates pretrained continuous text representations with pretrained visual flow backbones. Text-conditioning alignment preserves the backbone's learned conditioning behavior, while joint Text VAE adaptation with reconstruction and reference-posterior regularization maintains text decodability, enabling joint learning of understanding and generation.
- **State-of-the-art performance across multimodal understanding, generation, and editing.** Instantiated on Flux2-9B (Black Forest Labs, 2025) and Z-Image (Cai et al., 2025), **CoUPLe** supports both MM-DiT and single-stream DiT architectures. Relative to the strongest compared unified baseline on each benchmark, **CoUPLe** improves MS-COCO captioning (CIDEr) by **12.0%**, VQAv2 by **9.4%**, MathVista by **5.8%**, and ImgEdit by **14.3%**, while achieving **90.29** on DPG-Bench. Following multi-image fine-tuning, **CoUPLe** also achieves **45.02** on MuirBench and **49.66** on MMIU, corresponding to relative gains of **11.5%** and **14.8%**, respectively.

2 Related Work

Hybrid multimodal models combine discrete text generation with continuous image synthesis, retaining modality-specific objectives (Zhou et al., 2025; Ma et al., 2025; Deng et al., 2025; Bill et al., 2026). Unified discrete models support a common generative process but require visual quantization (Swerdlow et al., 2025; Shi et al., 2025; Yang et al., 2025; Li et al., 2026). Continuous models avoid this quantization but generate text latents without blockwise causality (Bao et al., 2023; Li et al., 2025b), while blockwise continuous approaches train their multimodal generative backbones from scratch (Guo et al., 2026; Zheng et al., 2026b). CoUPLe brings causal blockwise generation to pretrained visual flow backbones through text-conditioning alignment and joint adaptation of decodable text representations, enabling multimodal understanding, multi-image reasoning, and reasoning-based image generation and editing. Appendix A provides an extended discussion.

3 CoUPLe: Unified Multimodal Flow Matching

We model text and image generation through a shared conditional flow-matching network (Lipman et al., 2023; Liu et al., 2023). Given a target latent $\mathbf{x}$ and a clean condition $\mathbf{c}$, we construct $\mathbf{x}_t = (1-t)\mathbf{x} + t\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $t \sim \mathcal{U}[0,1]$, with data at $t=0$ and noise at $t=1$. The network learns the corresponding velocity by minimizing $\mathcal{L}_{\mathrm{FM}}(\mathbf{x} \mid \mathbf{c}) = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[\|v_\theta(\mathbf{x}_t, t; \mathbf{c}) - (\boldsymbol{\epsilon} - \mathbf{x})\|_2^2\right]$. Modality-specific projections connect text and image latents to the shared backbone. Both directions share θ, with conditioning latents kept clean and the loss restricted to target positions. At inference, we initialize the target with Gaussian noise and integrate $d\mathbf{x}_t/dt = v_\theta(\mathbf{x}_t, t; \mathbf{c})$ from $t=1$ to $t=0$.

3.1 Continuous Representations and Blockwise Generation

We generate both modalities in decodable continuous latent spaces. For images, a frozen autoencoder maps an image to its latent representation, $\mathbf{z}_{\mathrm{I}} = E_{\mathrm{I}}(X_{\mathrm{I}})$, and reconstructs it as $\widehat{X}_{\mathrm{I}} = D_{\mathrm{I}}(\mathbf{z}_{\mathrm{I}})$. Image generation requires reconstruction detail, whereas image understanding benefits from semantic features. We therefore distinguish the autoencoder latent $\mathbf{z}_{\mathrm{I}}$ from semantic conditioning features $\mathbf{h}_{\mathrm{I}}$ produced by a separate frozen encoder. For text, the pretrained T2I checkpoint provides conditioning features without a corresponding text-reconstruction decoder. We introduce a decodable text representation using the pretrained CoLA Text VAE (Guo et al., 2026), comprising an encoder $q_\phi(\mathbf{z} \mid X_{\mathrm{T}})$ that maps text to a continuous latent $\mathbf{z}_{\mathrm{T}}$, and a decoder $p_\omega(X_{\mathrm{T}} \mid \mathbf{z}_{\mathrm{T}})$ that reconstructs the token sequence. The text decoder receives only text latents and multimodal conditioning is incorporated through the shared flow backbone.

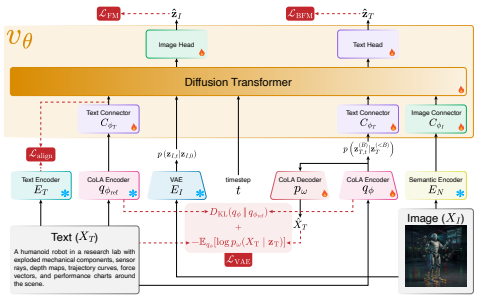


Figure 3: **CoUPLe Stage I.** Image-target and blockwise text-target flow matching share the velocity network v_θ. Flame and snowflake icons denote trainable and frozen modules, respectively.

Text conditioning and text generation impose distinct requirements on the latent representation: the former must remain compatible with the pretrained visual backbone, while the latter must support continuous flow modeling and reliable token reconstruction. We therefore initialize two encoder copies from the pretrained Text VAE: a frozen reference encoder $q_{\phi_{\text{ref}}}$ and a target encoder q_ϕ adapted during Stage I. For image generation, we construct $\widetilde{\mathbf{z}}_{\text{T}} = C_{\psi_T}(\boldsymbol{\mu}_{\phi_{\text{ref}}}(X_{\text{T}}))$, where the learned connector C_{ψ_T} maps the reference posterior mean $\boldsymbol{\mu}_{\phi_{\text{ref}}}$ to the backbone's text-conditioning space. For image-conditioned text generation, we sample $\mathbf{z}_{\text{T}} \sim q_\phi(\mathbf{z} \mid X_{\text{T}})$ as the continuous target, using modality-specific projections to interface with the shared backbone and p_ω to decode generated latents into tokens. This separation allows the text-target representation to adapt jointly with the flow model while anchoring image-generation conditioning to a fixed encoder.

Blockwise generation. We adopt causal blockwise flow matching from (Guo et al., 2026) to generate text through the shared multimodal backbone. We partition a target of L latent positions into $B = \lceil L/C \rceil$ contiguous blocks of size C, with a potentially shorter final block, and optimize

$$\mathcal{L}_{\text{BFM}}(\mathbf{x} \mid \mathbf{c}) = \frac{1}{B}\sum_{b=1}^{B} \mathcal{L}_{\text{FM}}\Big(\mathbf{x}^{(b)} \;\Big|\; \Big[\mathbf{c}; \text{sg}\Big(\mathbf{x}^{(<b)}\Big)\Big]\Big), \tag{1}$$

where $\mathbf{c}$ denotes the clean multimodal context and $\text{sg}(\cdot)$ prevents gradients from reaching the text encoder through preceding clean blocks. At inference, text is generated causally across blocks, with parallel refinement within each block, until a decoded end-of-sequence marker is reached. Images are generated as a single latent block ($B = 1$), recovering the standard flow-matching objective. Appendix G details the causal factorization, parallel training procedure, attention mask, and sampling.

3.2 Multistage Multimodal Training

Training proceeds through bidirectional image–text pretraining (Stage I), visual instruction tuning (Stage II), and interleaved rationale–image training for instruction-guided editing (Stage III).

Stage I: Bidirectional Pretraining. Stage I equips the pretrained visual backbone with text-generation capabilities while maintaining compatible conditioning for image synthesis. Given paired data $(X_{\text{I}}, X_{\text{T}}) \sim p_{\text{data}}$, we jointly train image generation conditioned on $\widetilde{\mathbf{z}}_{\text{T}}$ and blockwise text generation conditioned on semantic image features $\mathbf{h}_{\text{I}}$ and preceding text blocks.

To establish conditioning compatibility, we align the connector output with the original text-encoder features during warmup. To maintain text decodability during joint adaptation, we combine reconstruction supervision with KL regularization toward the frozen reference posterior:

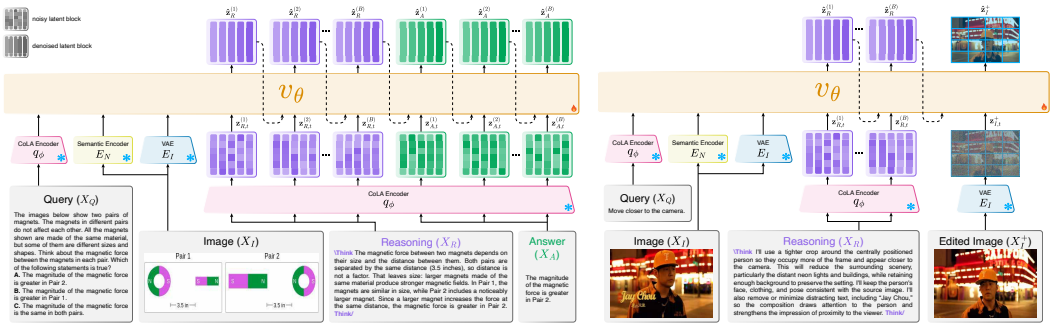
$$\begin{aligned} \mathcal{L}_{\text{align}} &= \mathcal{D}\Big(\widetilde{\mathbf{z}}_{\text{T}}, \text{sg}\Big[E_{\text{T}}^{\text{orig}}(X_{\text{T}})\Big]\Big), \\ \mathcal{L}_{\text{VAE}} &= -\mathbb{E}_{q_\phi}[\log p_\omega(X_{\text{T}} \mid \mathbf{z}_{\text{T}})] + \beta D_{\text{KL}}(q_\phi \,\|\, q_{\phi_{\text{ref}}}). \end{aligned} \tag{2}$$

Here, $\mathcal{D}$ denotes normalized feature regression, and both posteriors are conditioned on X_{T}. The resulting objective is

$$\mathcal{L}_{\text{Bi-Direct}} = \mathbb{E}_{p_{\text{data}}}\Big[\mathcal{L}_{\text{FM}}(\mathbf{z}_{\text{I}} \mid \widetilde{\mathbf{z}}_{\text{T}}) + \mathcal{L}_{\text{BFM}}(\mathbf{z}_{\text{T}} \mid \mathbf{h}_{\text{I}}) + \lambda_{\text{VAE}}\mathcal{L}_{\text{VAE}} + \lambda_{\text{align}}\mathbb{I}[s \le S_{\text{w}}]\mathcal{L}_{\text{align}}\Big], \tag{3}$$

where $\mathbf{z}_{\text{I}}$ denotes the image-VAE target, s is the optimization step, and S_{w} is the alignment warmup duration. We optimize v_θ, including modality-specific projections, the connector $[C_{\psi_T}, C_{\psi_V}]$, and the target Text VAE (q_ϕ, p_ω), while the visual interfaces and both reference text encoders remain frozen. Latent spaces are normalized to unit variance, and the two directional losses use equal coefficients.

Stage II: Instruction Tuning. Stage II adapts the model for image-conditioned instruction following and reasoning. We initialize v_θ from the Stage I checkpoint and freeze the image autoencoder, both text encoder copies, the decoder p_ω, and the connector C_{ψ_T}, so that only v_θ, which includes the modality-specific projections, is updated. Each example $(X_{\text{I}}, X_{\text{Q}}, X_{\text{Y}}) \sim p_{\text{SFT}}$ consists of an image, an instruction, and a target response, respectively. The instruction and response are encoded by



A. Stage II: Instruction Tuning **B. Stage III: Interleaved Training**

Figure 4: **CoUPLe Stages II–III. (A)** Image and instruction latents condition block-causal flows over reasoning and answer text. **(B)** Interleaved training couples rationale generation with image-target flow matching conditioned on the rationale, using the shared velocity network v_θ.

q_ϕ into the same continuous text space, producing $\mathbf{z}_\mathrm{Q}$ and $\mathbf{z}_\mathrm{Y}$, which enter the backbone through modality-specific projections. The frozen semantic image encoder provides image features $\mathbf{h}_\mathrm{I}$. The Stage II objective is

$$\mathcal{L}_\mathrm{IT} = \mathbb{E}_{p_\mathrm{SFT}}[\mathcal{L}_\mathrm{BFM}(\mathbf{z}_\mathrm{Y} \mid [\mathbf{h}_\mathrm{I}; \mathbf{z}_\mathrm{Q}])]\,. \tag{4}$$

When a rationale is supervised, $X_\mathrm{Y} = X_\mathrm{R} \oplus X_\mathrm{A}$ concatenates the rationale and answer using a `<think>` delimiter; otherwise, $X_\mathrm{Y} = X_\mathrm{A}$. The rationale therefore precedes the answer in the same blockwise target sequence. We retain the image-generation term of Eq. (3) alongside Eq. (4), so that a single checkpoint continues to serve both directions.

Stage III: Interleaved Text-Image Generation. Stage III extends blockwise flow matching to instruction-guided editing by modeling rationale and image targets within a causally ordered sequence. Each example $(X_\mathrm{I}, X_\mathrm{Q}, X_\mathrm{R}, X_\mathrm{I}^+) \sim p_\mathrm{III}$ contains a source image, an editing instruction, a supervised rationale, and the edited image. The conditioning context is $\mathbf{c}_\mathrm{III} = [\mathbf{h}_\mathrm{I}; \mathbf{z}_\mathrm{I}; \mathbf{z}_\mathrm{Q}; \widetilde{\mathbf{z}}_\mathrm{Q}]$, where $\mathbf{h}_\mathrm{I}$ and $\mathbf{z}_\mathrm{I}$ denote the source image's semantic features and VAE latents, respectively. The instruction is encoded through both text interfaces, preserving the conditioning conventions established for text and image generation. The target sequence comprises B rationale blocks followed by a single edited-image block, yielding

$$\mathcal{L}_\mathrm{Interleaved} = \mathbb{E}_{p_\mathrm{III}}\Big[\mathcal{L}_\mathrm{BFM}(\mathbf{z}_\mathrm{R} \mid \mathbf{c}_\mathrm{III}) + \lambda_\mathrm{III}\mathcal{L}_\mathrm{FM}\big(\mathbf{z}_\mathrm{I}^+ \mid [\mathbf{c}_\mathrm{III}; \mathrm{sg}(\mathbf{z}_\mathrm{R})]\big)\Big], \tag{5}$$

where $\mathbf{z}_\mathrm{R}$ denotes the encoded rationale, $\mathbf{z}_\mathrm{I}^+$ denotes the target image's VAE latents, and λ_III balances rationale and image supervision. During training, each rationale block conditions on $\mathbf{c}_\mathrm{III}$ and preceding clean rationale blocks, while the image target conditions on the same context and the complete supervised rationale. The attention mask prevents rationale blocks from accessing the edited-image target, while sg detaches the rationale latents supplied as image conditions. We initialize v_θ from Stage II and update only its parameters, including modality-specific projections, while keeping the encoders, decoders, and text-conditioning connector frozen.

At inference, we first generate the rationale blockwise and then use its sampled latents as fixed conditions while generating the edited image as a single target block. Thus, training uses supervised rationale latents, whereas inference uses model-generated latents. Decoding, guidance, and stopping procedures are detailed in Appendix B.

4 Experiments

Tasks and benchmarks. Image understanding (**TI2T**), text-to-image generation (**T2I**), and instruction-guided editing (**TI2I**) differ only in which latent is the condition and which is the target. For TI2T, we report CIDEr on MS-COCO (Chen et al., 2015) captioning, accuracy on VQAv2 (Antol et al., 2015), GQA (Hudson & Manning, 2019), MMMU (Yue et al., 2024), and MathVista (Lu et al., 2024), and the standard scores on MME (Fu et al., 2025a) and MM-VET (Yu et al., 2023). For T2I, we report GenEval (Ghosh et al., 2023) and DPG-Bench (Hu et al., 2024) for prompt alignment, and COCO-30K (Chen et al., 2015) FID for fidelity, both in the standard Inception feature space and in a DINOv2 feature space (DINOv2-FID). For TI2I, we report ImgEdit (Ye et al., 2025) and GEdit-Bench-EN (Liu et al., 2025). We report two instantiations, **CoUPLe** (Flux2-9B) (Black Forest Labs, 2025) and **CoUPLe** (Z-Image) (Cai et al., 2025), each initializing the shared backbone from

Table 1: **Image understanding (TI2T) comparison.** Higher is better. Best and second-best results are **bold** and underlined; highlighted rows denote our models.

