UniOwl: Bridging Multimodal Understanding and Generation with Adaptive Conditioning

Anonymous authors
Paper under double-blind review

Abstract

Unified multimodal models aim to support visual understanding and image generation within a single architecture, but diffusion-based generation introduces a key trade-off: while semantic guidance should adapt across denoising steps, invoking a large language model at each step is computationally prohibitive. Existing methods rely on either static conditions or costly step-wise language computation. We introduce UniOwl, a unified multimodal framework for joint understanding and generation with Adaptive Dual-Channel Conditioning. UniOwl first extracts step-invariant semantic condition tokens from LLM hidden states, computed once and shared across all denoising steps, then refines them at each step with lightweight latent- and timestep-aware residuals. In parallel, visual condition tokens are derived from the current denoising latent via a shared visual encoder. The fused semantic and visual tokens provide dynamic conditioning signals for the diffusion model, while generation supervision on the visual pathway ensures consistency between generation and understanding. To further encourage generation-understanding consistency, we introduce a cycle consistency objective that feeds generated images back into the understanding pathway and encourages their output distributions to be consistent with the original predictions. Extensive experiments show that UniOwl achieves superior generation quality and semantic consistency over existing unified models, while maintaining competitive multimodal understanding performance.

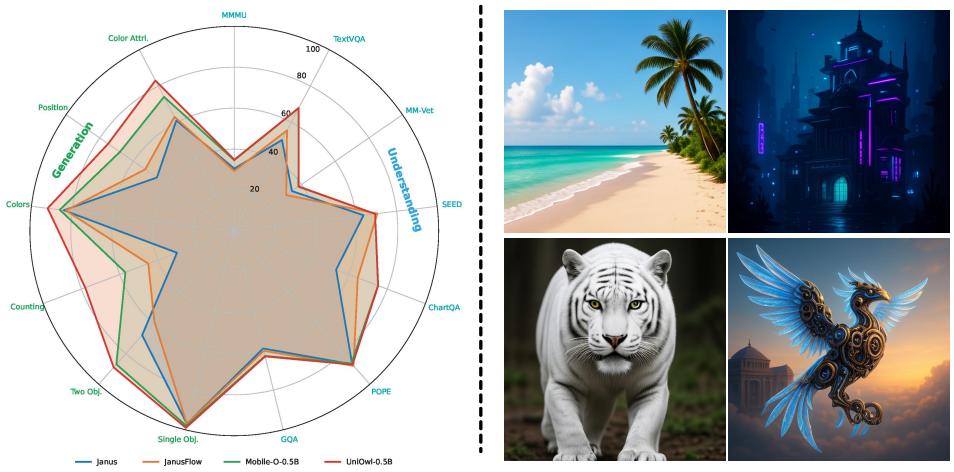
1 Introduction

Unified multimodal models aim to support both visual understanding and image generation in a single framework. This goal is important because understanding and generation are naturally connected. A model that understands images should also be able to generate images that follow the same semantic structure. A model that generates images should also preserve information that can be correctly interpreted by the understanding pathway. Building such a unified model can reduce system complexity, improve parameter sharing, and enable stronger interaction between perception, reasoning, and generation.

Recent works have explored different ways to unify these two abilities. Multimodal understanding models, such as LLaVA (Liu et al., 2023), BLIP-2 (Li et al., 2023b), and InstructBLIP (Dai et al., 2023), connect a visual encoder with a large language model and achieve strong visual reasoning ability. Text-to-image models, such as Stable Diffusion (Rombach et al., 2022), PixArt-α (Chen et al., 2024a), and SDXL (Podell et al., 2023), use diffusion models to generate high-quality images from prompts. More recent unified models, such as Show-o (Xie et al., 2025), SEED-X (Ge et al., 2024), Janus (Wu et al., 2025), JanusFlow (Ma et al., 2025), Emu3 (Wang et al., 2024), BLIP3-o (Chen et al., 2025a), and Mobile-O (Shaker et al., 2026), further attempt to combine understanding and generation within one architecture. These methods show that a shared multimodal system is possible, and the design of generation conditions is a key factor for unified modeling.

However, existing methods still face a conditioning dilemma. Diffusion-based generation needs conditions that adapt to the current denoising state (Ho et al., 2020; Rombach et al., 2022). Early denoising steps need coarse semantic and layout guidance. Later steps need fine-grained attributes and

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) Comparison Results

(b) Sample images by our UniOwl

Figure 1: Overview of UniOwl's performance and generation capabilities. (a) Radar chart comparing unified multimodal models on visual understanding and image generation benchmarks. UniOwl achieves strong performance across both types of tasks. (b) Sample images generated by UniOwl, demonstrating its ability to synthesize diverse content while maintaining semantic consistency with input prompts.

visual details. Static conditions are efficient, but they remain fixed across all denoising steps. Dynamic conditions can provide stronger guidance, but they often require repeated language-side computation or extra modules inside the diffusion loop. This increases inference cost. Another challenge is train-inference consistency. Some designs use clean target-image features during training, but these features are not available during inference. This creates a gap between training and generation. In addition, many unified models do not fully align the visual pathway for both understanding and generation. For example, Janus (Wu et al., 2025) decouples visual encoding pathways to handle the representation conflict between understanding and generation, while Mobile-O (Shaker et al., 2026) connects an LLM-based understanding pathway with a diffusion-based generation pathway through a semantic connector. These designs are effective, but diffusion-time adaptive conditioning remains underexplored.

To address these issues, we introduce UniOwl, a unified framework for visual understanding and image generation. UniOwl uses diffusion-adaptive semantic and visual conditioning. Specifically, the semantic channel extracts condition tokens from LLM hidden states once and refines them at each denoising step using the current noisy latent and timestep. This gives the generator dynamic semantic guidance without repeatedly running the LLM. The visual channel derives visual tokens from the current denoising latent through the shared visual tokenizer. Since the noisy latent exists during both training and inference, this design avoids train-inference mismatch. The refined semantic tokens and visual tokens are fused by a gated module and used to condition the DiT denoiser (Xie et al., 2024). We further introduce Cycle Consistency Learning, which feeds generated images back into the understanding pathway and encourages the model to interpret them consistently with the original inputs. As shown in Fig. 1, our UniOwl achieves strong performance on both visual understanding and image generation tasks, while maintaining semantic consistency between the two modalities.