Method	Text Gen.	Image Gen.	General Understanding			Reasoning-Based Understanding			
			MS-COCO CIDEr ↑	VQAv2 Acc. ↑	MME ↑	GQA Acc. ↑	MMMU Acc. ↑	MathVista Acc. ↑	MM-VET ↑
Show-o2 (Xie et al., 2025)	AR	Discrete Diff.	–	69.40	1097	58.00	–	25.57	47.12
BAGEL (Deng et al., 2025)	AR	FM	67.32	74.32	1312	61.74	58.22	73.11	67.22
MUDDiT (Shi et al., 2025)	Discrete Diff.	Discrete Diff.	60.22	65.66	941	53.48	51.11	33.42	54.34
MammothModa-2 (Shen et al., 2025)	AR	FM	69.23	75.12	1444	64.23	67.72	75.72	68.37
DDIT (Li et al., 2025d)	Discrete Diff.	Diffusion	56.20	60.10	1124	59.20	–	17.44	44.67
UniDisc (Swerdlow et al., 2025)	Discrete Diff.	Discrete Diff.	46.81	52.34	1000	53.11	49.08	–	49.09
OmniGen2 (Wu et al., 2026)	AR	FM	63.29	67.88	1299	67.67	61.51	65.77	65.23
FUDOKI (Wang et al., 2025c)	DFM	DFM	60.18	64.21	1308	70.77	63.04	69.78	59.38
Lavida-O (Li et al., 2026)	MDM	MDM	–	66.11	1349	**77.09**	45.10	56.90	65.05
CoUPLe (Flux2-9B)	FM	FM	75.23	81.01	1488	76.52	67.59	78.45	70.01
CoUPLe (Z-Image)	FM	FM	**77.52**	**82.17**	**1512**	76.43	**68.19**	**80.12**	**72.35**

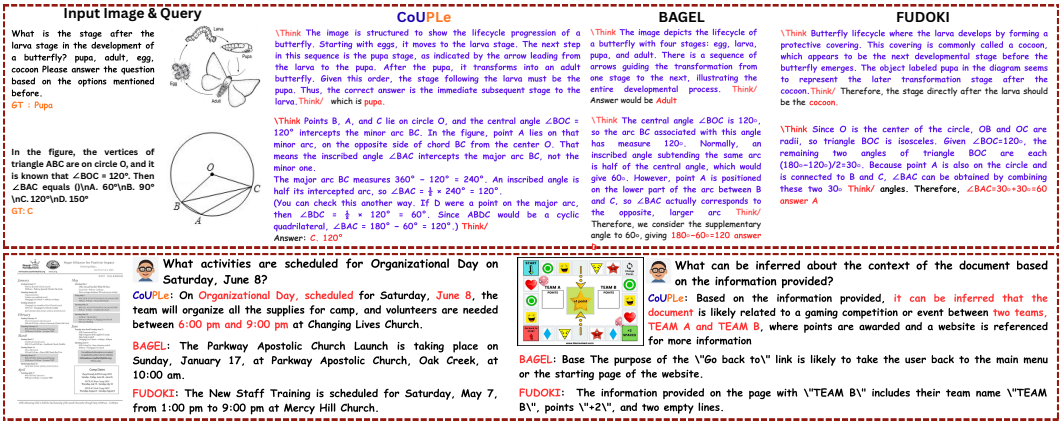


Figure 5: **Qualitative comparison on complex image understanding.** CoUPLe correctly follows the visual lifecycle progression, resolves the geometric relationship, and produces a fine-grained image-grounded caption. In contrast, BAGEL and FUDOKI either select incorrect answers, introduce inconsistent reasoning, or omit salient visual details.

the pretrained text-to-image checkpoint it is named after. Implementation details are in Appendix C, and data and protocols in Appendix D.

4.1 Image Understanding

Baselines. BAGEL (Deng et al., 2025), Show-O2 (Xie et al., 2025) OmniGen2 (Wu et al., 2026), and MammothModa-2 Shen et al. (2025) pair an autoregressive text head with a continuous flow image generator. All four unify the architecture but not the generative process. MUDDiT (Shi et al., 2025), DDIT, UniDisc, FUDOKI (Wang et al., 2025c), and Lavida-O (Li et al., 2026) instead generate text over discrete states, through discrete diffusion, discrete flow matching, or masked diffusion.

CoUPLe (Z-Image) leads on six of the seven benchmarks in Table 1. CoUPLe gains over the strongest baseline are 8.29 CIDEr, 7.05 VQAv2, 0.47 MMMU, 4.40 MathVista, and 68 MME points, and trails only Lavida-O on GQA. Among the four baselines that retain an autoregressive text head, it is best on every reported benchmark, including GQA by 8.76 points. Continuous text generation therefore supports strong image understanding without a separate autoregressive head. The Flux2-9B and Z-Image instantiations differ by at most 2.34 points on the accuracy benchmarks and 24 points on MME, showing that the result transfers across T2I backbones. The discrete-state models trail by larger margins: Lavida-O is 23.09 points behind on MMMU and 23.22 on MathVista, while FUDOKI is 17.34 CIDEr and 17.96 VQAv2 points behind. Together, these results show that a decodable continuous text state can support broad visual understanding within the same flow used for image generation.

Figure 5 shows the same capability in diagram reasoning and document understanding. In the lifecycle and geometry examples, CoUPLe follows relations across visual elements to reach the correct answer rather than relying on isolated object recognition. In the document examples, it locates the relevant evidence in dense content and integrates it to answer the question, while the baselines often copy unrelated details or produce inconsistent answers.

Table 2: **Image synthesis (T2I) and editing (TI2I) comparison.** ↑/↓ indicate higher/lower is better. Best and second-best results are **bold** and underlined, respectively.

Method	Image Generation					ImgEdit				GEdit-Bench-EN		
	Inf. Time (Sec.) ↓	GenEval ↑	DPG-Bench ↑	COCO-30K FID ↓	COCO-30K DINOv2-FID ↓	Add ↑	Extract ↑	Remove ↑	Overall ↑	SC ↑	PQ ↑	Overall ↑
Flux2-9B (Black Forest Labs, 2025)	12.22	0.84	81.23	8.29	8.88	4.57	3.22	3.54	4.08	**7.89**	**7.73**	6.59
Show-o2 (Xie et al., 2025)	**7.82**	0.78	75.78	14.14	15.15	–	–	–	–	–	–	–
BAGEL (Deng et al., 2025)	16.72	0.85	87.74	9.12	11.27	3.56	1.70	2.62	3.20	7.36	6.83	6.52
MUDDiT (Shi et al., 2025)	10.19	0.57	80.09	13.33	16.92	–	–	–	–	–	–	–
MammothModa-2 (Shen et al., 2025)	9.96	0.81	87.53	8.88	10.05	4.57	**3.38**	3.34	4.06	7.77	7.32	6.82
DDIT (Li et al., 2025d)	8.88	0.72	70.77	12.34	11.55	–	–	–	–	–	–	–
UniDisc (Swerdlow et al., 2025)	9.02	0.68	71.28	12.09	13.78	–	–	–	–	–	–	–
OmniGen2 (Wu et al., 2026)	10.87	0.82	81.78	10.92	10.67	3.57	1.77	3.20	3.44	7.16	6.77	6.41
FUDOKI (Wang et al., 2025c)	10.11	0.85	83.21	11.21	10.78	–	–	–	–	–	–	–
Lavida-O (Li et al., 2026)	12.19	**0.89**	83.34	**6.69**	9.23	4.11	2.04	**3.98**	3.81	–	–	5.98
CoUPLe (Flux2-9B)	8.02	0.88	89.88	7.89	8.08	**4.61**	3.31	3.65	**4.64**	**7.89**	7.64	**6.87**
CoUPLe (Z-Image)	7.98	**0.89**	**90.29**	7.75	**7.57**	4.55	3.32	3.74	4.59	7.82	7.61	6.79

Multi-image understanding. CoUPLe conditions causal blockwise text generation jointly on the query and visual representations of all input images. After fine-tuning all methods on Mantis-Instruct (Jiang et al., 2024), CoUPLe achieves 45.02% on MuirBench (Wang et al., 2025b) and 49.66% on MMIU (Meng et al., 2025) (Table 3), exceeding MammothModa-2 by 5.20 and 7.37 percentage points, respectively. Figure 6 illustrates query-conditioned evidence selection across candidate images. More results can be found in Appendix I.1.



Figure 6: CoUPLe identifies St. Basil's Cathedral among four candidate images through query-conditioned reasoning.

4.2 Image Generation and Editing

Table 2 evaluates image synthesis through compositional alignment (GenEval), dense-prompt adherence (DPG-Bench), and distributional fidelity in Inception and DINOv2 feature spaces. Compared with its pretrained initialization, CoUPLe (Flux2-9B) improves GenEval from 0.84 to 0.88 and DPG-Bench from 81.23 to 89.88, while reducing FID from 8.29 to 7.89 and DINOv2-FID from 8.88 to 8.08. Thus, extending the backbone to continuous text generation preserves its synthesis capability, with improvements across all four reported metrics. Among the evaluated unified models, CoUPLe (Z-Image) achieves the highest DPG-Bench score (90.29) and lowest DINOv2-FID (7.57), matches LaViDa-O on GenEval (0.89), and ranks second on FID (7.75). Relative to BAGEL, the strongest competing model on DPG-Bench, it gains 2.55 points in prompt adherence while reducing FID and DINOv2-FID by 1.37 and 3.70, respectively. Image-generation latency for all methods, including every baseline and both CoUPLe variants, is measured on a single NVIDIA A100 40GB GPU. CoUPLe (Flux2-9B) and CoUPLe (Z-Image) require 8.02 and 7.98 seconds per image, respectively.

For instruction-guided editing, the shared backbone generates rationale latents followed by the edited-image latent (§3.2). CoUPLe (Flux2-9B) obtains the highest overall scores on ImgEdit (4.64) and GEdit-Bench-EN (6.87), exceeding its pretrained initialization by 0.56 and 0.28 points. Operation-specific results reveal complementary strengths: LaViDa-O leads removal (3.98), whereas MammothModa-2 leads extraction (3.38). Figure 7 illustrates simultaneous clothing-color and illumination changes, localized clock-hand modification, and generation of astronomical and mathematical scenes with multiple compositional constraints. Additional qualitative comparisons are provided in Appendices H.6 and I, with category-level GenEval and DPG-Bench results in Appendix H.1.

Table 3: **Multi-image understanding**.

Method	MuirBench	MMIU
BAGEL	37.67	40.19
MammothModa-2	39.82	42.29
LaViDa-O	29.27	30.33
OmniGen2	33.42	39.98
FUDOKI	30.11	32.39
CoUPLe	**45.02**	**49.66**

Can generation help understanding? Table 4(A) varies Stage-I supervision. Joint pretraining optimizes $\mathcal{L}_{\text{FM}}(z_I \mid \tilde{z}_T)$ and $\mathcal{L}_{\text{BFM}}(z_T \mid h_I)$; text-only and image-only controls remove the image-target and text-target losses, respectively. Full-pipeline variants retain Stages II–III, including Stage-II image-generation training; Stage-I-only variants omit both later stages. Compared with text-only Stage I, joint pretraining improves MMMU/MathVista by 2.18/3.80 points and GenEval by 0.10, while reducing COCO-FID by 2.03 after Stages II–III. At the Stage-I checkpoint, the

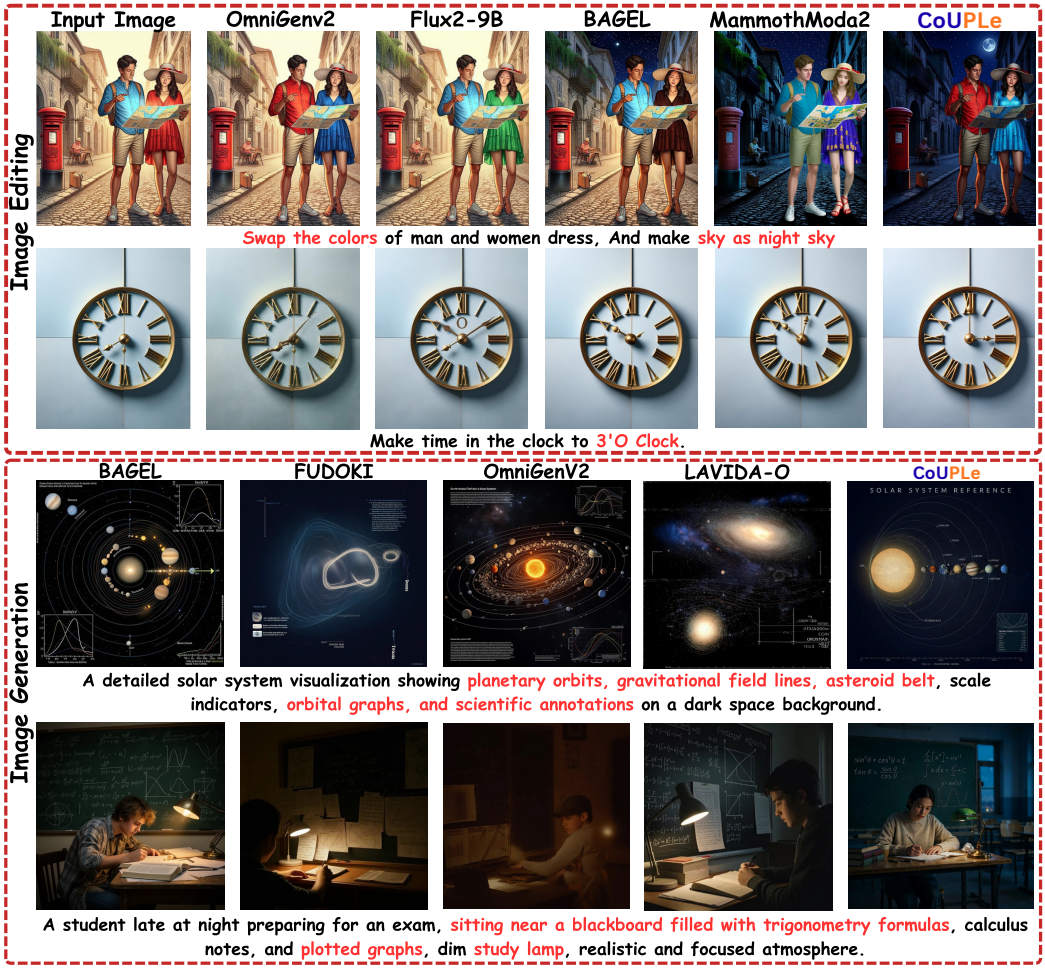


Figure 7: **Qualitative comparison on instruction-guided Image editing (TI2I) and Image Generation (T2I)**. Zoom to see the quality and differences.

corresponding gains are 2.92/1.41 points, 0.11 GenEval, and a 3.26 FID reduction. Image-only Stage I followed by Stages II–III also underperforms the full model (65.01/59.82 MMMU/MathVista, 0.85 GenEval, and 8.82 FID). These results support transfer from image-target optimization to text-response accuracy and complementary benefits from bidirectional supervision. Task-specific image-generation fine-tuning for counting and text rendering provides further evidence of transfer to understanding (Appendix E). Frozen-backbone ImageNet linear probing (Appendix F) and image–text feature-affinity visualization (Appendix G.1) further characterize visual representations and cross-modal alignment after joint training.

5 Ablations

Training strategy. Table 4(B) examines backbone initialization, adaptation, interleaved training, and causal text generation. Removing Stage III eliminates the rationale-to-image objective of Eq. (5), reducing MMMU/MathVista by 1.92/2.20 points and GenEval to 0.86 while increasing FID from 7.75 to 8.89. Training from scratch raises FID to 13.34; restricting adaptation to LoRA raises it to 18.27, compared with 7.75 for full adaptation of the pretrained T2I backbone. Finally, replacing blockwise causal flow with joint full-sequence denoising reduces MMMU/MathVista to 52.01/54.42 (GenEval 0.83; FID 8.88). The large understanding drop supports conditioning each generated text block on completed prefix blocks.

Ablation based on Stage-I loss. The two auxiliary losses support the two roles of text in the shared flow. Removing the Text-VAE loss in Table 4 (C) lowers MMMU and MathVista to 63.37 and 73.37 and raises COCO-FID to 13.11. Reconstruction and latent regularization keep the generated text latent

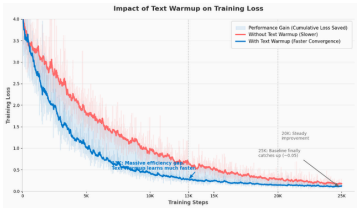


Figure 8: **Text warmup.** Warmup alignment reaches a comparable loss in half the steps.

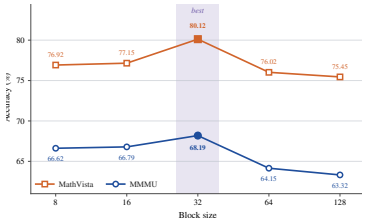


Figure 9: **Block size.** MMMU and MathVista accuracy peak at a block size of 32.

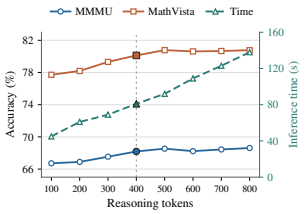


Figure 10: **Token budget.** Accuracy saturates beyond 500 tokens.

decodable while adaptation to the shared backbone (Heek et al., 2026). Warmup alignment maps the text condition to the representation expected by the pretrained T2I model. Removing it has a smaller effect on understanding but reduces GenEval by 0.06 and increases COCO-FID by 2.93. It also doubles the number of steps needed to reach a comparable training loss (Figure 8). Two losses let's text become a generation target without disrupting its role as the condition for image generation.