Our contributions are summarized as follows. First, we identify the conditioning dilemma in unified diffusion-based understanding and generation, where the model needs dynamic, consistent, and efficient generation guidance. Second, we propose diffusion-adaptive semantic conditioning, which keeps expensive LLM reasoning outside the denoising loop while still adapting semantic conditions at each diffusion step. Third, we introduce generation-driven visual conditioning, which aligns the shared visual tokenizer with image generation and preserves train-inference consistency. Finally, we propose Cycle Consistency Learning to encourage semantic alignment between generated images and the understanding pathway. Experiments show that UniOwl improves image generation quality and semantic consistency while maintaining strong visual understanding performance.

2 Related Work

Multimodal Understanding. Multimodal understanding aims to generate descriptions, answers, or reasoning from visual inputs. Recent progress is largely driven by connecting pretrained visual encoders with large language models. A common design, adopted by LLaVA (Liu et al., 2023), BLIP-2 (Li et al., 2023b), and InstructBLIP (Dai et al., 2023), encodes images with ViT-based visual encoders (Dosovitskiy et al., 2021) and maps visual features into the LLM embedding space through lightweight connectors. This architecture has become a standard design for multimodal large language models. LLaVA-1.5 (Liu et al., 2024b) further demonstrates that simple projection layers can achieve strong image-text alignment on benchmarks such as MMBench, MME, and SEED-Bench (Li et al., 2023a). Later methods improve this paradigm through stronger connectors, larger visual backbones, and broader training data. For example, Qwen-VL (Bai et al., 2023) compresses visual tokens before LLM processing, while InternVL (Chen et al., 2024b) scales the visual encoder and alignment pipeline for fine-grained perception. Another line of work focuses on efficient multimodal understanding. LLaVA-Phi and LLaVA-v1.5-Phi-1.5 explore compact multimodal assistants with smaller language models (Zhu et al., 2024). MobileVLM (Chu et al., 2023) and MobileVLM-V2 (Chu et al., 2024) target efficient VLMs for resource-limited scenarios. Recent models further reduce visual processing cost: LLaVA-OneVision (Li et al., 2025) improves transfer across image and video tasks, SmolVLM-0.5B (Marafioti et al., 2025) studies sub-billion-scale multimodal models, and FastVLM-0.5B (Vasu et al., 2025) improves the latency-accuracy trade-off with fewer visual tokens and efficient high-resolution vision encoding. However, these models mainly focus on discriminative multimodal tasks, while image generation is usually handled by separate generative models. In contrast, UniOwl uses the visual pathway for both understanding and generation, allowing generation supervision to shape the shared visual tokenizer through dynamic visual conditioning.

Text-to-Image Generation. Text-to-image generation has evolved from GAN-based methods to diffusion and autoregressive models. Early approaches such as StackGAN (Zhang et al., 2017) and AttnGAN (Xu et al., 2018) explore multi-stage adversarial refinement and text-image alignment, while diffusion models (Ho et al., 2020), especially latent diffusion models (Rombach et al., 2022), have become dominant for high-quality synthesis. Stable Diffusion (Rombach et al., 2022) and SDXL (Podell et al., 2023) improve generation quality through latent diffusion scaling, while recent methods focus on efficiency and scalability. PixArt-α (Chen et al., 2024a) studies efficient Transformer-based diffusion training, SANA-0.6B (Xie et al., 2024) improves high-resolution synthesis with efficient architectures, and SnapGen (Hu et al., 2024) targets mobile text-to-image generation through compact models and few-step sampling. Beyond diffusion, LlamaGen (Sun et al., 2024) explores autoregressive image generation with next-token prediction. However, these models are mainly designed for generation only: text conditions are typically extracted once and remain static during denoising, without interaction with visual understanding. In contrast, UniOwl unifies understanding and generation by extracting semantic conditions from LLM hidden states, refining them across denoising steps, and introducing dynamic visual conditioning through the shared visual pathway.

Unified Multimodal Understanding and Generation. Recent multimodal foundation models aim to support visual understanding and image generation within a unified framework. One line of work integrates modalities through discrete token prediction. Emu3 tokenizes text, images, and videos into a discrete space and trains a single Transformer with next-token prediction (Wang et al., 2024). LWM (Liu et al., 2024a), Show-o (Xie et al., 2025), SEED-X (Ge et al., 2024), and BLIP3-o (Chen et al., 2025a) further explore unified multimodal modeling with different combinations of autoregressive modeling, diffusion, and visual representations. Another line of work focuses on bridging the representation gap between understanding and generation. Janus (Wu et al., 2025) decouples visual encoding pathways, while JanusFlow (Ma et al., 2025) combines autoregression and rectified flow. Mobile-O (Shaker et al., 2026) connects an LLM-based understanding pathway and a diffusion-based generation pathway through a semantic connector. Existing unified models therefore follow different paradigms, including integrated multimodal transformers and modular vision-language-diffusion frameworks. UniOwl follows the modular paradigm and focuses on diffusion-time adaptive conditioning. It extracts reusable semantic condition tokens from LLM hidden states, refines them with latent- and timestep-aware updates, and introduces generation-driven visual conditioning to provide feedback from the evolving diffusion state.