Table 4: **CoUPLe ablation analysis**

Configuration	MMMU Acc. ↑	Math Vista ↑	Gen Eval ↑	COCO FID ↓
CoUPLe	**68.19**	**80.12**	**0.89**	**7.75**
(A) Can generation help understanding? Yes, by our CoUPLe model				
Stage I (Text Gen.) + Stages II–III	66.01	76.32	0.79	9.78
Stage I (Image Gen.) + Stages II–III	65.01	59.82	0.85	8.82
Only Stage I (Joint)	59.23	49.13	0.73	12.01
Only Stage I (Text Gen.)	56.31	47.72	0.62	15.27
(B) Ablations on training strategy				
Training from Scratch	63.39	74.29	0.78	13.34
LoRA Fine-Tuning	48.23	39.71	0.72	18.27
w/o Interleaved Training	66.27	77.92	0.86	8.89
w/o Blockwise Causal Learning	52.01	54.42	0.83	8.88
(C) Ablations on training objectives				
w/o $\mathcal{L}_{\text{VAE}}$	63.37	73.37	0.79	13.11
w/o $\mathcal{L}_{\text{align}}$	67.12	78.89	0.83	10.68
(D) Ablations on vision-encoder configuration				
RAE + RAE	66.05	76.72	0.84	11.37
VAE + VAE	50.34	67.82	0.87	8.72
RAE + VAE	**68.19**	**80.12**	**0.89**	**7.75**

Condition and target favor different image representations. The two image roles place different demands on their representations. Using representation autoencoder (RAE) features (Zheng et al., 2026a) for both roles retains most of the understanding performance but increases COCO-FID by 3.62 in Table 4 (D). Using the image VAE of the T2I checkpoint for both roles keeps COCO-FID within 0.97 of the full model but lowers MMMU from 68.19 to 50.34. RAE features provide the semantic image condition needed for text generation, whereas VAE latents provide the reconstruction space needed for image generation. Combining the RAE condition with the VAE target is best on all four metrics. The shared flow therefore benefits from role-specific representations rather than forcing the same image latent into both directions. A qualitative example is provided in Appendix H.2.

Effect of block size. Block size determines how many text latent positions are refined jointly before updating the causal prefix. As shown in Figure 9, a block size of 32 achieves the best accuracy on MMMU (68.19) and MathVista (80.12). Increasing the size to 64 lowers both scores to 64.15 and 76.02, respectively, with further degradation at 128. Smaller blocks (8–16) also underperform while requiring more sequential block-generation steps. These results favor a moderate block size that balances parallel latent refinement with conditioning on completed reasoning prefixes. We therefore use 32 throughout. More ablation for text-encoder and reasoning based editing is in Appendix H.5

Effect of the reasoning-token budget. Figure 10 shows diminishing accuracy gains with longer reasoning sequences. Increasing the budget from 100 to 400 tokens improves MMMU and MathVista by 1.47 and 2.40 percentage points, respectively. Accuracy largely saturates beyond 500 tokens: extending the budget from 500 to 800 increases latency by 50% with negligible accuracy changes. A budget of 400–500 tokens therefore provides a favorable accuracy–latency trade-off in this setting.

6 Conclusion

We introduce **CoUPLe**, a unified continuous-flow framework that extends pretrained text-to-image models to multimodal understanding and generation. Decodable text latents and causal blockwise flow matching enable a shared backbone to generate both text and images under a common flow formulation. Experiments across understanding, synthesis, and editing demonstrate its effectiveness, while ablations support the benefits of joint training and text-latent adaptation. Future work will explore more efficient blockwise sampling and adaptive reasoning budgets to improve the accuracy–latency trade-off.

AI USE STATEMENT

We used the LLM for refining the sentences and making them grammatically correct. We used the AI agent for optimizing our code and running experiments faster.

REPRODUCIBILTY STATEMENT

We ensure reproducibility by providing all training and evaluation hyperparameters, along with full implementation details, in Appendix C. These materials are sufficient to replicate our experiments. In line with our commitment to open-source research, we will make our codes and model weights publicly available.

ETHICS STATEMENT

To ensure ethical compliance, our training data is carefully selected from public sources and take deliberate measures to minimize potential biases, fully aligning with universal ethical guidelines. We explicitly emphasize that this framework is not intended for misuse in achieving harmful purposes; downstream users are encouraged to adhere to ethical principles when applying the technology. Additionally, all authors declare no conflicts of interest related to this work.

REFERENCES

Manoj Acharya, Kushal Kafle, and Christopher Kanan. TallyQA: Answering complex counting questions. *arXiv preprint arXiv:1810.12440*, 2018.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. LLaVA-OneVision-1.5: Fully open framework for democratized multimodal training, 2025.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Kaichen Zhang, Wenkang Zhang, Zheng Cheng, Nansen Zhang, Chunsheng Wu, Chunjiang Ge, Zimin Ran, Dehua Song, Chunyuan Li, Shikun Feng, Ming Hu, Zhangquan Chen, Junbo Niu, Bo Li, Ziyong Feng, Ziwei Liu, Zongyuan Ge, and Jiankang Deng. LLaVA-OneVision-2: Towards next-generation perceptual intelligence, 2026.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Marianne Arriola, Aaron Gokaslan, Justin Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sahoo, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. In *International Conference on Learning Representations*, volume 2025, pp. 50726–50753, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025.

Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, pp. 1692–1717. PMLR, 2023.

Eric Tillmann Bill, Enis Simsar, Alessio Tonioni, and Thomas Hofmann. Fullflow: Upgrading text-to-image flow matching models for bidirectional vision–language generation, 2026.

Black Forest Labs. FLUX.2: Frontier visual intelligence, 2025.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Yeonwoo Cha, Semin Kim, Jinhyeon Kwon, and Seunghoon Hong. Flowbind: Efficient any-to-any generation with bidirectional flows. In *International Conference on Learning Representations*, volume 2026, pp. 60231–60260, 2026.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Zhihong Chen, Xuehai Bai, Yang Shi, Chaoyou Fu, Huanyu Zhang, Haotian Wang, Xiaoyan Sun, Zhang Zhang, Liang Wang, Yuanxing Zhang, et al. Opengpt-4o-image: A comprehensive dataset for advanced image generation and editing. *arXiv preprint arXiv:2509.24900*, 2025.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. *arXiv preprint arXiv:2409.17146*, 2024.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2025a.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, et al. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *Advances in Neural Information Processing Systems*, 38, 2025b.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T. Q. Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, pp. 133345–133385. Curran Associates, Inc. doi: 10.52202/079017-4239.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Hongcan Guo, Qinyu Zhao, Yian Zhao, Shen Nie, Rui Zhu, Qiushan Guo, Feng Wang, Tao Yang, Hengshuang Zhao, Guoqiang Wei, and Yan Zeng. Continuous latent diffusion language model, 2026.

Jiawei Guo, Tianyu Zheng, Yizhi Li, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Graham Neubig, Wenhu Chen, and Xiang Yue. MAmmoTH-VL: Eliciting multimodal reasoning with instruction tuning at scale. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13869–13920. Association for Computational Linguistics, 2025.

Jonathan Heek, Emiel Hoogeboom, Thomas Mensink, and Tim Salimans. Unified latents (ul): How to train your latents. *arXiv preprint arXiv:2602.17270*, 2026.

Keya Hu, Linlu Qiu, Yiyang Lu, Hanhong Zhao, Tianhong Li, Yoon Kim, Jacob Andreas, and Kaiming He. ELF: Embedded language flows, 2026.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019.

Inclusion AI, Tiwei Bie, Haoxing Chen, Tieyuan Chen, Zhenglin Cheng, Long Cui, Kai Gan, Zhicheng Huang, Zhenzhong Lan, Haoquan Li, Jianguo Li, Tao Lin, Qi Qin, Hongjun Wang, Xiaomei Wang, Haoyuan Wu, Yi Xin, and Junbo Zhao. LLaDA2.0-Uni: Unifying multimodal understanding and generation with diffusion large language model, 2026.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research (TMLR)*, 2024.

Jiaxiu Jiang, Jingjing Ren, Wenbo Li, Bo Wang, Haoze Sun, Yijun Yang, Jianhui Liu, Yanbing Zhang, Shenghe Zheng, Yuan Zhang, Haoyang Huang, Nan Duan, and Wangmeng Zuo. Textldm: Language modeling with continuous latent diffusion, 2026.

Jaehyeong Jo and Sung Ju Hwang. Continuous diffusion model for language modeling. In *Advances in Neural Information Processing Systems*, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025a.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Zichun Liao, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. Omniflow: Any-to-any generation with multi-modal rectified flows. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13178–13188. IEEE, 2025b.

Shufan Li, Jiuxiang Gu, Kangning Liu, Zhe Lin, Zijun Wei, Aditya Grover, and Jason Kuen. Lavida-o: Elastic large masked diffusion models for unified multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2026, pp. 50571–50603, 2026.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-LM improves controllable text generation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4328–4343, 2022.

Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, Yuyin Zhou, and Cihang Xie. What if we recaption billions of web images with LLaMA-3? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35957–35976. PMLR, 2025c.

Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. Dual diffusion for unified image generation and understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2779–2790, June 2025d.

Jiajun Liang, Yucheng Liao, Yukang Cao, Jiazhe Wei, Ken Li, Wende Tan, Jiankun Zhang, ZY Cui, Jingkang Yang, Liucheng Guo, Shiqi Yang, B. Yang, Caifeng Shan, Ziwei Liu, and Chenyang Si. AURORA-LM: Autoencoding unified representation for continuous-latent diffusion language modeling, 2026.

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

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.

Justin Lovelace, Varsha Kishore, Chao Wan, Eliot Shekhtman, and Kilian Q. Weinberger. Latent diffusion for language generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 56998–57025, 2023.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024.

Xu Ma, Yitian Zhang, Qihua Dong, and Yun Fu. Fine-T2I: An open, large-scale, and diverse dataset for high-quality T2I fine-tuning, 2026.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7739–7751. IEEE, 2025.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, et al. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 38405–38453, 2025.

Yusu Qian, Eli Bocek-Rivele, Liangchen Song, Jialing Tong, Yinfei Yang, Jiasen Lu, Wenze Hu, and Zhe Gan. Pico-Banana-400K: A large-scale dataset for text-guided image editing, 2025.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 25278–25294, 2022.

Tao Shen, Xin Wan, Taicai Chen, Rui Zhang, Junwen Pan, Dawei Lu, Fanding Lei, Zhilin Lu, Yunfei Yang, Chen Cheng, et al. Mammothmoda2: A unified ar-diffusion framework for multimodal understanding and generation. *arXiv preprint arXiv:2511.18262*, 2025.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu Edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8871–8879, 2024.

Qingyu Shi, Jinbin Bai, Zhuoran Zhao, Wenhao Chai, Kaidong Yu, Jianzong Wu, Shuangyong Song, Yunhai Tong, Xiangtai Li, Xuelong Li, and Shuicheng Yan. Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model, 2025.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. *arXiv preprint arXiv:1904.08920*, 2019.

Onkar Susladkar, Tushar Prakash, Gayatri Deshmukh, Kiet A. Nguyen, Jiaxun Zhang, Adheesh Juvekar, Tianshu Bao, Lin Chai, Sparsh Mittal, Inderjit S. Dhillon, and Ismini Lourentzou. Best of both worlds: Multimodal reasoning and generation via unified discrete flow matching, 2026.

Alexander Swerdlow, Mihir Prabhudesai, Siddharth Gandhi, Deepak Pathak, and Katerina Fragkiadaki. Unified multimodal discrete diffusion. *arXiv preprint arXiv:2503.20853*, 2025.

Alex Jinpeng Wang, Dongxing Mao, Jiawei Zhang, Weiming Han, Zhuobai Dong, Linjie Li, Yiqi Lin, Zhengyuan Yang, Libo Qin, Fuwei Zhang, Lijuan Wang, and Min Li. TextAtlas5M: A large-scale dataset for dense text image generation. *arXiv preprint arXiv:2502.07870*, 2025a.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *International Conference on Learning Representations*, volume 2025, pp. 62624–62650, 2025b.

Jin Wang, Yao Lai, Aoxue Li, Shifeng Zhang, Jiacheng Sun, Ning Kang, Chengyue Wu, Zhenguo Li, and Ping Luo. Fudoki: Discrete flow-based unified understanding and generation via kinetic-optimal velocities. *Advances in Neural Information Processing Systems*, 38:81402–81440, 2025c.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need, 2024.

Luis Wiedmann, Orr Zohar, Amir Mahla, Xiaohan Wang, Rui Li, Thibaud Frere, Leandro von Werra, Aritra Roy Gosthipaty, and Andrés Marafioti. FineVision: Open data is all you need, 2025.

Chenyuan Wu, Jiahao Wang, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, et al. Omnigen2: Towards instruction-aligned multimodal generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21964–21975, 2026.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *Advances in Neural Information Processing Systems*, 38:47490–47518, 2025.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *Advances in Neural Information Processing Systems*, 38:138867–138907, 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. *Advances in Neural Information Processing Systems*, 38, 2025.

Kyeongmin Yeo, Yunhong Min, and Minhyuk Sung. MUNI: Multimodal unified latent diffusion for coherent any-to-any generation, 2026.

Junchao Yi, Rui Zhao, Jiahao Tang, Weixian Lei, Linjie Li, Qisheng Su, Zhengyuan Yang, Lijuan Wang, Xiaofeng Zhu, and Alex Jinpeng Wang. Flowinone: Unifying multimodal generation as image-in, image-out flow matching. In *European Conference on Computer Vision*, pp. 240–258. Springer, 2026.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. AnyEdit: Mastering unified high-quality image editing for any idea. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26125–26135, 2025.

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

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. UltraEdit: Instruction-based fine-grained image editing at scale. In *Advances in Neural Information Processing Systems*, volume 37, pp. 3058–3093, 2024.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, volume 2026, pp. 35791–35820, 2026a.

Guangting Zheng, Yiyuan Zhang, Tao Yang, Yunpeng Chen, Rui Zhu, Jiajun Deng, and Yanyong Zhang. Genfirst: Generation before reconstruction for stable end-to-end latent generative modeling. *arXiv preprint*, 2026b.

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. In *International Conference on Learning Representations*, 2025.

APPENDIX

A Related Work **17**

B Inference and Decoding **18**

B.1 Text-to-Image Generation 18

B.2 Image-Conditioned Text Generation 19

B.3 Reasoning-Conditioned Image Editing 19

C Implementation Details **20**

C.1 Training 20

C.2 Evaluation 22

D Training Data **22**

E Generation Helps Understanding **24**

F understanding help generation **25**

G Causal Blockwise Flow Matching: Training and Inference **25**

G.1 Visualizing Image–Text Affinities 27

H Additional Qualitative and Quntitative Analysis **27**

H.1 Fine-Grained Evaluation of Text-to-Image Generation 27

H.2 Visual Representations for Generation and Understanding 29

H.3 OCR-Centric Visual Understanding 29

H.4 Qualitative Examples of Long-Form Visual Reasoning 30

H.5 Ablation on text autoencoders. 30

H.6 Additional Qualitative Comparisons for Generation and editing . . . 31

I Additional Qualitative Results on Image-to-Text and Text-to-Image Generation **32**

I.1 Additional visual results Multi-Image Understanding 32

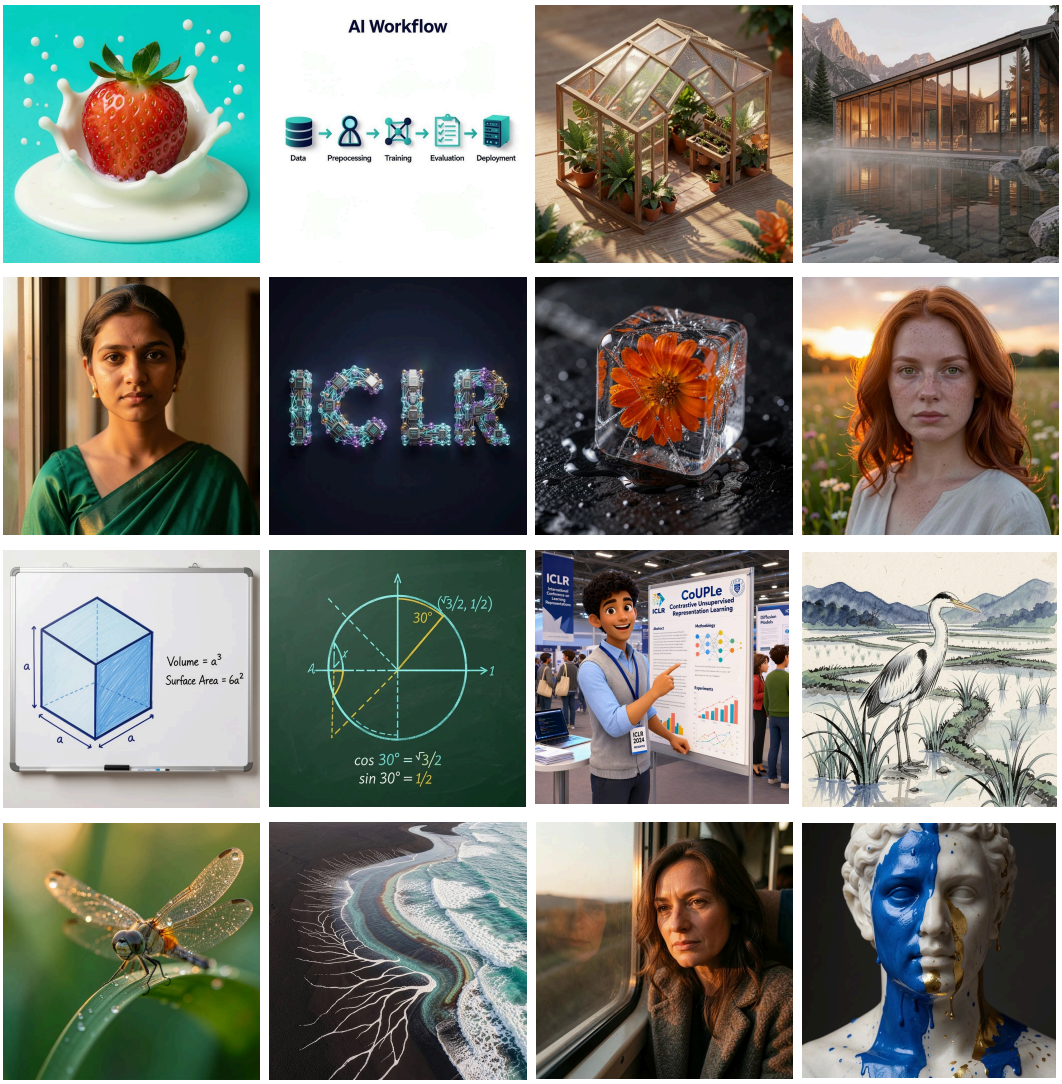


Figure 11: **Text-to-image generation results with CoUPLe**

A Related Work

Unified multimodal generative models. Existing unified models adopt different combinations of state spaces and generative objectives. Transfusion, JanusFlow, and BAGEL combine autoregressive language modeling with continuous image diffusion or flow matching (Zhou et al., 2025; Ma et al., 2025; Deng et al., 2025). D-DiT and FullFlow instead pair continuous image generation with discrete text diffusion or insertion processes (Li et al., 2025d; Bill et al., 2026). Discrete multimodal priors model text and quantized visual tokens through next-token prediction, as in Chameleon and Emu3 (Chameleon Team, 2024; Wang et al., 2024), or masked diffusion, as in UniDisc, MUDDiT, MMaDA, and LaViDa-O (Swerdlow et al., 2025; Shi et al., 2025; Yang et al., 2025; Li et al., 2026). UniDFlow and FUDOKI adopt discrete flow matching for multimodal generation (Gat et al.; Susladkar et al., 2026; Wang et al., 2025c). These categorical processes remain discrete in state even when parameterized in continuous time. **CoUPLe** applies continuous velocity regression to both text and image targets, preserving the pretrained image generator's continuous latent interface.

Continuous language and multimodal generation. Flow matching and rectified flow provide velocity-regression formulations for continuous generative modeling (Lipman et al., 2023; Liu et al., 2023). Diffusion-LM learns continuous token embeddings for iterative denoising (Li et al., 2022),

while Latent Diffusion for Language Generation models compact autoencoded representations using a pretrained language encoder–decoder (Lovelace et al., 2023). Subsequent approaches explore distinct representations: RDLM incorporates categorical-distribution geometry (Jo & Hwang, 2025), ELF performs flow matching in embedding space (Hu et al., 2026), and CoLA-DLM, TextLDM, and AURORA-LM learn decodable continuous text latents (Guo et al., 2026; Jiang et al., 2026; Liang et al., 2026). For multimodal generation, UniDiffuser models joint and conditional distributions through continuous diffusion (Bao et al., 2023), and OmniFlow extends rectified flows to multiple modalities (Li et al., 2025b). MUNI jointly learns a shared flow prior with subset-conditioned encoders and modality-specific decoders (Yeo et al., 2026). These studies establish continuous representations as viable generative interfaces; **CoUPLe** investigates their use for instruction-conditioned understanding, reasoning, and image synthesis within an adapted visual flow backbone.

Causal and blockwise generative priors. Block Diffusion combines autoregressive factorization across blocks with discrete diffusion within each block (Arriola et al., 2025). MMaDA and LLaDA2.0-Uni also employ blockwise text generation in multimodal discrete diffusion systems (Yang et al., 2025; Inclusion AI et al., 2026). In continuous language modeling, CoLA-DLM and AURORA-LM generate latent blocks conditioned on preceding context (Guo et al., 2026; Liang et al., 2026). GenFirst further combines blockwise continuous text flow matching with bidirectional text–image generation (Zheng et al., 2026b). These formulations establish that causal factorization can operate over continuous targets. **CoUPLe** incorporates this structure into pretrained visual flow backbones: text blocks condition on multimodal inputs and previously generated blocks, while image latents form a single target block. Evaluation extends to instruction-conditioned understanding, reasoning, and editing.

Representation learning and adaptation of pretrained components. REPA improves diffusion training by aligning denoiser features with pretrained visual representations (Yu et al., 2024), while representation autoencoders construct decodable latent spaces from semantic encoders (Zheng et al., 2026a). Unified Latents jointly optimizes latent representations with a diffusion prior and decoder (Heek et al., 2026). GenFirst stabilizes joint latent learning through entropy regularization and a generation-before-reconstruction schedule, followed by training a fresh prior on the adapted latent spaces (Zheng et al., 2026b). Reusing pretrained specialists is likewise established by D-DiT and FullFlow (Li et al., 2025d; Bill et al., 2026). **CoUPLe** adapts an existing T2I flow backbone through text-conditioning alignment, reference-regularized Text VAE optimization, and semantic visual conditioning. The original image VAE is retained for synthesis, while the adapted text representation supports continuous generation and decoding for multimodal understanding.

B Inference and Decoding

We perform inference with the trained shared velocity network and fixed modality-specific encoders, decoders, and projections. Text-to-image generation (T2I), image-conditioned text generation (TI2T), and instruction-guided editing (TI2I) use the same continuous flow formulation, with task-dependent conditioning sequences and target latent spaces. All benchmark inference is performed using eight NVIDIA H100 GPUs in BF16 precision.

B.1 Text-to-Image Generation

Given a prompt X_{T}, we construct the text condition using the frozen reference Text-VAE encoder and the learned connector:

$$\widetilde{\mathbf{z}}_{\mathrm{T}} = C_{\psi}(\boldsymbol{\mu}_{\phi_{\mathrm{ref}}}(X_{\mathrm{T}}))\,. \tag{6}$$

The reference posterior mean provides the deterministic text representation, while the connector maps it to the conditioning interface of the pretrained T2I backbone. We initialize an image target from Gaussian noise with the latent shape required by the requested output resolution. The prompt condition remains fixed while the complete image latent is generated as a single target block using 35 integration steps and image CFG. Writing $\mathcal{S}_{\theta}^{m}$ for the numerical sampler, the resulting image is

$$\widehat{\mathbf{z}}_{\mathrm{I}} = \mathcal{S}_{\theta}^{\mathrm{I}}\left(\boldsymbol{\epsilon}_{\mathrm{I}}; \widetilde{\mathbf{z}}_{\mathrm{T}}\right), \qquad \widehat{X}_{\mathrm{I}} = D_{\mathrm{I}}^{\mathrm{VAE}}\left(\widehat{\mathbf{z}}_{\mathrm{I}}\right). \tag{7}$$

Here, $D_{\mathrm{I}}^{\mathrm{VAE}}$ denotes the pretrained image-VAE decoding interface, including any inverse latent normalization required by the checkpoint. The semantic image encoder is not required for this task.

B.2 IMAGE-CONDITIONED TEXT GENERATION

Given an image X_{I} and an instruction X_{Q}, we extract semantic image features $\mathbf{h}_{\mathrm{I}} = E_{\mathrm{I}}^{\mathrm{sem}}(X_{\mathrm{I}})$ and encode the instruction through the frozen, adapted Text-VAE encoder q_ϕ to obtain $\mathbf{z}_{\mathrm{Q}}$. The instruction uses the same latent convention as clean instruction conditions during training. The external condition is

$$\mathbf{c}_{\mathrm{U}} = [\mathbf{h}_{\mathrm{I}}; \mathbf{z}_{\mathrm{Q}}]. \tag{8}$$

These representations enter the backbone through their respective projections. The T2I conditioning connector is not used to construct $\mathbf{z}_{\mathrm{Q}}$. We generate the response as a sequence of latent blocks, with $C = 32$ latent positions per block. At block b, we sample fresh Gaussian noise $\boldsymbol{\epsilon}^{(b)}$ and condition the text sampler on the image, the instruction, and all previously completed response latents:

$$\widehat{\mathbf{z}}_{\mathrm{Y}}^{(b)} = \mathcal{S}_\theta^{\mathrm{T}}\left(\boldsymbol{\epsilon}^{(b)}; [\mathbf{c}_{\mathrm{U}}; \widehat{\mathbf{z}}_{\mathrm{Y}}^{(<b)}]\right). \tag{9}$$

Each block uses 20 integration steps and text CFG. The current block can attend to the clean external condition, its completed latent prefix, and its own positions. Its positions are updated in parallel during each integration step. Previously generated blocks remain fixed and are not noised again. At the end of sampling, the completed block is appended to the latent prefix for the next block.

Following the blockwise decoding procedure described in the main method, we decode each completed block using the Text-VAE decoder p_ω at temperature $\tau = 0.95$. Generation terminates when the decoded output first contains the end-of-sequence marker; content after that marker is discarded. When a response contains reasoning, the rationale and final answer belong to the same causally generated text sequence. Block size counts latent positions and does not, without a specified tokenizer mapping, equal the number of decoded text tokens. The generated target latents are supplied directly to the Text-VAE decoding interface.

B.3 REASONING-CONDITIONED IMAGE EDITING

Given a source image $X_{\mathrm{I}}^{\mathrm{src}}$ and an editing instruction X_{Q}, we encode the source image using both the semantic encoder and the pretrained image-VAE encoder:

$$\mathbf{h}_{\mathrm{I}}^{\mathrm{src}} = E_{\mathrm{I}}^{\mathrm{sem}}\left(X_{\mathrm{I}}^{\mathrm{src}}\right), \qquad \mathbf{z}_{\mathrm{I}}^{\mathrm{src}} = E_{\mathrm{I}}^{\mathrm{VAE}}\left(X_{\mathrm{I}}^{\mathrm{src}}\right). \tag{10}$$

The semantic features provide context for interpreting the requested edit, while the VAE latents represent the source image in the same latent space as the edited-image target. Both source-image representations remain clean throughout rationale generation and image sampling. The instruction is represented through both text interfaces: the adapted encoder provides $\mathbf{z}_{\mathrm{Q}}$, while the frozen reference encoder and connector provide $\widetilde{\mathbf{z}}_{\mathrm{Q}} = C_\psi(\mu_{\phi_{\mathrm{ref}}}(X_{\mathrm{Q}}))$. The clean editing condition is therefore

$$\mathbf{c}_{\mathrm{E}} = [\mathbf{h}_{\mathrm{I}}^{\mathrm{src}}; \mathbf{z}_{\mathrm{I}}^{\mathrm{src}}; \mathbf{z}_{\mathrm{Q}}; \widetilde{\mathbf{z}}_{\mathrm{Q}}]. \tag{11}$$

This notation collects the two source-image representations and the two instruction representations that form the clean Stage-III context.

We first generate the rationale using the text sampler in Eq. (9), replacing $\mathbf{c}_{\mathrm{U}}$ with $\mathbf{c}_{\mathrm{E}}$. After the rationale terminates, its completed latent sequence $\widehat{\mathbf{z}}_{\mathrm{R}}$ is retained as an additional clean condition. We then initialize the edited-image target from Gaussian noise and sample

$$\widehat{\mathbf{z}}_{\mathrm{I}}^{\mathrm{tgt}} = \mathcal{S}_\theta^{\mathrm{I}}(\boldsymbol{\epsilon}_{\mathrm{I}}; [\mathbf{c}_{\mathrm{E}}; \widehat{\mathbf{z}}_{\mathrm{R}}]), \qquad \widehat{X}_{\mathrm{I}}^{\mathrm{tgt}} = D_{\mathrm{I}}^{\mathrm{VAE}}\left(\widehat{\mathbf{z}}_{\mathrm{I}}^{\mathrm{tgt}}\right). \tag{12}$$

The rationale uses the text sampling settings above, and the edited-image target uses 35 integration steps with image CFG. During image sampling, the semantic source-image features, source VAE latents, instruction representations, and completed rationale latents remain fixed. The image target attends to the completed rationale; the rationale is generated before the image target and cannot depend on its future sampled content. Consistent with the latent conditioning in the Stage-III objective, the generated rationale latents are passed directly to the image sampler. Decoded rationale text provides a textual output and a termination signal; it is not decoded and re-encoded to construct a replacement image condition. The edited image is generated in the original image-VAE target space and decoded using its corresponding decoder.

Classifier-free guidance. At each integration step, we contrast the velocity prediction under the full condition c^+ with a reference prediction under c^-:

$$\widehat{v}_\theta(z_t, t; c^+, c^-) = v_\theta(z_t, t; c^-) + w_m\left[v_\theta(z_t, t; c^+) - v_\theta(z_t, t; c^-)\right], \tag{13}$$

where $m \in \{\mathrm{I}, \mathrm{T}\}$ denotes the target modality. Both predictions use the same noisy target z_t and flow time t. We use $w_{\mathrm{I}} = 6.0$ for image generation and editing, and $w_{\mathrm{T}} = 4.0$ for text generation; $w_m = 1$ recovers the ordinary conditional prediction. The guided velocity is integrated from $t = 1$ to $t = 0$. Below, $\varnothing$ denotes the corresponding null or omitted conditioning input, using the representation and masking convention supported during training.

Guidance for image generation. For text-to-image synthesis, the full branch conditions on the input prompt, while the reference branch uses an empty prompt or a supplied negative prompt. Let X_{T}^{-} denote the negative prompt when available and the empty prompt otherwise. The two conditions are

$$c_{\mathrm{I}}^{+} = \widetilde{z}_{\mathrm{T}} = C_{\psi}(\mu_{\phi_{\mathrm{ref}}}(X_{\mathrm{T}})), \qquad c_{\mathrm{I}}^{-} = \widetilde{z}_{\mathrm{T}}^{-} = C_{\psi}\big(\mu_{\phi_{\mathrm{ref}}}(X_{\mathrm{T}}^{-})\big). \tag{14}$$

We apply Eq. (13) to the current image latent using these conditions. Both prompts pass through the same frozen reference encoder and learned connector.

Guidance for text generation. For response block b, the full branch conditions on the image, instruction, and previously generated response blocks. The reference branch retains the instruction but removes both the image condition and the response prefix:

$$c_{\mathrm{T}}^{+,b} = \left[h_{\mathrm{I}};\, z_{\mathrm{Q}};\, \widehat{z}_{\mathrm{Y}}^{(<b)}\right], \qquad c_{\mathrm{T}}^{-,b} = [\varnothing_{\mathrm{I}};\, z_{\mathrm{Q}};\, \varnothing_{\mathrm{P}}]. \tag{15}$$

Here, $\varnothing_{\mathrm{P}}$ indicates that preceding response blocks are inaccessible to the reference prediction. Eq. (13) is evaluated on the same current noisy block $z_{\mathrm{Y},t}^{(b)}$ in both branches. Previously completed blocks remain fixed and available to the full branch throughout sampling.

Guidance for image editing. After generating the rationale $\widehat{z}_{\mathrm{R}}$, we guide image sampling by contrasting the complete editing condition with a reference branch that removes the rationale and replaces the editing instruction with empty or negative text:

$$c_{\mathrm{E}}^{+} = [h_{\mathrm{I}}^{\mathrm{src}};\, z_{\mathrm{I}}^{\mathrm{src}};\, z_{\mathrm{Q}};\, \widetilde{z}_{\mathrm{Q}};\, \widehat{z}_{\mathrm{R}}], \qquad c_{\mathrm{E}}^{-} = \left[h_{\mathrm{I}}^{\mathrm{src}};\, z_{\mathrm{I}}^{\mathrm{src}};\, z_{\mathrm{Q}}^{-};\, \widetilde{z}_{\mathrm{Q}}^{-};\, \varnothing_{\mathrm{R}}\right]. \tag{16}$$

The representations z_{Q}^{-} and $\widetilde{z}_{\mathrm{Q}}^{-}$ encode the same reference text through the adapted Text-VAE encoder and the frozen-reference connector interface, respectively. Both source-image representations are retained in both branches. We apply Eq. (13) to the noisy edited-image latent, with the sampled rationale supplied directly as a latent condition only to the full branch.