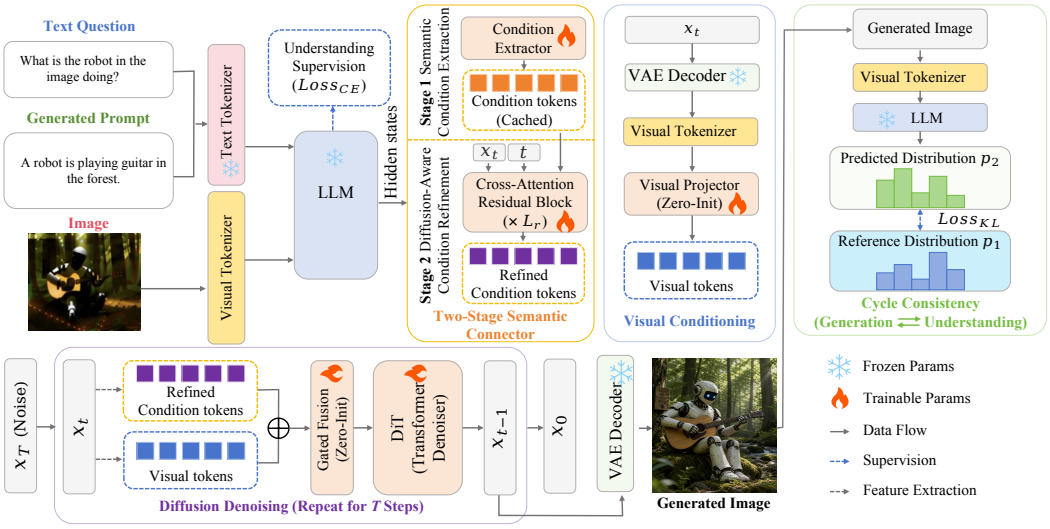


Figure 2: Overview of UniOwl. UniOwl unifies visual understanding and image generation with adaptive dual-channel conditioning. LLM hidden states are converted into condition tokens and refined at each diffusion step using the current noisy latent x_t and timestep t. In parallel, x_t is decoded and passed through the shared visual tokenizer to obtain visual tokens. The refined condition tokens and visual tokens are fused to guide the DiT denoiser. A cycle consistency objective feeds generated images back into the understanding pathway to encourage semantic alignment.

3 METHOD

3.1 OVERALL ARCHITECTURE

The overall architecture of UniOwl is shown in Fig. 2. Our proposed UniOwl contains three main components: a LLM backbone *i.e.,* Qwen2-0.5B (Yang et al., 2024) for multimodal reasoning, a shared visual tokenizer for visual encoding, and a DiT-based diffusion generator for image synthesis. In addition, a standard text tokenizer is used to encode both the understanding question and the generation prompt. The shared visual tokenizer is used in three places: visual understanding, generation-driven visual conditioning, and consistency learning. Specifically, for visual understanding, UniOwl takes a text question and an input image as input. The text question is converted into text tokens, while the image is converted into image tokens by the shared visual tokenizer. These tokens are fed into the LLM backbone, which predicts the textual answer and is optimized with a next-token prediction loss. And for image generation, UniOwl uses the generation prompt as the textual condition. The prompt is processed by the same text tokenizer and passed through the LLM backbone. The resulting hidden states provide semantic information for the diffusion generator. The key challenge is how to convert these hidden states into effective denoising conditions. A useful condition should be adaptive, consistent, and efficient. It should adapt to different denoising stages, remain available in the same form during training and inference, and avoid repeated LLM computation inside the diffusion loop.

A good condition should satisfy three requirements. First, it should be adaptive. Different denoising steps need different information. Early steps need coarse layout and object-level semantics. Later steps need attributes, texture, and local details. Second, it should be consistent. The condition should be computed in the same way during training and inference. Third, it should be efficient. Running the LLM at every denoising step is expensive. Our UniOwl addresses these requirements with adaptive dual-channel conditioning. The semantic channel extracts condition tokens from LLM hidden states and refines them at each denoising step. We denote the initial condition tokens as $\mathbf{C}^0$ and the refined condition tokens at timestep t as $\mathbf{C}_t^{\text{ref}}$. The visual channel derives visual tokens $\mathbf{V}_t$ from the current noisy latent x_t. These two token streams are fused into $\mathbf{C}_t^{\text{fuse}}$, which is used as the only condition for the DiT denoiser. This design keeps expensive language reasoning outside the denoising loop while still providing dynamic semantic and visual guidance throughout generation. We further introduce a

cycle consistency objective, which feeds generated images back into the understanding pathway and encourages semantic alignment between generation and understanding.

3.2 Diffusion-adaptive Semantic Conditioning

The semantic channel provides language-derived guidance for image generation. Existing unified understanding-generation models connect LLM outputs to the generator in a fixed or pre-computed way. For example, Show-o (Xie et al., 2025) uses a single Transformer for autoregressive modeling and discrete diffusion, SEED-X (Ge et al., 2024) supports multi-granularity comprehension and generation, Janus (Wu et al., 2025) decouples visual encoding pathways, JanusFlow (Ma et al., 2025) combines autoregression with rectified flow, and Mobile-O (Shaker et al., 2026) uses a semantic connector. While effective, these approaches tie the generation condition to a static representation. They cannot adapt to the evolving state of the diffusion denoiser. Diffusion generation is dynamic. Early timesteps require global layout and object-level semantics. Middle timesteps emphasize attributes, relations, and coarse structures. Later timesteps require fine details, textures, and precise visual attributes. A fixed condition cannot respond to these changing requirements or compensate for under-generated regions.

To address this, we introduce the diffusion-adaptive semantic conditioning strategy, which provides dynamic, timestep-aware semantic guidance for image generation. The approach first extracts reusable condition tokens from LLM hidden states and then refines them at each denoising step using the current DiT features. It comprises two stages: semantic condition extraction and diffusion-aware condition refinement. Specifically, given the LLM hidden states $\mathbf{H}$, a lightweight condition extractor produces the base tokens,

$$\mathbf{C}^0 = f_{\text{cond}}(\mathbf{H}), \quad \mathbf{C}^0 \in \mathbb{R}^{B\times N\times D}, \tag{1}$$

where B is the batch size, N is the number of condition tokens, and D is the condition dimension. These tokens are computed once and cached for all diffusion steps. They encode stable semantic information from the prompt and multimodal context, such as object categories, relations, style, and global intent. However, the base tokens do not observe the current generation state. To enable step-wise adaptation, a lightweight refinement module is applied at each timestep t. It treats the base tokens as semantic queries and attends to the current DiT feature $\mathbf{X}_t$ concatenated with a timestep embedding, producing a residual update,

$$\mathbf{C}_t^{\text{ref}} = \mathbf{C}^0 + \gamma f_{\text{refine}}(\mathbf{C}^0, \mathbf{X}_t, t), \tag{2}$$

where $\mathbf{C}_t^{\text{ref}}$ denotes the refined condition tokens and γ is a learnable scalar initialized to zero. The timestep embedding informs the module which stage of denoising is in progress, enabling different attention patterns at early versus late steps.