C Implementation Details

C.1 Training

Software and distributed training. We implement CoUPLe in PyTorch 2.11 with CUDA 13.0. We use the Hugging Face `Diffusers` library for flow-scheduling and model utilities, and `Accelerate` for distributed execution. Unless stated otherwise, all stages use bfloat16 (`BF16`) mixed-precision training.

Stage I: Bidirectional Pretraining. We initialize the shared velocity network from the corresponding pretrained text-to-image backbone and jointly optimize the image- and text-generation directions for $25{,}000$ parameter-update steps. We use AdamW with a peak learning rate of 1×10^{-6}. The learning rate is linearly warmed up for the first $4{,}000$ updates and subsequently decayed using a cosine-annealing schedule. Training is distributed across five 8-GPU nodes, totaling 40 NVIDIA H100 GPUs with 80 GB of memory per GPU. Each GPU processes a batch of 16 paired examples, and gradients are accumulated over 4 micro-steps.

Consistent with the continuous flow-matching formulation, training times are sampled over $t \in [0, 1]$, while the implementation scheduler is parameterized using $1{,}000$ nominal flow-time levels. We use modality-specific scheduler shifts of 12 and 10 for the image- and text-target flows, respectively. We further employ a progressive-resolution curriculum, increasing the training resolution from 256×256 to 1024×1024 pixels. Distributed training is performed using PyTorch FSDP2 with `BF16` precision.

Stage II: Instruction Tuning. Starting from the Stage-I checkpoint, we fine-tune CoUPLe for image-conditioned instruction following and multimodal reasoning for $15{,}000$ parameter-update steps.

Table 5: **Joint bidirectional pretraining configuration.** Hyperparameters used during Stage-I image–text flow training.

Hyperparameter	Configuration
Parameter-update steps	25,000
Hardware	5×8 NVIDIA H100 GPUs, 80 GB each (40 GPUs)
Optimizer	AdamW
Peak learning rate	1×10^{-6}
Learning-rate warmup	4,000 updates
Learning-rate schedule	Cosine annealing
Batch size	16 examples per GPU
Gradient accumulation	4 micro-steps
Flow-time levels	1,000
Flow-scheduler shift	12 for images and 10 for text
Resolution curriculum	$256 \times 256 \rightarrow 1024 \times 1024$
Numerical precision	`BF16`
Distributed API	FSDP2

Table 6: **Image-understanding SFT configuration.** Hyperparameters used for multimodal instruction and reasoning fine-tuning.

Hyperparameter	Configuration
Parameter-update steps	15,000
Hardware	3×8 NVIDIA H200, 144 GB (24 GPUs)
Optimizer	AdamW
AdamW coefficients	$\beta_1 = 0.98,\ \beta_2 = 0.95$
Peak learning rate	1×10^{-5}
Learning-rate warmup	1,000 updates
Learning-rate schedule	Linear decay
Batch size	12 examples per GPU
Gradient accumulation	4 micro-steps
Numerical precision	`BF16`
Distributed API	FSDP2
Memory optimization	Gradient checkpointing

Following the causal blockwise formulation, the image and instruction latents remain clean and condition every response block, whereas only the current response block is noised and optimized using the blockwise flow-matching objective. The image autoencoders, Text VAE encoders, Text VAE decoder, and text-conditioning connector remain frozen, while the shared velocity network is optimized.

We use AdamW with $\beta_1 = 0.98$, $\beta_2 = 0.95$, and a peak learning rate of 1×10^{-5}. The learning rate is linearly warmed up for 1,000 updates and subsequently linearly decayed. Training is performed on three 8-GPU nodes, totaling 24 NVIDIA H200 GPUs with 144 GB of memory per GPU. We use a batch size of 12 examples per GPU and accumulate gradients over 4 micro-steps. This stage uses PyTorch FSDP2, `BF16` mixed precision, and gradient checkpointing. The larger-memory H200 GPUs are used because rationale-augmented responses result in longer multimodal sequences and a substantially larger activation-memory footprint.

Stage III: Interleaved Text–Image Generation. We further post-train CoUPLe on interleaved, reasoning-augmented instruction-guided editing data for 10,000 parameter-update steps. In this setting, the source image and instruction provide the clean conditioning context, while the rationale and the target-image latent are jointly optimized with the text- and image-space flow-matching terms of Eq. (5), weighted equally ($\lambda_{\mathrm{III}} = 1$). Training is distributed across two 8-GPU nodes, totaling 16 NVIDIA H100 GPUs with 80 GB of memory per GPU. We use a batch size of 16 examples per GPU and accumulate gradients over 6 micro-steps.

We optimize the model using AdamW with $\beta_1 = 0.99$, $\beta_2 = 0.94$, and a learning rate of 1×10^{-5}. The learning rate is controlled using a cosine-annealing schedule. This stage uses the PyTorch FSDP1 API and `BF16` mixed-precision training.

Table 7: **Reasoning-guided editing SFT configuration.** Hyperparameters used for instruction-based image-editing fine-tuning.

Hyperparameter	Configuration
Parameter-update steps	10,000
Hardware	2×8 NVIDIA H100, 80 GB (16 GPUs)
Optimizer	AdamW
AdamW coefficients	$\beta_1 = 0.99,\ \beta_2 = 0.94$
Learning rate	1×10^{-5}
Learning-rate schedule	Cosine annealing
Target balance λ_{III}	1
Batch size	16 examples per GPU
Gradient accumulation	6 micro-steps
Numerical precision	`BF16`
Distributed API	FSDP1

Table 8: **Evaluation configuration.** Sampling hyperparameters used for image and text generation.

Hyperparameter	Configuration
Hardware	8 NVIDIA H100 GPUs
Numerical precision	`BF16`
Image integration steps	35
Text integration steps	20 per block
Image CFG scale	6.0
Text CFG scale	4.0
Text-decoder temperature	0.95

C.2 Evaluation

All benchmark inference is performed using eight NVIDIA H100 GPUs in `BF16` precision. For image synthesis, we integrate the learned image velocity field using 35 sampling steps and apply classifier-free guidance (CFG) with a scale of $w_{\mathrm{I}} = 6.0$. For text generation, continuous latent blocks are generated sequentially from left to right using 20 flow-integration steps per block and a CFG scale of $w_{\mathrm{T}} = 4.0$. Thus, generating a response containing B latent blocks requires $20B$ flow-integration steps.After all latent blocks have been generated, the completed continuous representation is decoded into discrete text tokens using the Text VAE decoder p_ω with a sampling temperature of $\tau = 0.95$. Evaluation is done on 8×H100 GPU. Inference speed for all models is calculated on 1×A100 40GB.

D Training Data

We construct a stage-specific data mixture that mirrors the optimization objectives introduced in §3. Tables 9 to 11 summarize the composition of each training stage. The statistics in these tables denote our training instances after filtering, sequence construction, and mixture allocation. In addition, we have carefully checked for data leackage between training and evaluation, to ensure that no data is overlapped with each other.

Stage I: Bidirectional Pretraining. We use approximately 305M image–caption pairs for joint image- and text-target flow matching, and 280M pairs selected from Recap-DataComp-1B (Li et al., 2025c). After applying our text-normalization and tokenization pipeline, we retain caption sequences containing 32–600 tokens, with an average length of approximately 200 tokens.

We complement this broad-coverage corpus with 5M pairs sampled from Fine-T2I (Ma et al., 2026), which provides fine-grained compositional, attribute-level, and stylistically controlled supervision. We further include 20M pairs from the LAION-Aesthetics subset of LAION-5B (Schuhmann et al., 2022) to increase visual quality and aesthetic diversity. The resulting mixture covers natural photographs, dense multi-object scenes, scene text, graphic design, and illustration- or animation-style imagery. Each pair contributes to both conditional directions: text-conditioned image flow matching and image-conditioned blockwise text flow matching.

Table 9: **Stage-I joint-pretraining data.** Effective contribution of each source to bidirectional image–text flow training.

Source	Pairs	Primary contribution
Recap-DataComp-1B (Li et al., 2025c)	280M	Broad semantic and visual coverage with re-captioned descriptions.
Fine-T2I (Ma et al., 2026)	5M	Fine-grained composition, attributes, style, and prompt fidelity.
LAION-Aesthetics (Schuhmann et al., 2022)	20M	Aesthetically filtered and visually diverse image–text pairs.
Total	**305M**	Joint image- and text-target pretraining.

Table 10: **Stage-II image-understanding data.** Effective composition of the TI2T instruction-tuning mixture.

Source	Examples	Primary supervision
LLaVA-OneVision (Li et al., 2025a)	3.2M	Single-image VQA, OCR, charts, documents, and visual reasoning.
Merged46M / LLaVA-OneVision-2 (An et al., 2025; Wiedmann et al., 2025; An et al., 2026)	45M	Broad multimodal instructions, grounding, science, code, and mathematics.
MAmmoTH-VL-Instruct-12M (Guo et al., 2025)	12M	Direct and rationale-enriched multimodal instruction responses.
Total	**60.2M**	Image-conditioned blockwise text-flow tuning.

Stage II: Instruction Tuning (TI2T). For image-conditioned text generation, we assemble approximately 60.2M multimodal instruction–response instances. We first use the 3.2M single-image instruction subset released with LLaVA-OneVision (Li et al., 2025a). This collection spans general visual question answering, document and chart understanding, OCR, captioning, mathematical reasoning, and geometric reasoning.

We additionally retain approximately 45M examples from the Merged46M instruction mixture used by LLaVA-OneVision-2 (An et al., 2025; Wiedmann et al., 2025; An et al., 2026). This collection is a broad multimodal instruction mixture obtained by combining LLaVA-OneVision-1.5-Instruct and FineVision. It provides supervision for knowledge-intensive VQA, grounding, counting, science, code, mathematics, charts, tables, OCR, and document understanding.

Finally, we incorporate 12M instruction–response examples from MAmmoTH-VL-Instruct-12M (Guo et al., 2025). Its task-aware rewriting and filtering procedure supplements direct-answer examples with rationale-enriched responses across a diverse range of visual perception and reasoning tasks.

We normalize all sources into a common triplet (X_I, X_Q, X_Y), where X_I is the input image, X_Q is the instruction, and X_Y is the textual target. Direct-answer examples use $X_Y = X_A$. When a verified rationale is available, we serialize the target as $X_Y = X_R \oplus X_A$, such that the rationale precedes the final answer in the same block-causal target sequence. This construction preserves concise perception supervision while exposing the model to longer, multi-step reasoning trajectories.

Stage III: Interleaved Text–Image Generation (TI2I). For image editing, we construct an effective mixture of approximately 11.98M training instances from five complementary sources. We use approximately 0.4M examples from Pico-Banana-400K (Qian et al., 2025), including single-turn, preference, and multi-turn editing examples. We allocate 0.08M constructed training instances derived from OpenGPT-4o-Image (Chen et al., 2025).

We further retain 4.5M pairs from AnyEdit (Yu et al., 2025), covering local, global, camera-movement, implicit, and visual-conditioned operations. The mixture additionally contains 5M examples selected from the Emu Edit training corpus (Sheynin et al., 2024) and 2M examples from UltraEdit (Zhao et al., 2024). UltraEdit increases the diversity of free-form and region-conditioned operations and broadens the linguistic coverage of editing instructions.

Table 11: **Instruction-guided image-editing data.** Contribution of each source to TI2I fine-tuning.

Source	Instances	Primary supervision
Pico-Banana-400K (Qian et al., 2025)	0.4M	Diverse single-turn, preference, Reasoning, and multi-turn editing behavior.
OpenGPT-4o-Image (Chen et al., 2025)	0.08M	Broad single-, Reasoning based Editing and Gen. and multi-turn edit transformations.
AnyEdit (Yu et al., 2025)	4.5M	Local, global, implicit, camera, and visual-conditioned edits.
Emu Edit (Sheynin et al., 2024)	5M	Multitask editing and recognition-to-generation examples.
UltraEdit (Zhao et al., 2024)	2M	Fine-grained Reasoning and linguistically diverse editing instructions.
Total	**11.98M**	Image-target flow-matching fine-tuning.

We augment each editing triplet $(X_I^{\text{src}}, X_Q, X_I^{\text{tgt}})$ with a synthetic textual edit plan X_R generated using the GPT-5.6 API and the prompt in Listing 1. Given the source image and editing instruction, the annotator identifies relevant visual content, specifies the requested modifications, and describes what should remain unchanged. The resulting example is $(X_I^{\text{src}}, X_Q, X_R, X_I^{\text{tgt}})$. During Stage III, X_R supervises blockwise text generation, while the target-image flow is conditioned on the source image, instruction, and clean plan latents.

Listing 1: **Prompt for generating** image-editing plans.

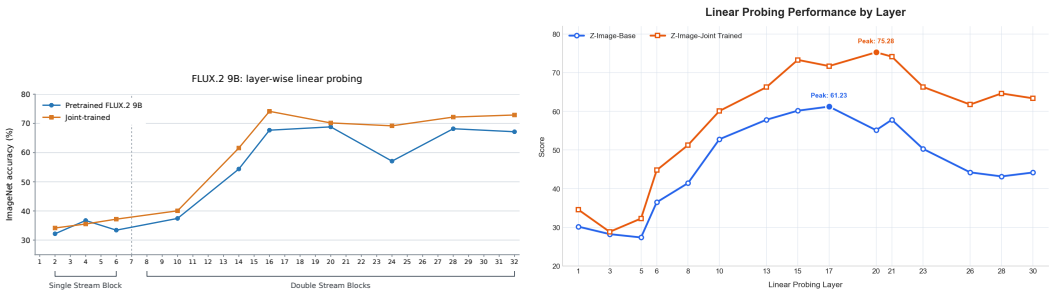
```
You are annotating image-editing examples. Given a source image and an
    editing instruction, write a concise, visually grounded plan for
    performing the requested edit.

Inputs:
- Source image: attached.
- Editing instruction: <EDITING_INSTRUCTION>

Describe:
1. The visible objects, attributes, or regions relevant to the
    instruction.
2. The changes needed to satisfy the instruction, including spatial or
    appearance changes when applicable.
3. The content and properties that should remain unchanged.

Requirements:
- Ground the plan in the source image and instruction.
- Describe only changes required by the instruction.
- Do not invent hidden details or assume unspecified attributes.
- Describe intended operations; do not claim that the edit has already
    been performed.
- Do not include quality scores, evaluation comments, or unrelated
    observations.

Output:
Return only the edit plan as one concise paragraph of 2-8 sentences (
    Based on editing complexity), without headings or bullet points.
```

E Generation Helps Understanding

We investigate whether task-specific image-generation fine-tuning benefits the understanding capabilities of **CoUPLe**. Starting from the trained unified model, we perform additional generative fine-tuning using 36K samples from PixMo-Count (Deitke et al., 2024) for counting and 50K samples from TextAtlas5M (Wang et al., 2025a) for text rendering. We subsequently evaluate visual counting on TallyQA (Acharya et al., 2018) and scene-text question answering on TextVQA (Singh et al., 2019). As shown in Table 12, generative fine-tuning improves TallyQA accuracy by 1.30 and 1.38 percentage points for the Flux-9B and Z-Image variants, respectively, while TextVQA accuracy increases by

Table 12: **Generation-to-understanding transfer.** Understanding accuracy (%) before and after task-specific image-generation fine-tuning, evaluated on TallyQA for counting and TextVQA for scene-text question answering. Δ denotes absolute improvement in percentage points (pp). Higher is better.

Evaluation Dataset	Generation Fine-tuning	Understanding accuracy (%) Ours (Flux-9B)	Ours (Z-Image)
TallyQA	Before	67.78 ± 0.08	70.11 ± 0.06
	After	$\mathbf{69.08 \pm 0.05}$	$\mathbf{71.49 \pm 0.07}$
	Δ (pp)	$+1.30$	$+1.38$
TextVQA	Before	56.67 ± 0.09	59.01 ± 0.08
	After	$\mathbf{58.01 \pm 0.06}$	$\mathbf{61.03 \pm 0.08}$
	Δ (pp)	$+1.34$	$+2.02$



(A) CoUPLe (Flux2-9B, MM-DiT) **(B) CoUPLe (Z-Image, single-stream DiT)**

Figure 12: **Understanding-to-generation transfer: layerwise ImageNet probing before and after joint multimodal training.** (A) Flux2-9B is evaluated after double-stream blocks (DSBs) and single-stream blocks (SSBs). (B) Z-Image is evaluated across its single-stream layers. Blue and orange curves denote the pretrained and jointly trained backbones, respectively. Backbone parameters remain frozen; only a two-layer classification head is trained.