This design gives an intuitive interpretation. The base tokens indicate *what should be generated*. The DiT feature indicates *what has been generated so far*. The refinement module dynamically emphasizes semantic concepts based on the current generation state and denoising stage. Early steps focus on global layout and main objects. Later steps focus on attributes, textures, and fine details. Zero initialization of γ ensures stability: initially, the model behaves like a static-conditioned generator, and the refinement gradually learns to adjust semantic guidance during training.

3.3 Generation-driven Visual Conditioning

The semantic channel provides prompt-level guidance. However, existing unified models, such as Janus (Wu et al., 2025), Mobile-O (Shaker et al., 2026), and SEED-X (Ge et al., 2024), often train the visual encoder primarily with understanding losses, without explicit supervision for generation. As a result, the visual features may not optimally guide image generation. A natural solution is to use ground-truth image features as generation conditions (Rombach et al., 2022; Chen et al., 2024a), but this introduces a train-inference mismatch because the target image is unavailable at inference time.

To address this, we further introduce generation-driven visual conditioning to provide timestep-aware visual guidance for image generation. At each diffusion step t, the current noisy latent x_t is decoded into an intermediate image and processed by the shared visual tokenizer and visual projector to produce a set of visual tokens,

$$\mathbf{V}_t = f_{\text{vis}}\big(\text{Tok}_{\text{vis}}(\text{Dec}_{\text{VAE}}(x_t))\big). \quad \mathbf{V}_t \in \mathbb{R}^{B\times N\times D}. \tag{3}$$

These tokens encode the current generation state and are available during both training and inference. This ensures train-inference consistency while allowing generation supervision to refine the visual tokenizer (Hu et al., 2024; Xie et al., 2024). After that, the visual tokens are fused with the diffusion-adaptive semantic tokens. At each timestep, the refined semantic tokens $\mathbf{C}_t^{\text{ref}}$ and visual tokens $\mathbf{V}_t$ are concatenated,

$$\mathbf{U}_t = \text{Concat}\big[\mathbf{C}_t^{\text{ref}}, \mathbf{V}_t\big], \quad \mathbf{U}_t \in \mathbb{R}^{B\times 2N\times D}. \tag{4}$$

The concatenated tokens are passed into a zero-initialized gated fusion module:

$$\mathbf{C}_t^{\text{fuse}} = f_{\text{gate}}(\mathbf{U}_t), \quad \mathbf{C}_t^{\text{fuse}} \in \mathbb{R}^{B\times N\times D}. \tag{5}$$

where the output $\mathbf{C}_t^{\text{fuse}}$ is the only condition for the DiT denoiser. The zero-initialized gate prevents unstable visual injection at the beginning of training. Initially, the generator relies mainly on semantic tokens. As training progresses, the gate gradually incorporates visual tokens, providing adaptive visual guidance. This aligns the generation pathway with the understanding pathway and preserves dynamic semantic guidance (Chen et al., 2024a; Rombach et al., 2022).

3.4 Cross-Task Consistency Learning

Adaptive dual-channel conditioning connects understanding and generation at the feature level. However, feature-level coupling does not guarantee semantic consistency in the generated image. The image may look realistic, but it may miss an object, attribute, or relation in the input. UniOwl introduces a cycle consistency loss. Specifically, given an image-text pair $(\mathbf{I}, q)$, UniOwl first runs the original image and question through the understanding pathway. This gives a reference output distribution p_1. The generation pathway then produces an image $\tilde{\mathbf{I}}$ from the corresponding generation prompt. We detach $\tilde{\mathbf{I}}$ and feed it back into the same understanding pathway with the question q. This gives a cycle-induced distribution p_2,

$$p_1 = p_\theta(y \mid \mathbf{I}, q), \quad p_2 = p_\theta(y \mid \text{sg}(\tilde{\mathbf{I}}), q), \tag{6}$$

where $\text{sg}(\cdot)$ denotes stop-gradient. We minimize the KL divergence between p_1 and p_2 on valid supervised token positions,

$$\mathcal{L}_{\text{cycle}} = \frac{1}{|\mathcal{M}|} \sum_{i\in\mathcal{M}} D_{\text{KL}}\left(\text{sg}(p_1^{(i)}) \| p_2^{(i)}\right), \tag{7}$$

where $\mathcal{M}$ denotes the valid token positions. This avoids supervising padding or ignored tokens and reduces memory cost. Finally, the full training objective is,

$$\mathcal{L} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{diff}}\mathcal{L}_{\text{diff}} + \lambda_{\text{cycle}}\mathcal{L}_{\text{cycle}}, \tag{8}$$

where $\mathcal{L}_{\text{ce}}$ is the next-token prediction loss for visual understanding. $\mathcal{L}_{\text{diff}}$ is the diffusion loss for image generation. $\mathcal{L}_{\text{cycle}}$ is the proposed cycle consistency loss.

3.5 Training Schemes

Stage 1: Pre-training. The goal of this stage is to train the condition extractor to effectively align LLM features with the DiT's conditional feature space. During pre-training, the parameters of the LLM, visual encoder, and VAE are frozen. The training uses 9 million image-caption pairs from the JourneyDB (Sun et al., 2023) and BLIP3o-Short-Caption (Chen et al., 2025a) datasets. This stage establishes a stable connection between language and visual features for subsequent generation.

Stage 2: Supervised Fine-Tuning (SFT). In this stage, we use high-quality image-caption pairs to further refine the condition extractor and cross-attention residual blocks. The LLM, visual encoder, and VAE remain frozen, while the condition extractor and residual blocks are trainable. The training data consists of 105K image-caption pairs from BLIP3o (Chen et al., 2025a) and ShareGPT-4o-Image (Chen et al., 2025b). This stage improves the alignment between semantic conditions and the DiT generator for more accurate and coherent generation.

Stage 3: Unified Understanding and Generation. The final stage jointly trains the model for both visual understanding and image generation. Trainable components include the condition extractor, cross-attention residual blocks, visual encoder, visual projector, and the LLM with LoRA adapters. Training samples are constructed as quadruplets $S = p, x_{\text{img}}, q, a$, where p is the generation prompt, x_{img} is the image, and (q, a) is the associated question-answer pair (Shaker et al., 2026). This stage allows the model to learn shared multimodal representations that support both tasks.

Table 1: Comparison of unified and understanding-only models on multimodal understanding benchmarks. "Und." and "Gen." refer to understanding and generation, respectively. Total parameters include the visual encoder, language model, and diffusion/UNet components. Best results are highlighted in bold.

Type	Method	# Params	MMMU	TextVQA	MM Vet	SEED	ChartQA	POPE	GQA	Avg
Und. Only > 1B	LLaVA-Phi (Zhu et al., 2024)	3.1B	–	48.6	28.9	–	–	85.0	–	–
	LLaVA-v1.5-Phi-1.5 (Zhu et al., 2024)	1.6B	30.7	–	–	–	–	84.1	56.5	–
	MobileVLM (Chu et al., 2023)	1.7B	–	41.5	–	–	–	84.5	56.1	–
	MobileVLM-V2 (Chu et al., 2024)	1.7B	–	52.1	–	–	–	84.3	59.3	–
	LLaVA-OV (Li et al., 2025)	1.6B	31.4	–	29.1	65.5	61.4	–	–	–
Und. Only ≤ 1B	Smol-VLM-0.5B (Marafioti et al., 2025)	0.6B	33.7	60.2	–	–	62.8	–	–	–
	FastVLM-0.5B (Vasu et al., 2025)	0.6B	33.3	68.0	37.5	69.3	71.6	81.1	62.7	60.5
Und. and Gen. > 2B	EMU3-8B (Wang et al., 2024)	9.0B	31.6	64.7	37.2	68.2	68.6	85.2	60.3	59.4
	BLIP3o-4B (Chen et al., 2025a)	7.1B	**46.6**	**78.0**	**60.1**	**73.8**	–	–	–	–
Und. and Gen. ≤ 2B	Janus (Wu et al., 2025)	2.1B	30.5	50.2	34.3	63.7	53.0	87.0	59.1	54.0
	Show-o (Xie et al., 2025)	1.5B	25.1	–	–	–	–	73.8	48.7	–
	Show-o-Clip-ViT (Xie et al., 2025)	1.6B	27.4	41.2	20.9	51.6	44.7	84.5	57.5	46.8
	JanusFlow (Ma et al., 2025)	2.1B	29.3	55.5	30.9	70.5	64.6	**88.0**	60.3	57.0
	Mobile-O-0.5B (Shaker et al., 2026)	1.6B	34.6	67.8	38.1	69.4	75.2	86.4	62.9	62.1
	UniOwl-0.5B (Ours)	1.7B	34.8	67.9	38.3	69.4	**75.3**	87.2	**63.1**	**62.3**

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS

For the understanding backbone, we adopt FastVLM-0.5B (Vasu et al., 2025), which pairs FastViT (Vasu et al., 2023) as the vision encoder with Qwen2-0.5B (Yang et al., 2024) as the language model. For image generation, we employ SANA-600M-512 (Xie et al., 2024), a DiT-based latent diffusion model. Specifically, the semantic branch exploits cached LLM hidden states augmented with lightweight timestep-aware attention residuals, and the visual branch reuses the shared vision encoder to extract features from decoded intermediate latents.

During training, each batch contains both understanding and generation samples. Understanding samples are supervised with the next-token prediction loss, while generation samples are supervised with the diffusion loss. For paired image-text samples, we additionally apply the cycle consistency loss. The LLM is updated through lightweight adapters, while the vision encoder, conditioning modules, and generation interfaces are optimized jointly. All experiments are conducted on a single node equipped with 8 NVIDIA A100 GPUs. The pre-training stage runs for 60k steps. The subsequent SFT stage runs for 25 epochs. And the unified understanding and generation step is trained for 10 epochs.

4.2 COMPARISON RESULTS

Multimodal Visual Understanding. Table 1 reports the comparison on multimodal understanding benchmarks. Among unified understanding-generation models under 2B parameters, UniOwl achieves the best average score of 62.3, improving over Mobile-O-0.5B by 0.2 points and outperforming Janus (Wu et al., 2025), Show-o (Xie et al., 2025), Show-o-CLIP-ViT (Xie et al., 2025), and JanusFlow (Ma et al., 2025) by a clear margin. UniOwl also performs competitively against understanding-only small models. It surpasses FastVLM-0.5B (Vasu et al., 2025) on the average score, while supporting image generation in addition to visual understanding. Compared with larger unified models, UniOwl achieves a better average score than EMU3-8B (Wang et al., 2024) with far fewer parameters. These results show that UniOwl preserves strong multimodal understanding ability while adding generation capability, suggesting that the proposed diffusion-adaptive semantic and visual conditioning does not compromise the understanding pathway.

Text-to-Image Generation. We evaluate text-to-image generation on GenEval (Ghosh et al., 2023) using the original prompts. As shown in Table 2, UniOwl achieves an overall score of 0.83, which is the competitive performance among unified understanding-generation models under 2B parameters. It improves over Mobile-O-0.5B (Shaker et al., 2026) from 0.74 to 0.83 and outperforms Show-o (Xie et al., 2025), Show-o-Clip-ViT (Xie et al., 2025), Janus (Wu et al., 2025), and JanusFlow (Ma et al., 2025) by clear margins. UniOwl also achieves the best score in every GenEval sub-task within this group, especially on Counting, Position, and Color Attribution, where compositional reasoning is important. Compared with generation-only models, UniOwl also remains highly competitive. It

Table 2: Results on the GenEval benchmark. Best results among Und. and Gen. $\leq$ 2B models are bolded.