1.34 and 2.02 percentage points. These consistent gains across both backbones support positive generation-to-understanding transfer, suggesting that adapting shared multimodal representations for object cardinality and text rendering also benefits visual counting and text-based question answering.

F UNDERSTANDING HELP GENERATION

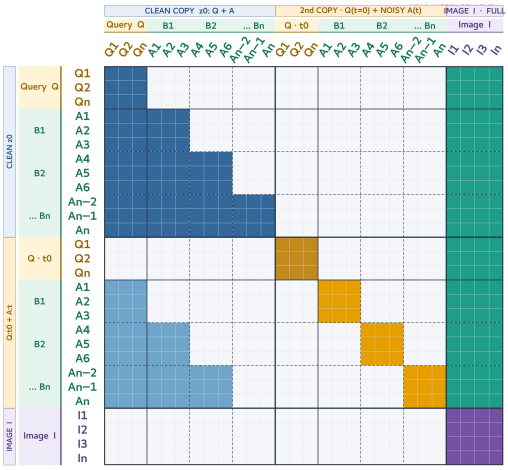
To assess the semantic information accessible in the learned image representations, ImageNet classification probes are trained on features extracted at different depths of CoUPLe (Flux2-9B), CoUPLe (Z-Image), and their pretrained counterparts. The DiT backbone remains frozen, and only a two-layer classification head is optimized. Flux2-9B is probed after double-stream and single-stream blocks (DSBs and SSBs), whereas Z-Image is evaluated across its single-stream layers.

As shown in Figure 12 (A), joint training primarily strengthens the SSB representations of Flux2-9B, increasing the best probing accuracy from approximately 69% to 74%, with smaller and nonuniform changes in the DSBs. For Z-Image, Figure 12 (B) shows improvements at every sampled layer: the best accuracy increases from 61.23% at layer 17 to 75.28% at layer 20, a gain of 14.05 percentage points between the respective maxima. These results suggest that joint multimodal training makes ImageNet category information more accessible to the classification probe across both architectures, although the strength of this information varies with depth.

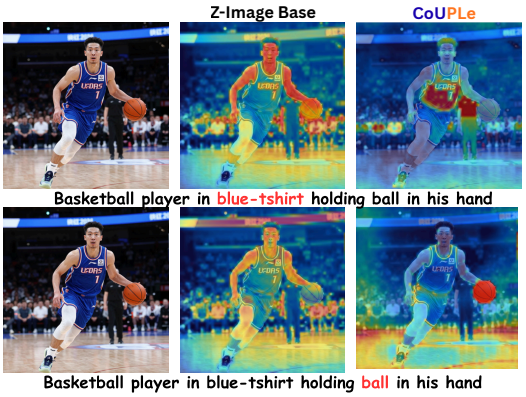
G CAUSAL BLOCKWISE FLOW MATCHING: TRAINING AND INFERENCE

Causal factorization. Given a text target $\mathbf{x}$ of L latent positions, we partition it into $B = \lceil L/C \rceil$ contiguous blocks, $\mathbf{x} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(B)})$, where the final block may contain fewer than C positions. The conditional distribution factorizes as

$$p(\mathbf{x} \mid \mathbf{c}) = \prod_{b=1}^{B} p\Big(\mathbf{x}^{(b)} \mid \mathbf{c}, \mathbf{x}^{(<b)}\Big). \tag{17}$$



(a) Structured attention for doubled-sequence training.



(b) Image–text affinity before and after adaptation.

Figure 13: **Attention structure and cross-modal feature analysis. (a)** The doubled text sequence combines clean and noised answer copies with clean question and image conditions to evaluate all blockwise objectives in one forward pass. The total sequence length is $2L + N_{\mathrm{I}}$. **(b)** Image–text affinity maps extracted from block 22 at the fifth sampling iteration of Z-Image Base and **CoUPLe** (Z-Image). The highlighted concepts are "blue-t-shirt" and "ball."

Each factor is modeled by conditional flow matching, with the external context and preceding blocks kept clean while the target block is noised. Averaging these conditional flow-matching losses yields the objective in Eq. (1).

Conditional log-likelihood estimation. Following CoLA-DLM (Guo et al., 2026), we combine the block-causal latent prior with the Text VAE decoder. For a fixed block partition and condition c,

$$p_{\theta,\omega}(X_T \mid c) = \int p_\omega(X_T \mid z_T) \prod_{b=1}^{B} p_\theta(z_T^{(b)} \mid c, z_T^{(<b)})\, dz_T. \tag{18}$$

Let $c_b = [c; z_T^{(<b)}]$ remain fixed and solve $\dot{y}_t^{(b)} = v_\theta(y_t^{(b)}, t; c_b)$ with $y_0^{(b)} = z_T^{(b)}$. Under standard flow regularity assumptions,

$$\log p_\theta(z_T \mid c) = \sum_{b=1}^{B} \left[\log \mathcal{N}(y_1^{(b)}; 0, I) + \int_0^1 \nabla_{y_t^{(b)}} \cdot v_\theta(y_t^{(b)}, t; c_b)\, dt \right]. \tag{19}$$

For posterior samples $z_T^{[k]} \sim q_\phi(z_T \mid X_T)$, define

$$\ell_k = \log p_\omega(X_T \mid z_T^{[k]}) + \log p_\theta(z_T^{[k]} \mid c) - \log q_\phi(z_T^{[k]} \mid X_T). \tag{20}$$

Then $\widehat{\mathcal{L}}_{\mathrm{ELBO}} = K^{-1} \sum_{k=1}^{K} \ell_k$ estimates a conditional log-likelihood lower bound.

Proof. Integrating the continuity equation along each conditional ODE and summing the block log-densities gives Eq. equation 19. Assuming sufficient posterior support, importance sampling gives $\mathbb{E}[\exp(\ell_k)] = p_{\theta,\omega}(X_T \mid c)$. Jensen's inequality therefore yields

$$\mathbb{E}\left[\widehat{\mathcal{L}}_{\mathrm{ELBO}}\right] = \mathbb{E}[\ell_k] \le \log \mathbb{E}[\exp(\ell_k)] = \log p_{\theta,\omega}(X_T \mid c). \tag{21}$$

Thus, the estimator targets the ELBO rather than the exact marginal log-likelihood. The flow-matching regression loss is not itself a negative log-likelihood.

Parallel training with a doubled text sequence. Although the distribution factorizes causally across blocks, all B block losses can be evaluated in one forward pass using clean and noised copies of the target sequence. For instruction-conditioned training, let L_{Q} denote the question latent length and $L_{\mathrm{A}} = L$ the answer latent length. We concatenate a clean question–answer copy, a second copy containing the clean question and noised answer blocks, and N_{I} clean image tokens. The resulting sequence has $2(L_{\mathrm{Q}} + L_{\mathrm{A}}) + N_{\mathrm{I}}$ positions. A structured attention mask, illustrated in Figure 13a, ensures that each noisy answer block receives exactly the context permitted by the causal factorization.

Structured attention. Within the clean copy, each answer block attends to the question, image tokens, and preceding or current clean answer blocks, but cannot attend to the noisy stream. Each noisy answer block attends only to the clean question, image tokens, preceding clean answer blocks, and its own noisy positions; its clean counterpart, future answer blocks, and other noisy blocks remain inaccessible. Each question copy attends only within its own question segment and to image tokens, while image tokens attend exclusively to image tokens. These restrictions prevent future answer information from being relayed through attention to clean-prefix or conditioning representations. Attention is bidirectional within each permitted block. Clean answer-prefix latents are detached at the text-encoder output, preventing gradients from reaching the encoder through the prefix. Velocity losses are evaluated only on noisy answer positions and averaged across blocks, implementing Eq. (1) without block-index sampling.

Blockwise inference. At inference, blocks are generated from left to right according to Eq. (17). For each block, we sample fresh Gaussian noise and integrate the velocity field from $t = 1$ to $t = 0$, conditioned on $\mathbf{c}$ and all previously completed blocks. Positions within the current block are refined in parallel, while completed blocks remain fixed. We decode each completed block using p_ω and stop when its decoded output contains an end-of-sequence marker, discarding content after that marker. The number of generated blocks therefore follows the response length. With N integration steps per block, a response containing B blocks requires BN integration steps. For image generation, the complete image latent forms a single target block ($B = 1$), recovering standard conditional flow matching.

G.1 Visualizing Image–Text Affinities

We compare Z-Image Base with **CoUPLe** (Z-Image) after joint multimodal training and supervised fine-tuning. For both models, we extract image and text features from transformer block $\ell = 22$ at the fifth sampling iteration, $k = 5$, providing a consistent layer and iteration for analysis. Let $\mathbf{F}_{\mathrm{I}}^{(22,5)} \in \mathbb{R}^{N_{\mathrm{I}} \times d}$ and $\mathbf{F}_{\mathrm{T}}^{(22,5)} \in \mathbb{R}^{N_{\mathrm{T}} \times d}$ denote the extracted features. Treating image positions as queries and text positions as keys, we compute the cross-modal affinity matrix

$$\mathbf{S}^{(22,5)} = \mathbf{F}_{\mathrm{I}}^{(22,5)} \left(\mathbf{F}_{\mathrm{T}}^{(22,5)} \right)^{\top}. \tag{22}$$

Scores associated with the selected text concept are mapped to the image-token spatial grid and resized to the image resolution. These feature affinities provide a diagnostic visualization rather than the post-mask attention probabilities of the blockwise training mask. As shown in Figure 13b, the adapted model exhibits more concept-specific spatial responses: "blue-t-shirt" produces a pronounced response over the jersey, while "ball" produces a distinct response over the basketball. The base model responds more broadly across the player and surrounding scene. These examples are consistent with more selective image–text alignment following joint training and fine-tuning, although residual background responses remain.

H Additional Qualitative and Quntitative Analysis

H.1 Fine-Grained Evaluation of Text-to-Image Generation

We provide category-level results on GenEval and DPG-Bench in Tables 13 and 14, respectively. GenEval (Ghosh et al., 2023) separates object composition, counting, color specification, spatial positioning, and attribute binding, while DPG-Bench (Hu et al., 2024) evaluates Global, Entity, Attribute, Relation, and Other categories for dense prompts. All scores are higher-is-better. Differences below are absolute score differences; a difference of 0.01 on GenEval corresponds to one percentage point.

GenEval: category-level performance. Both **CoUPLe** variants achieve 1.00 on Two Objects and share the best Colors score of 0.95. CoUPLe (Z-Image) also matches the best Single Object score of 1.00, while **CoUPLe** (Flux2-9B) achieves the highest Counting score of 0.86. Relative to its Flux2-9B initialization, CoUPLe (Flux2-9B) improves five categories and preserves Colors: Position increases from 0.48 to 0.85, Attribute Binding from 0.67 to 0.75, and Counting from 0.83 to 0.86. These results show that the adapted checkpoint retains basic object and color fidelity while improving more demanding compositional constraints.

GenEval: comparison with all unified baselines. Both variants outperform MUDDiT and OmniGen2 in all six categories, with Counting gains of 0.08/0.06 over MUDDiT and 0.22/0.20 over

Table 13: **Category-level GenEval results.** Scores are reported on a $[0, 1]$ scale. Best and second-best distinct scores in each column are shown in **bold** and underlined, respectively, including ties. Attr. Binding denotes attribute binding.

Method	Single Obj.	Two Obj.	Counting	Colors	Position	Attr. Binding
Flux2-9B (Black Forest Labs, 2025)	0.97	0.98	0.83	**0.95**	0.48	0.67
BAGEL (Deng et al., 2025)	0.98	0.95	0.84	**0.95**	0.78	0.77
MUDDiT (Shi et al., 2025)	0.74	0.72	0.78	0.77	0.42	0.56
MammothModa-2 (Shen et al., 2025)	**1.00**	0.97	0.63	0.89	**0.90**	**0.82**
OmniGen2 (Wu et al., 2026)	0.95	0.93	0.64	0.81	0.73	0.74
FUDOKI (Wang et al., 2025c)	0.98	0.95	0.73	0.94	0.88	0.78
LaViDa-O (Li et al., 2026)	0.98	0.98	0.75	0.94	0.84	0.75
CoUPLe (Flux2-9B)	0.99	**1.00**	**0.86**	**0.95**	0.85	0.75
CoUPLe (Z-Image)	**1.00**	**1.00**	0.84	**0.95**	0.84	0.78

Table 14: **Category-level DPG-Bench results.** Higher is better. Best and second-best distinct scores in each column are shown in **bold** and underlined, respectively.

Method	Global	Entity	Attribute	Relation	Other
Flux2-9B (Black Forest Labs, 2025)	80.11	82.42	81.98	82.78	81.21
BAGEL (Deng et al., 2025)	**89.42**	91.43	90.42	92.34	88.78
MUDDiT (Shi et al., 2025)	84.78	79.88	75.57	84.23	78.77
MammothModa-2 (Shen et al., 2025)	81.16	**92.99**	90.16	**94.35**	84.81
OmniGenV2 (Wu et al., 2026)	–	–	86.43	91.23	–
FUDOKI (Wang et al., 2025c)	80.55	89.73	88.05	93.66	78.00
LaViDa-O (Li et al., 2026)	81.28	88.92	88.78	91.04	80.23
CoUPLe (Flux2-9B)	84.44	90.12	90.12	93.38	87.72
CoUPLe (Z-Image)	85.67	92.23	**93.23**	93.92	**90.12**

OmniGen2 for the Flux2-9B/Z-Image variants, respectively. Against BAGEL, each variant improves four categories; Z-Image ties the remaining two, whereas Flux2-9B ties Colors but trails Attribute Binding by 0.02. Compared with MammothModa-2, both variants improve Two Objects and Colors by 0.03 and 0.06, while Counting increases by 0.23/0.21. However, MammothModa-2 remains stronger on Position (0.90) and Attribute Binding (0.82). Both variants also exceed FUDOKI on Single Object, Two Objects, Counting, and Colors, but trail its Position score of 0.88; Z-Image matches its Attribute Binding score, while Flux2-9B trails by 0.03. Relative to LaViDa-O, each variant improves five categories and ties one: Flux2-9B ties Attribute Binding, and Z-Image ties Position. Thus, CoUPLe's strongest advantages lie in object composition and counting, while precise positioning and attribute assignment remain areas for improvement.

DPG-Bench: dense-prompt alignment. **CoUPLe** (Z-Image) ranks first in Attribute (93.23) and Other (90.12), and second in Global (85.67), Entity (92.23), and Relation (93.92). Relative to the pretrained Flux2-9B checkpoint, **CoUPLe** (Flux2-9B) improves Global, Entity, Attribute, Relation, and Other by 4.33, 7.70, 8.14, 10.60, and 6.51 points, respectively. The largest gains occur in relational and attribute-level constraints, complementing the improvements in compositional generation observed on GenEval.

DPG-Bench: comparison with all unified baselines. Against BAGEL, **CoUPLe** (Z-Image) improves Entity, Attribute, Relation, and Other by 0.80, 2.81, 1.58, and 1.34 points, respectively, but trails Global by 3.75 points; the Flux2-9B variant improves only Relation, by 1.04 points. Compared with MammothModa-2, Z-Image improves Global, Attribute, and Other by 4.51, 3.07, and 5.31 points, while trailing Entity and Relation by 0.76 and 0.43 points. Flux2-9B exceeds MammothModa-2 on Global and Other but trails the remaining categories. Z-Image outperforms MUDDiT and FUDOKI in all five categories, whereas Flux2-9B improves four categories against each, with small deficits on Global against MUDDiT (0.34) and Relation against FUDOKI (0.28). Both variants outperform LaViDa-O in every category, with Relation gains of 2.34/2.88 points for Flux2-9B/Z-Image. They also exceed OmniGen2 on both available categories: Attribute improves by 3.69/6.80 points and Relation by 2.15/2.69 points. The remaining OmniGen2 categories are unavailable and

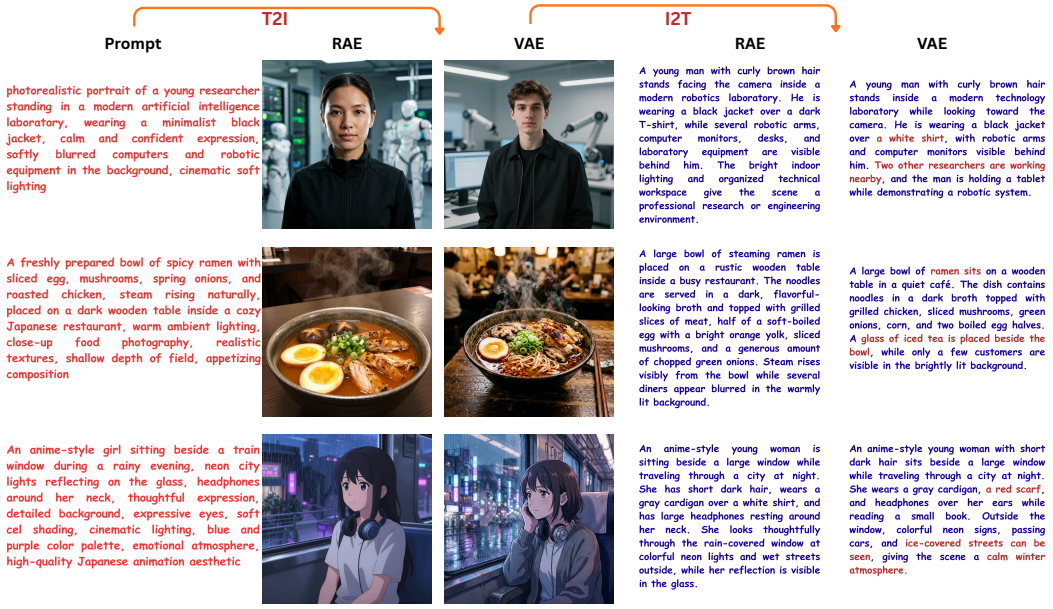


Figure 14: **Qualitative comparison of VAE and RAE visual representations.** Left: text-to-image generation using RAE and VAE representations for the same prompts. Right: image-to-text outputs using RAE and VAE visual conditioning. RAE-conditioned captions more closely match the visible content in these examples, while VAE-conditioned captions introduce unsupported attributes and objects.

are excluded from comparison. Global alignment and selected entity–relation constraints remain gaps to the strongest baselines despite CoUPLe (Z-Image)'s leading Attribute and Other scores.