Type	Method	# Total Params	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
Gen. Only > 1B	LlamaGen (Sun et al., 2024)	3.8B	0.71	0.34	0.21	0.58	0.07	0.04	0.32
	LDM (Rombach et al., 2022)	1.5B	0.92	0.29	0.23	0.70	0.02	0.05	0.37
	PixArt-α (Chen et al., 2024a)	4.9B	0.98	0.50	0.44	0.80	0.08	0.07	0.48
	SDXL (Podell et al., 2023)	3.0B	0.98	0.74	0.39	0.85	0.15	0.23	0.55
	SDv2.1 (Rombach et al., 2022)	1.9B	0.98	0.51	0.44	0.85	0.07	0.17	0.50
	SANA-0.6B (Xie et al., 2024)	2.6B	0.99	0.77	0.62	0.88	0.21	0.47	0.66
Gen. Only $\leq$ 1B	SDv1.5 (Rombach et al., 2022)	1.0B	0.97	0.38	0.35	0.76	0.04	0.06	0.43
	SnapGen (Hu et al., 2024)	0.4B	1.00	0.84	0.60	0.88	0.18	0.45	0.66
Und. and Gen. > 2B	SEED-X (Ge et al., 2024)	16.0B	0.97	0.58	0.26	0.80	0.19	0.14	0.49
	LWM (Liu et al., 2024a)	7.2B	0.93	0.41	0.46	0.79	0.09	0.15	0.47
Und. and Gen. $\leq$ 2B	Janus (Wu et al., 2025)	2.1B	0.97	0.68	0.30	0.84	0.46	0.42	0.61
	Show-o (Xie et al., 2025)	1.5B	0.98	0.80	0.66	0.84	0.31	0.50	0.68
	Show-o-Clip-ViT (Xie et al., 2025)	1.6B	0.98	0.85	0.67	0.81	0.28	0.55	0.69
	JanusFlow (Ma et al., 2025)	2.1B	0.97	0.59	0.45	0.83	0.53	0.42	0.63
	Mobile-O-0.5B (Shaker et al., 2026)	1.6B	0.98	0.87	0.57	0.86	0.68	0.49	0.74
	UniOwl-0.5B (Ours)	1.7B	**0.99**	**0.89**	**0.78**	**0.92**	**0.75**	**0.65**	**0.83**

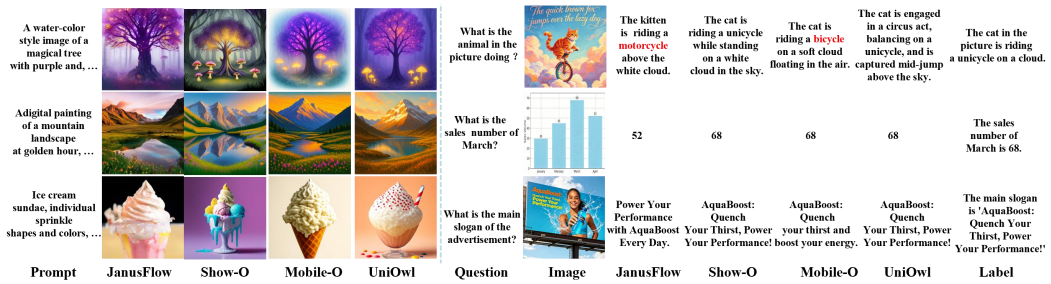


Figure 3: Visual comparison of text-to-image generation (left) and visual understanding (right) among different unified models. Our UniOwl generates images with more accurate semantic alignment and visual fidelity compared to the comparison models.

surpasses larger diffusion models such as SDXL (Podell et al., 2023) and PixArt-α (Chen et al., 2024a), and also improves over efficient generators such as SnapGen (Hu et al., 2024) and SANA-0.6B (Xie et al., 2024) in overall score. These results show that UniOwl can produce stronger compositional generations while keeping a compact unified architecture. The gains come from the proposed diffusion-adaptive semantic and visual conditioning. The semantic channel refines LLM-derived condition tokens across denoising steps, while the visual channel injects generation-state-aware visual tokens from the shared visual tokenizer. Together, they provide dynamic guidance for the DiT denoiser and improve semantic fidelity in generated images.

4.3 Ablation Study

Table 3: Ablation of dual condition architecture.

Base	Semantic	visual	Und.	Gen.
✓	✗	✗	60.6	0.67
✓	✓	✗	61.3	0.81
✓	✓	✓	62.3	0.83

Effects of dual-condition architecture in UniOwl. We conduct ablation studies to investigate the impact of diffusion-adaptive semantic conditioning and generation-driven visual conditioning. As shown in Table 3, we first evaluate the base semantic condition without refinement (Base), which provides static prompt-level guidance. Next, we enable diffusion-adaptive semantic conditioning (Variant 1) to refine the semantic tokens at each denoising step. The results show that this refinement improves understanding from 60.6 to 61.3 and generation from 0.67 to 0.81, indicating that step-aware semantic guidance helps the model better capture dynamic content during diffusion. We then incorporate generation-driven visual conditioning (Variant 2), which extracts visual tokens from the current noisy latent and fuses them with the refined semantic tokens. This further boosts understanding to 62.3 and generation to 0.83. The denoising dynamics of UniOwl are further illustrated in Fig. 4, which visualizes intermediate generation results for two different prompts. As denoising progresses, the generated content evolves from coarse structures to increasingly recognizable objects with more refined shapes and textures. This progression illustrates how UniOwl exploits the evolving generation state throughout the denoising process, complementing step-wise semantic guidance with

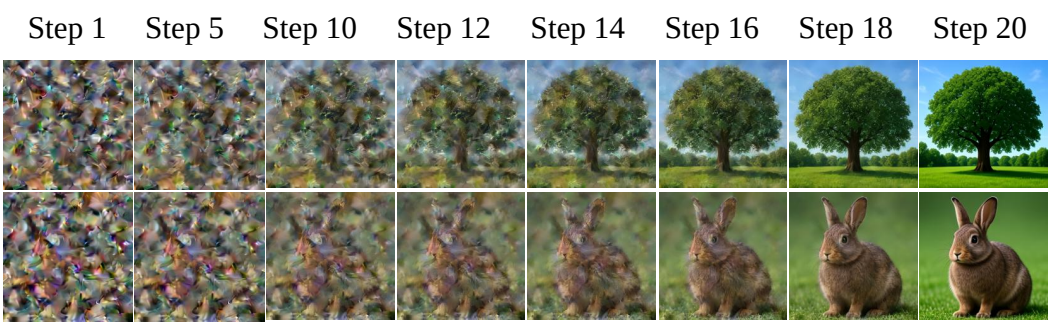


Figure 4: Visualization of intermediate generation steps under diffusion. The top and bottom rows show two different prompts progressively denoised by UniOwl. Early steps provide coarse structure and layout, while later steps refine local details. The images illustrate the effect of generation-driven visual conditioning: the evolving latent allows the model to adapt semantic guidance dynamically and produce more coherent and accurate images as denoising progresses.

visual information derived from the current noisy latent. Together with the quantitative improvements in Table 3, these results support the effectiveness of generation-driven visual conditioning. Overall, the ablation suggests that both components provide complementary benefits: diffusion-adaptive semantic conditioning provides dynamic semantic guidance, while generation-driven visual conditioning offers complementary visual feedback, enabling strong performance on both visual understanding and text-to-image generation benchmarks.

Effects of cross-task consistency learning. We perform an ablation to examine the impact of jointly training understanding and generation tasks with consistency learning. Without consistency learning, UniOwl achieves 61.4 on understanding and 0.76 on generation. Adding cross-task consistency learning improves understanding to 62.3 and generation to 0.83. The results suggest that the consistency objective not only improves generation quality but also benefits visual understanding, indicating that UniOwl can leverage generation feedback to strengthen shared multimodal representations.

Training Schemes. We conduct ablation studies to evaluate the impact of our staged training strategy. As shown in Table 4, Stage 1 pre-training alone provides a baseline performance of 60.5 for understanding and 0.58 for generation. Adding Stage 2 supervised fine-tuning improves generation significantly to 0.74, while understanding remains stable at 60.5. Incorporating Stage 3 unified training, where the model jointly learns understanding and generation on quadruplet samples p, x_{img}, q, a, further boosts both understanding to 62.3 and generation to 0.83. These results demonstrate that each stage contributes progressively: pre-training establishes the base condition extractor, supervised fine-tuning aligns semantic and visual features, and joint understanding-generation training effectively leverages multimodal interactions to maximize performance.

Table 4: Effect of training schemes.

Stage 1	Stage 2	Stage 3	Und.	Gen.
✓	✗	✗	60.5	0.58
✓	✓	✗	60.5	0.74
✓	✓	✓	62.3	0.83

5 Conclusion

In this paper, we introduce UniOwl, a unified framework for visual understanding and image generation. UniOwl overcomes the challenge of providing dynamic, consistent, and efficient diffusion conditioning while leveraging shared visual representations. Our Diffusion-Adaptive Semantic Conditioning adapts LLM-derived semantics to the current denoising state, and Generation-Driven Visual Conditioning aligns the visual tokenizer with generation supervision. A gated fusion module combines these channels, and Cycle Consistency Learning provides semantic-level consistency signals between generated images and the understanding pathway. Experiments show that UniOwl achieves competitive performance among compact unified models, demonstrating that dual-channel conditioning and cross-modal supervision effectively bridge perception and generation in a single compact model.

AI USE STATEMENT

In this work, we used generative AI tools to assist with migrating the manuscript into the ICLR 2027 LaTeX template and checking the resulting formatting. We reviewed all AI-assisted changes and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 3 describes the UniOwl architecture and training methodology. Section 4 provides implementation details and experimental results. Additional architecture and training data details are provided in Appendices A and B. These descriptions are intended to facilitate the reproduction of our results.

REFERENCES

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *eprint arXiv: 2308.12966*, 2023.

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

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *Proceedings of International Conference on Learning Representations*, 2024a.

Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation. *arXiv preprint arXiv:2506.18095*, 2025b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. Mobilevlm: A fast, strong and open vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*, 2023.

Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Proceedings of Advances in Neural Information Processing Systems*, pp. 49250–49267, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of International Conference on Learning Representations*, 2021.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of Advances in Neural Information Processing Systems*, pp. 6840–6851, 2020.

Dongting Hu, Jierun Chen, Xijie Huang, Huseyin Coskun, Arpit Sahni, Aarush Gupta, Anujraaj Goyal, Dishani Lahiri, Rajesh Singh, Yerlan Idelbayev, et al. Snapgen: Taming high-resolution text-to-image models for mobile devices with efficient architectures and training. *arXiv preprint arXiv:2412.09619*, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of International Conference on Machine Learning*, pp. 19730–19742, 2023b.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention. *arXiv preprint arXiv:2402.08268*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of Advances in Neural Information Processing Systems*, pp. 34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024b.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7739–7751, 2025.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Abdelrahman Shaker, Ahmed Heakl, Jaseel Muhammad, Ritesh Thawkar, Omkar Thawakar, Senmao Li, Hisham Cholakkal, Ian Reid, Eric P Xing, Salman Khan, et al. Mobile-o: Unified multimodal understanding and generation on mobile device. *arXiv preprint arXiv:2602.20161*, 2026.

Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. In *Advances in neural information processing systems*, pp. 49659–49678, 2023.

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

Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. Fastvit: A fast hybrid vision transformer using structural reparameterization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5785–5795, 2023.

Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokula Santhanam, James Gabriel, Peter Grasch, Oncel Tuzel, et al. Fastvlm: Efficient vision encoding for vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19769–19780, 2025.

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. *arXiv preprint arXiv:2409.18869*, 2024.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12966–12977, 2025.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *Proceedings of the International Conference on Learning Representations*, 2025.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1316–1324, 2018.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5907–5915, 2017.

Yichen Zhu, Minjie Zhu, Ning Liu, Zhiyuan Xu, and Yaxin Peng. Llava-phi: Efficient multi-modal assistant with small language model. In *Proceedings of the International Workshop on Efficient Multimedia Computing under Limited*, pp. 18–22, 2024.

APPENDIX

This supplementary material provides additional details and analyses to complement the main paper. We organize the content as follows:

- Section A: Detailed architecture of UniOwl.
- Section B: Training data and curriculum details.
- Section C: Additional qualitative results.
- Section D: Efficiency Analysis.
- Section E: Limitations and future work.

A ARCHITECTURE DETAILS

A.1 TWO-STAGE MCP CONDITIONING

The diffusion-adaptive semantic conditioning in UniOwl is implemented via a two-stage Mobile Conditioning Projector (MCP). We provide a detailed description of each stage here.

Stage 1: MCP Backbone. Given the LLM hidden states $\mathbf{H} \in \mathbb{R}^{B\times L\times d_{\text{llm}}}$, where L is the sequence length and d_{llm} is the LLM hidden dimension, the MCP backbone performs a one-time projection to produce base condition tokens:

$$\mathbf{C}^0 = f_{\text{mcp}}(\mathbf{H}) \in \mathbb{R}^{B\times N\times 2304}, \tag{9}$$

where N is the number of condition tokens and 2304 is the DiT's cross-attention dimension. f_{mcp} consists of a cross-attention layer that compresses the LLM sequence into N learnable query slots, followed by a feed-forward network. This stage runs only once per generation request, outside the denoising loop.