H.2 Visual Representations for Generation and Understanding

Figure 14 compares VAE and RAE representations for text-to-image generation and image-conditioned captioning. For synthesis, retaining the original VAE preserves the latent space and decoding interface learned by the pretrained T2I backbone. The displayed VAE-based samples exhibit detailed textures, lighting, and background structure across photographic and stylized scenes. For understanding, the RAE's pretrained semantic encoder provides features that support visually grounded descriptions. RAE-conditioned captions more accurately preserve visible attributes and scene content, while VAE-conditioned captions introduce unsupported details, such as additional researchers, an iced-tea glass, and a red scarf. These qualitative examples illustrate the complementary roles of the two representations in CoUPLe: the VAE maintains compatibility with the pretrained visual generation prior, while RAE features provide semantic conditioning for text generation.

H.3 OCR-Centric Visual Understanding

Quantitative comparison. We evaluate **CoUPLe** on OCRBench v2 (Fu et al., 2025b) to assess text recognition, spatial grounding, structured extraction, and reasoning over textual content in images. As reported in Table 15, **CoUPLe** achieves the highest overall score of **57.4**, exceeding MammothModa-2, the strongest overall baseline in the comparison, by **1.3** points. It also attains the highest reported scores in seven of the eight capability categories. Relative to MammothModa-2, CoUPLe improves text recognition by 1.90 points, relation extraction by 2.34 points, and knowledge reasoning by 2.00 points. For element parsing, it reaches 41.64, exceeding the strongest baseline, LaViDa-O, by 2.44 points. The improvements therefore extend beyond recognizing strings to extracting their relationships and recovering structured content. Visual text understanding is comparable to MammothModa-2 (80.24 versus 80.20). Text spotting remains a limitation: although CoUPLe ranks second at 12.23, it trails LaViDa-O by 1.27 points. This contrast indicates that joint localization and transcription remains challenging despite strong performance in extraction and semantic interpretation.

Qualitative comparison. Figure 15 provides complementary examples of question-conditioned visual text understanding. In the mobile-interface example, **CoUPLe** recovers the complete username,

Table 15: **Comparison on OCRBench v2.** Recog., Ref., Spot., Extr., Pars., Calc., Und., and Reas. denote text recognition, text referring, text spotting, relation extraction, element parsing, mathematical calculation, visual text understanding, and knowledge reasoning, respectively. Higher scores are better. Best and second-best results are shown in **bold** and underlined, respectively.

Method	Recog.	Ref.	Spot.	Extr.	Pars.	Calc.	Und.	Reas.	Avg.
OmniGen2 (Wu et al., 2026)	68.80	25.70	1.20	80.20	30.40	38.20	73.20	56.20	46.7
Qwen2.5-VL (Bai et al., 2025)	64.40	38.20	5.70	91.00	37.80	44.20	76.80	62.60	55.7
Show-o2 (Xie et al., 2025)	61.30	36.50	2.40	87.20	33.40	40.70	72.70	65.70	48.7
BAGEL (Deng et al., 2025)	65.80	37.10	3.30	90.50	38.50	41.30	75.20	66.40	52.2
MammothModa-2 (Shen et al., 2025)	68.20	39.50	11.40	92.20	39.10	50.20	80.20	68.10	56.1
FUDOKI (Wang et al., 2025c)	61.20	26.70	0.00	77.50	36.30	43.40	71.10	55.50	46.5
LaViDa-O (Li et al., 2026)	61.20	39.50	**13.50**	79.30	39.20	47.70	75.50	59.30	51.9
CoUPLe	**70.10**	**40.18**	12.23	**94.54**	**41.64**	**50.90**	**80.24**	**70.10**	**57.4**

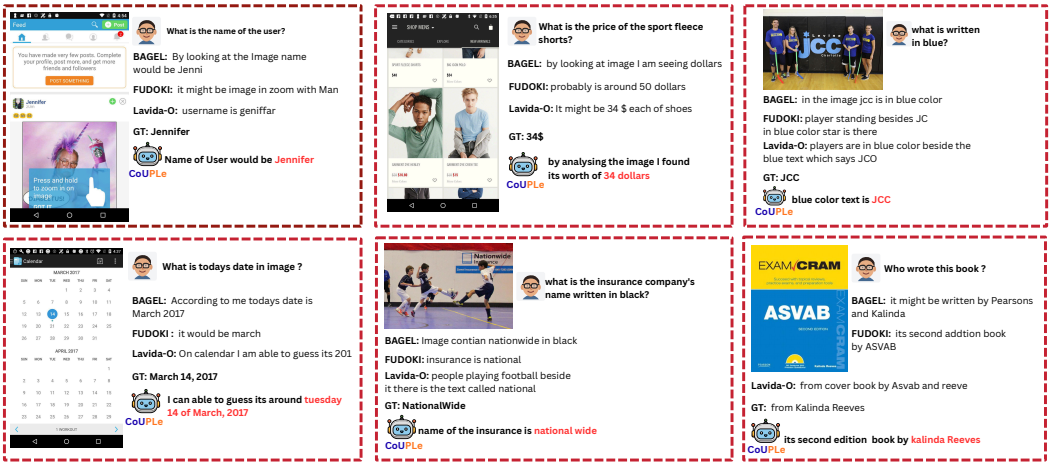


Figure 15: **Qualitative comparison on OCR-centric visual understanding.** Each example presents an input image, a question, model responses, and an annotated answer. The selected examples span mobile interfaces, product listings, scene text, a calendar, and a book cover. The username, calendar, and book-cover examples illustrate differences in transcription completeness, date interpretation, and identification of the requested textual field.

whereas BAGEL truncates the name and LaViDa-O produces an incorrect transcription. In the calendar example, it identifies the full selected date, March 14, 2017, while the compared responses omit the day or produce incomplete date information. In the book-cover example, **CoUPLe** identifies Kalinda Reeves as the author; the compared responses confuse the requested author field with publisher or title information. These examples illustrate the ability to select question-relevant textual evidence from surrounding visual content and preserve the details needed to answer the query.

H.4 Qualitative Examples of Long-Form Visual Reasoning

Figures 17 and 18 illustrate long-form visual reasoning by **CoUPLe** across geometry diagrams, charts, and infographics. The generated responses extract visual quantities and articulate intermediate calculations or geometric relations before producing a final answer. The triangle-similarity example correctly derives $BC = 40\,\mathrm{cm}$, demonstrating a successful multi-step calculation grounded in the diagram. Other examples reveal limitations, including contradictory intermediate deductions and unsupported aggregation of distinct quantities. These qualitative results complement the quantitative understanding evaluation by illustrating both successful reasoning and failure cases, highlighting the need to assess the consistency of generated explanations alongside final-answer accuracy.

H.5 Ablation on Text Autoencoders.

Before adopting the pretrained CoLA-VAE, we trained a T5-based text autoencoder and additionally evaluated a Qwen2.5-VL-based alternative. Our formulation places two requirements on the text

Table 16: Ablation on the text autoencoder. We evaluate multimodal understanding using MMMU and MathVista accuracy (%), and image generation using GenEval and COCO-FID. Best results in **bold**.

Text autoencoder	MMMU ↑	MathVista ↑	GenEval ↑	COCO-FID ↓
T5-AE	62.37	73.01	0.86	8.98
Qwen2.5-VL-AE	55.71	61.78	0.83	8.16
CoLA-VAE	**68.19**	**80.12**	**0.89**	**7.75**

Table 17: **Reasoning ablation for image editing.** Rationale conditioning improves the reported scores on both benchmarks. Best results in **bold**.

Variant	ImgEdit ↑	GEdit-Bench-EN ↑
CoUPLe w/o reasoning	4.31	6.36
CoUPLe	**4.59**	**6.79**
Absolute improvement	+0.28	+0.43

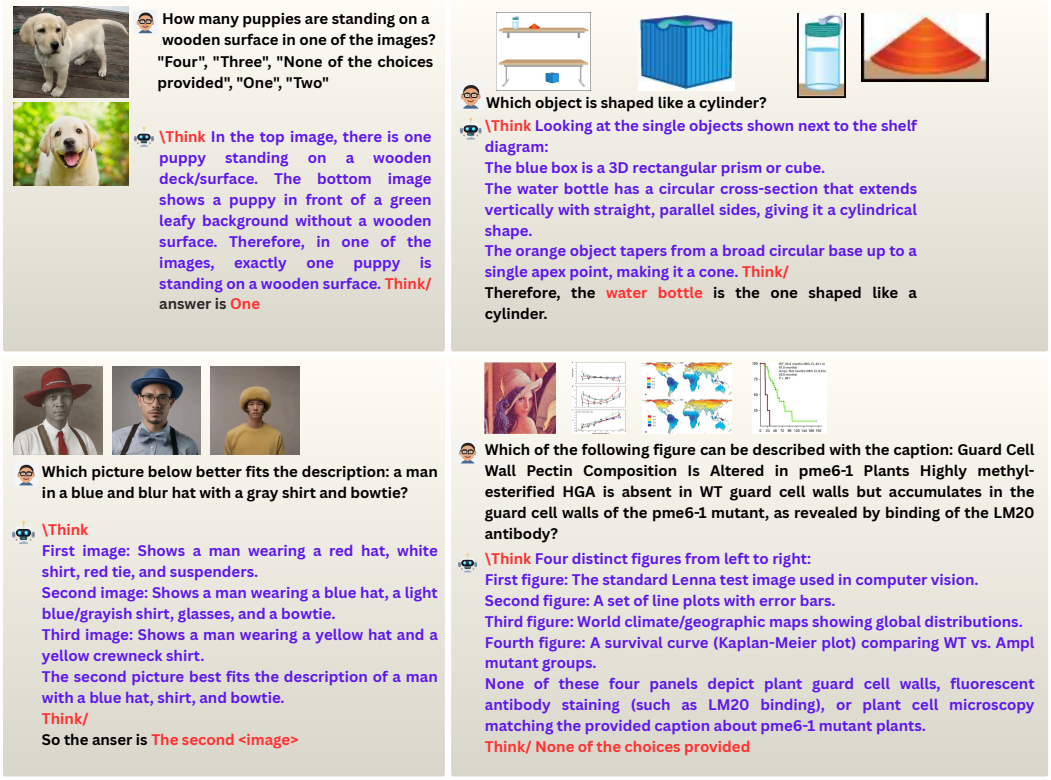


Figure 16: **Qualitative multi-image understanding with CoUPLe.** Each example shows multiple input images, a question, and the generated rationale and answer. The examples cover image-conditioned counting, shape recognition, attribute-based image selection, and scientific caption-to-figure matching, including rejection of mismatched candidates.

interface: providing semantic conditions for image synthesis and defining decodable continuous targets for blockwise flow matching. As shown in Table 16, T5-AE supports effective image conditioning, achieving a GenEval score of 0.86, but underperforms CoLA-VAE on multimodal understanding. Replacing T5-AE with CoLA-VAE improves MMMU and MathVista by 5.82 and 7.11 percentage points, respectively, while also improving image generation. CoLA-VAE achieves the best results across all four metrics, suggesting that representations suitable for semantic conditioning are not necessarily equally effective as continuous generation targets. These findings support our choice of CoLA-VAE, whose target latents are jointly adapted to the shared flow while preserving decodability through reconstruction and reference-posterior regularization.

H.6 Additional Qualitative Comparisons for Generation and Editing

Figures 19 and 20 compare CoUPLe with the evaluated baselines on instruction-guided editing and text-to-image generation. For editing, CoUPLe accurately replaces the sign text with "ICLR 2027," whereas Flux2-9B (Black Forest Labs, 2025) omits the year and OmniGenV2 (Wu et al., 2026), BAGEL (Deng et al., 2025), and MammothModa-2 (Shen et al., 2025) produce incorrect character sequences. In the silver-donut example, CoUPLe produces a reflective appearance while retaining the mask, central decoration, and surrounding hand; Flux2-9B primarily produces a gray fabric

appearance, while MammothModa-2 substantially alters the object's defining features. During blue-tone recoloring, **CoUPLe** also retains the displayed year "2023," which is changed in MammothModa-2's output. For image generation, **CoUPLe** and BAGEL reproduce the complete requested coffee-shop menu, while FUDOKI (Wang et al., 2025c) truncates an entry and OmniGen2 and LaViDa-O (Li et al., 2026) exhibit spelling or legibility errors. The robot example further shows **CoUPLe** integrating mechanical components, trajectory curves, and graphical annotations into the scene, whereas LaViDa-O omits most of the requested analytical overlays. The poster and paper-cut examples additionally illustrate typographic organization and the requested visual styles. These comparisons highlight instruction-specific text fidelity, retention of source content during editing, and composition of multiple prompt attributes within the displayed examples.

Effect of reasoning on image editing. We compare **CoUPLe** with a variant without rationale conditioning. As shown in Table 17, incorporating reasoning improves ImgEdit and GEdit-Bench-EN scores by 0.28 and 0.43 points respectively, suggesting a modest benefit across both benchmarks. Figure 21 provides qualitative examples: rationale-conditioned editing reduces neon illumination when transitioning to daylight and produces a visibly deflated basketball following a puncture.

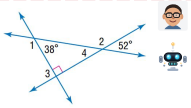
I Additional Qualitative Results on Image-to-Text and Text-to-Image Generation

Image-to-text generation. Figure 22 presents image-conditioned descriptions generated by **CoUPLe** for natural scenes and text-rich imagery. The responses integrate object identities, appearance attributes, spatial relations, and embedded text into coherent descriptions. For example, the street-scene description identifies the left-pointing "ONE WAY" sign and its position above the yellow traffic-signal housings, while the beverage example describes the relative placement of the glass and bottle together with foam, reflections, and background context. The winery and event-poster examples further illustrate the integration of scene description with visible lettering, including the winery name and the poster's event date. These examples illustrate fine-grained visual grounding in descriptive text generation.

Text-to-image generation. Figure 23 presents images generated by **CoUPLe** from detailed prompts specifying visual content, appearance, spatial composition, and embedded text. Across postcards, book covers, infographics, product photographs, and geometric diagrams, the outputs capture multiple requested constraints within a single composition. The postcard combines large "KENTUCKY" lettering with scenic inserts and an aged print appearance, while the book cover renders the requested title against a crimson background. The infographic reproduces the requested title and broad multi-panel organization, and the chalkboard example includes the two block arrangements and numerical annotations. These results illustrate qualitative adherence to dense prompts, although small lettering and fine spatial details remain imperfect.

I.1 Additional visual results Multi-Image Understanding

Figure 16 presents qualitative examples of image-conditioned counting, attribute-based image selection, shape recognition, and caption-to-figure matching, illustrating how the model selects and compares visual evidence across the supplied images.



Find the angle measure of \angle 4.

\Think The diagram shows two intersecting lines. We are given that one angle (marked as 52°) and a right-angled triangle is present within the intersecting lines. One of the angles of this triangle is 38°. We know that the sum of angles in a triangle is 180° and angles on a straight line sum up to 180°. Vertical angles are equal.

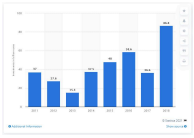
2. **Find the missing angle in the right-angled triangle:** The triangle has angles 38° and 90°. The third angle, which we can call angle 5, can be found using the property that the sum of angles in a triangle is 180°: angle 5 = 180° - 90° - 38° = 52°

3. **Identify angle pairs:** Observe that angle 2 (52°) and angle 5 (52°) are vertical angles. This confirms the consistency of our triangle calculation.

4. **Use the property of linear pairs:** Angles 2 and 3 are a linear pair (they are adjacent and form a straight line), so their sum is 180°: angle 3 = 180° - angle 2 = 180° - 52° = 128°

5. **Find angle 4:** Angles 3 and 4 are vertical angles. Since vertical angles are equal: angle 4 = angle 3 = 128° However this contradicts the triangle information. Angle 3 and this unlabeled angle are supplementary, therefore angle 3 is 180 - 128 = 52 degrees. Since angles 3 and 4 are vertical angles, angle 4 = angle 3 = 52°
Think/

So the answer is 52°



In what year did the amount of foreign direct investment outflow in France fluctuate?

\Think To find the year with the most significant fluctuation in France's foreign direct investment (FDI) outflow, we need to analyze the provided bar chart.

First, we carefully extract the FDI outflow data for each year from the chart: * 2011: 37 billion euros * 2012: 27.6 billion euros * 2013: 15.3 billion euros * 2014: 37.5 billion euros * 2015: 48 billion euros * 2016: 58.6 billion euros * 2017: 36.6 billion euros * 2018: 86.8 billion euros

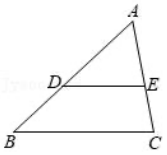
Next, let's calculate the change in FDI outflow from one year to the next. This helps quantify the fluctuations: * 2011-2012: -9.4 billion euros * 2012-2013: -12.3 billion euros * 2013-2014: +22.2 billion euros * 2014-2015: +10.5 billion euros * 2015-2016: +10.6 billion euros * 2016-2017: -22 billion euros * 2017-2018: +50.2 billion euros

Now, we analyze these changes to identify the years with notable fluctuations. A large positive or negative change indicates significant fluctuation. We are looking for the year with the largest absolute change in FDI outflow.

Comparing the absolute values of the year-to-year changes, we find that the largest change occurred between 2016 and 2017 (-22 billion euros), representing a substantial drop in FDI outflow after a period of consistent growth. While the change from 2017 to 2018 is also large (+50.2 billion), it represents a sharp increase after a significant decrease; the decrease in 2017 is a more significant fluctuation in that it represents a marked reversal of trend. Think/

Based on our analysis of the year-to-year changes in FDI outflow, the year with the most significant fluctuation is 2017

Figure 17: **Long-form visual reasoning on geometry and chart questions.**



In triangle ABC, given that DE is parallel to BC, AD:DB = 12:8 and DE = 24cm, what is the length of BC? Choices: A: 32.0 B: 40.0 C: 48.0 D: 72.0

\Think
1. Since DE is parallel to BC, ΔADE and ΔABC are similar triangles. This means that their corresponding angles are equal, and their corresponding sides are proportional.

2. The given ratio AD:DB = 12:8 can be simplified by dividing both parts by their greatest common divisor (4): AD:DB = 3:2. This indicates that AD is 3/5 of AB and DB is 2/5 of AB.

3. The Basic Proportionality Theorem (Thales' theorem) states that if a line is drawn parallel to one side of a triangle, intersecting the other two sides, then it divides those sides proportionally. In our case, this means that: AD/AB = DE/BC = AE/AC

4. We are given AD:DB = 3:2 and DE = 24 cm. We can use the proportion DE/BC = AD/AB. Since AD/DB = 3/2, the ratio AD/AB is 3/(3+2) = 3/5. Therefore: 24/BC = 3/5

5. To find BC, we can cross-multiply and solve the equation: 24 * 5 = 3 * BC 120 = 3 * BC BC = 120 / 3 BC = 40 cm Therefore, the length of BC is 40 cm.
Think/
The length of BC is 40.0 cm. Therefore, the correct answer is B.



How many study options does Australia offer for international students?

\Think
To calculate the total number of study options available to international students in Australia, we need to analyze the data provided in the infographic. The infographic clearly states:

1.Australia has more than 1200 institutions offering education.

2.There are over 22,000 courses available across these institutions. The question asks for the *total number of study options*. While we can't definitively say each institution offers each course, it's reasonable to assume that the combination of institutions and courses provides a reasonable proxy for the total number of distinct study options available.

A student has a choice among 1200 institutions and within those, a choice of 22000 courses. Adding the numbers is not the ideal method, but it is the only information provided.

The assumption that courses are independent from the institution and are additive in terms of options is made to address the problem. Therefore, a reasonable estimate of the total number of study options can be obtained by adding the number of institutions and the number of courses.

It's not a perfect metric as it might overcount in cases of multiple institutions offering the same course; However, given the limited data, this approach provides a possible estimate: Total Study Options ≈ Number of Institutions + Number of Courses Total Study Options ≈ 1200 + 22000 = 23200 Note: The reference reasoning incorrectly states 12,000 institutions. The infographic clearly shows 1200 institutions.
Think/

information provided in the infographic, Australia offers approximately 23,200 study options for international students.

Figure 18: **Long-form visual reasoning on geometry and an infographic.**

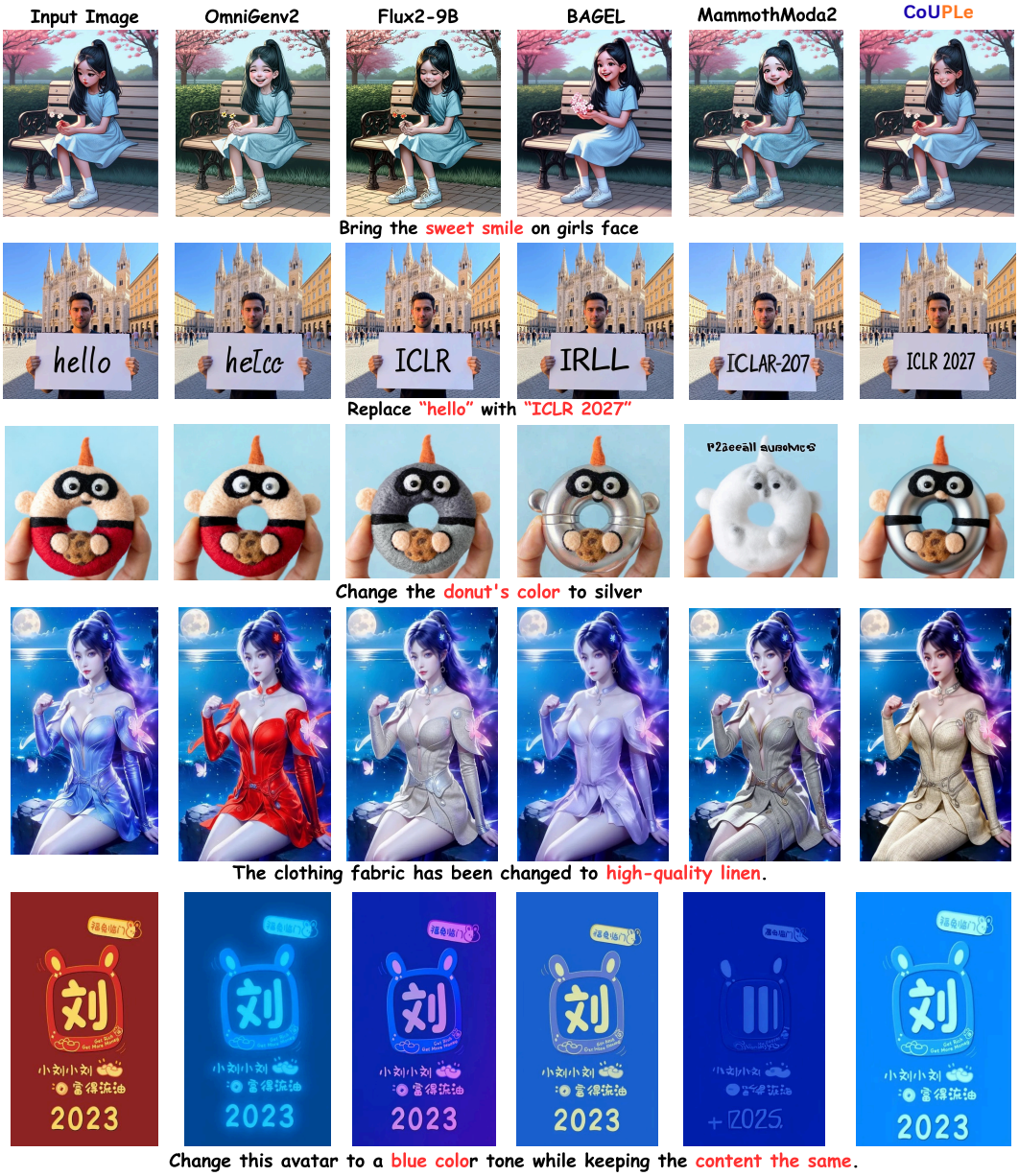


Figure 19: **Qualitative comparison on instruction-guided image editing.** Columns show the input image and outputs from OmniGen2, Flux2-9B, BAGEL, MammothModa-2, and **CoUPLe**. Rows evaluate facial-expression modification, exact text replacement, object recoloring, fabric modification, and global color-tone adjustment. Each instruction appears below its corresponding row.



Figure 20: **Qualitative comparison on text-to-image generation.** Columns compare BAGEL, FUDOKI, OmniGen2, LaViDa-O, and **CoUPLe** under the same prompts. Rows cover a technical scene with graphical annotations, a coffee-shop menu, a rainy street photograph, layered paper-cut artwork, and a typographic poster. The examples illustrate differences in prompt-attribute coverage, text rendering, scene composition, and visual style.



Figure 21: **Qualitative comparison with and without reasoning.** From left to right: input image and instruction, editing rationale, output with reasoning, and output without reasoning.

(a) Explain the given image.



This low-angle street photograph centers on a **black-and-white "ONE WAY" sign with a left-pointing arrow**. Beneath it, **two large yellow traffic-signal housings** are mounted on yellow horizontal supports, with their backs facing the camera. The broader housing on the right bears **dark, handwritten graffiti**, while the narrower housing on the left has a plain rear panel and a weathered underside. **Green evergreen branches** rise through the foreground and partially obscure the mounting structure. Behind the signals, **brown brick buildings with metal fire escapes** fill the left and lower background. A curved streetlight extends into the pale, overcast sky at the upper right. Another directional sign is partially cropped along the top edge, and part of a traffic light appears at the bottom left.

(b) Caption the given image



A warmly lit indoor display features a **long wooden beam engraved "STONES THROW WINERY"** above a counter. Its uneven edges and visible grain contrast with a **textured golden-yellow wall**, whose cracks and mottled finish give the setting a rustic appearance. Below it, **dark wine bottles arranged in orderly groups** stand in staggered rows. Their black labels display white numerical combinations: **"50/50," "60/40," "70/30," and "90/10."** Some bottles have black neck coverings, while others have pale gold ones. **Upright wooden presentation boxes and clear stemmed glasses** separate the groups. Near the center, a gray knitted bottle cover with a white top and red accent adds a decorative detail. White ceiling beams cross the upper edge, and warm lighting emphasizes the wood and amber wall.

(c) Caption the given image



A close-up photograph shows a **clear, stemmed glass filled with amber-colored beer** beside a **tall, uncapped brown glass bottle** on a wooden table. The glass stands on the left and has a gently flared rim, a rounded lower bowl, a short stem, and a broad circular base. A **thick layer of white foam** rests above the beer, with traces clinging to the inner glass. The bottle on the right extends higher than the glass; its **rear label contains a barcode and small white text**. Bright reflections trace both curved surfaces. **Wooden chairs and softly blurred seated people** occupy the background. Above them, a large tan canopy and an elevated metal railing frame the covered seating area.

(d) Describe the image in detailed.



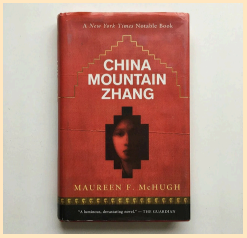
This vertical event poster uses a **vivid orange, black, white, and red palette**. Its central illustration combines a **skateboard, a stylized rib cage, and a red anatomical heart**, while two orange hands grip the lower end of the board. Broad white brush-like strokes sit behind the illustration, and a small red flag appears at the top. Distressed lettering presents **"SKATE" and "CONTR2L," with a red numeral 2**. At the lower left, **"21 Julio" and "Carrera 19 / calle 22D"** give the date and address. The right side advertises **"Bandas en Vivo," "Muro de escalada," and "Graffiti."** A red strip announces free registration and the categories "Amateur, Libre y Chicas." The white footer contains organizational logos, including "BOGOTÁ HUMANA."

Figure 22: **Qualitative image-to-text generation.** Input images and instructions are shown on the left, with descriptions generated by **CoUPLe** on the right. The examples cover a street scene, a winery display, a beverage photograph, and an event poster, illustrating descriptions of object attributes, spatial relations, scene context, and visible text. Red text highlights selected details in the responses.

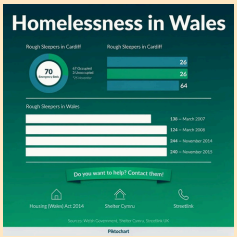
This vintage-style illustrated postcard reads **"GREETINGS FROM"** above the oversized word **"KENTUCKY."** The main lettering curves across the composition with cream outlines, orange upper edges, and a broad black shadow. **Small scenic illustrations fill the letters,** including a large house, a waterfall, a horse-racing scene, a rocky interior, a riverboat, and grazing cattle. Pastel blue, yellow, green, and peach areas form the background, with diagonal bands near the top. A **cluster of yellow goldenrod flowers** appears at the lower left beside the label **"State Flower the Goldenrod."** At the lower right, a **domed government building** stands behind a green lawn and trees, accompanied by the text **"State Capital in Frankfort."** An off-white border and textured printed surface reinforce the postcard's aged appearance.



The photographed book cover is dominated by a **textured crimson-red background** with cream, gold, and black graphic elements. The title **"CHINA MOUNTAIN ZHANG"** is arranged in three large uppercase lines across the upper half, over a **fine gold outline of an angular, roof-like structure.** A small endorsement at the top reads **"A New York Times Notable Book."** Below the title, a dark stepped shape frames a **sideways portrait of a person's face,** tinted red and partly obscured by shadow. A narrow strip of small yellow-toned portraits runs across the lower section. The author's name, **"MAUREEN F. McHUGH,"** appears prominently above a dark bottom band containing a review quotation. Slight edge wear, faint horizontal lines, and uneven lighting emphasize the physical printed cover.



Teal-to-green infographic is titled **"Homelessness in Wales."** Its upper section, "Rough Sleepers in Cardiff," combines a doughnut chart and horizontal bars. The doughnut shows **70 emergency beds: 67 occupied and 3 unoccupied,** with a note specifying 25 November. The neighboring bars report **26, 26, and 64 people sleeping on the street.** Under "Rough Sleepers in Wales," four white bars give **138 in March 2007, 124 in March 2008, 244 in November 2014, and 240 in November 2015.** White dividers organize the charts and lower information area. A ribbon reading **"Do you want to help? Contact them!"** introduces columns headed "Housing (Wales) Act 2014," "Shelter Cymru," and "Streetlink." Source credits appear near the bottom, followed by a white footer with the Piktochart logo.



A warmly lit close-up centers on a **brown glass bottle with an orange, cream, and olive-green label.** The upper badge reads **"30 spinnakers ANNIVERSARY."** The main illustration shows a **pumpkin-headed scarecrow with stitched facial features,** wearing a hat and dark coat, extending its arms, and holding a bottle. Radiating lines evoke fields behind it, with a small barn to the left and pumpkins along the bottom. Large lettering identifies the drink as **"SPICED PUMPKIN ALE."** The lower label lists **"5.0% ABV," "650 ML," and "15 IBU,"** followed by **"CRAFT BREWED SINCE 1984."** Moisture and bright reflections are visible on the bottle's shoulder. A tall glass of dark liquid stands out of focus behind it on the left, creating a softly blurred background.



A dark chalkboard displays a hand-drawn puzzle titled "GEOMETRY PROBLEM:" in large white chalk lettering. Two arrangements of a tall rectangular block and a smaller square block appear side by side. In the left arrangement, the blocks rest on the same baseline, with the small square positioned against the rectangle's lower-left side. A pink vertical measurement marker labeled "400" extends from the square's top to the rectangle's top, indicating the difference between their heights. In the right arrangement, the small square sits on top of the tall rectangle, aligned with its left edge. A second pink marker labeled "460" spans from the rectangle's base to the square's top, indicating their combined height. Beneath this arrangement, the partially visible expression "$H_1 + H_2 =$" reinforces the addition relationship. Near the bottom, a separate square outline followed by "= ?" identifies the smaller block's height as the unknown.

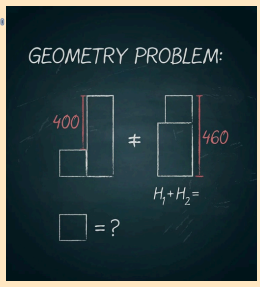


Figure 23: **Qualitative text-to-image generation from detailed prompts.** Each row pairs an input prompt (left) with an image generated by **CoUPLe** (right). The examples span a vintage postcard, a book cover, an infographic, a product photograph, and a geometric diagram. Highlighted prompt phrases identify selected content, appearance, text, and layout constraints.