Stage 2: DiT Step-level Attention Residual. At each denoising timestep t, a lightweight refinement module computes a residual update conditioned on the current noisy latent x_t and timestep embedding $\mathbf{e}_t$:

$$\Delta\mathbf{C}_t = f_{\text{attn_res}}(\mathbf{C}^0,\, x_t,\, \mathbf{e}_t), \tag{10}$$

$$\mathbf{C}_t^{\text{ref}} = \mathbf{C}^0 + \gamma \cdot \Delta\mathbf{C}_t, \tag{11}$$

where γ is a scalar parameter initialized to zero. The refinement module $f_{\text{attn_res}}$ uses cross-attention where $\mathbf{C}^0$ serves as queries and the flattened DiT features derived from x_t serve as keys and values. The timestep embedding is injected via AdaLN conditioning. Zero-initialization of γ ensures that training starts from a stable fixed-condition baseline and the adaptive behavior is learned progressively.

Complexity Analysis. The two-stage design keeps the expensive LLM computation outside the denoising loop. Stage 1 is computed once with $O(L \cdot N)$ attention complexity. Stage 2 incurs only $O(N \cdot H_{\text{dit}})$ cost per step, where H_{dit} is the spatial dimension of the DiT feature map. This is significantly cheaper than re-running the full LLM at each denoising step.

A.2 GENERATION-DRIVEN VISUAL ALIGNMENT (GDVA)

The GDVA module extracts visual conditioning signals from the current noisy latent, ensuring train-inference consistency. The pipeline proceeds as follows:

VAE Decoding. The noisy latent x_t is decoded using the frozen VAE decoder: $\hat{\mathbf{I}}_t = \text{Dec}_{\text{VAE}}(x_t/s_{\text{vae}})$, where s_{vae} is the VAE scaling factor.

Normalization and Resizing. The decoded pixels are normalized to $[0, 1]$, resized to 384×384 via bilinear interpolation, and standardized with the vision encoder's image mean/std.

Visual Feature Extraction. The preprocessed image is passed through the shared FastViT encoder and the multimodal projector: $\mathbf{F}_t = f_{\text{proj}}(\text{FastViT}(\hat{\mathbf{I}}_t))$.

Table 5: Training data statistics per stage.

Stage	Dataset	# Samples	Task
Stage 1	JourneyDB (Sun et al., 2023)	4M	Generation
	BLIP3o-Short-Caption (Chen et al., 2025a)	5M	Generation
	Total	9M	Generation
Stage 2	BLIP3o (Chen et al., 2025a)	60K	Generation
	ShareGPT-4o-Image (Chen et al., 2025b)	45K	Generation
	Total	105K	Generation
Stage 3	BLIP3o (Chen et al., 2025a)	60K	Joint
	ShareGPT-4o-Image (Chen et al., 2025b)	45K	Joint
	Total	105K	Joint

Visual Conditioning Projection. A dedicated VisGenProjector further maps visual features to the DiT's conditioning space: $\mathbf{V}_t = f_{\text{vis_gen}}(\mathbf{F}_t) \in \mathbb{R}^{B\times N_{\text{vis}}\times 2304}$.

The VisGenProjector is a two-layer MLP with GeLU activation, which transforms the multimodal projector's output dimension to 2304 to match the DiT's cross-attention dimension.

A.3 GenHiGate Fusion Module

The fused conditioning token $\mathbf{C}_t^{\text{fuse}}$ is produced by a gated fusion module (GenHiGate) that combines the refined semantic tokens and visual tokens:

$$\mathbf{U}_t = \text{Concat}\big[\mathbf{C}_t^{\text{ref}},\ \mathbf{V}_t\big] \in \mathbb{R}^{B\times(N+N_{\text{vis}})\times 2304}, \tag{12}$$

$$\mathbf{g} = \sigma(f_{\text{gate_proj}}(\mathbf{U}_t)) \in \mathbb{R}^{B\times N\times 1}, \tag{13}$$

$$\mathbf{C}_t^{\text{fuse}} = \mathbf{g}\odot\tilde{\mathbf{C}}_t^{\text{ref}} + (1-\mathbf{g})\odot\tilde{\mathbf{V}}_t, \tag{14}$$

where $\tilde{\mathbf{C}}_t^{\text{ref}}$ and $\tilde{\mathbf{V}}_t$ are linearly projected versions of the semantic and visual tokens respectively, and σ denotes the sigmoid function. All projection parameters are initialized to zero, so the gate starts near 0.5 and the fusion is learned from data. The output $\mathbf{C}_t^{\text{fuse}} \in \mathbb{R}^{B\times N\times 2304}$ is the sole cross-attention conditioning input for the DiT denoiser.

B Training Data Details

B.1 Pre-training Data (Stage 1)

Stage 1 uses 9 million image-caption pairs to train the MCP Stage 1 backbone. The data comes from two sources:

- **JourneyDB** (Sun et al., 2023): A large-scale dataset of AI-generated images with high-quality captions covering diverse styles and content.
- **BLIP3o-Short-Caption** (Chen et al., 2025a): A subset of the BLIP3o dataset with concise, diverse image captions. Short captions are used at this stage to establish a stable semantic-to-visual alignment.

Only the generation loss $\mathcal{L}_{\text{diff}}$ is applied during Stage 1. The LLM, visual encoder, and VAE are all frozen, so only the MCP backbone parameters are updated.

B.2 Supervised Fine-tuning Data (Stage 2)

Stage 2 uses 105K high-quality image-caption pairs from:

- **BLIP3o** (Chen et al., 2025a) (60K samples): Long, detailed captions generated by a strong captioner, covering complex compositional scenes.

- **ShareGPT-4o-Image** (Chen et al., 2025b) (45K samples): Detailed captions produced by GPT-4o, providing rich descriptions of fine-grained visual attributes.

This stage introduces MCP Stage 2 (the DiT-step attention residual) and continues to train with only the generation loss.

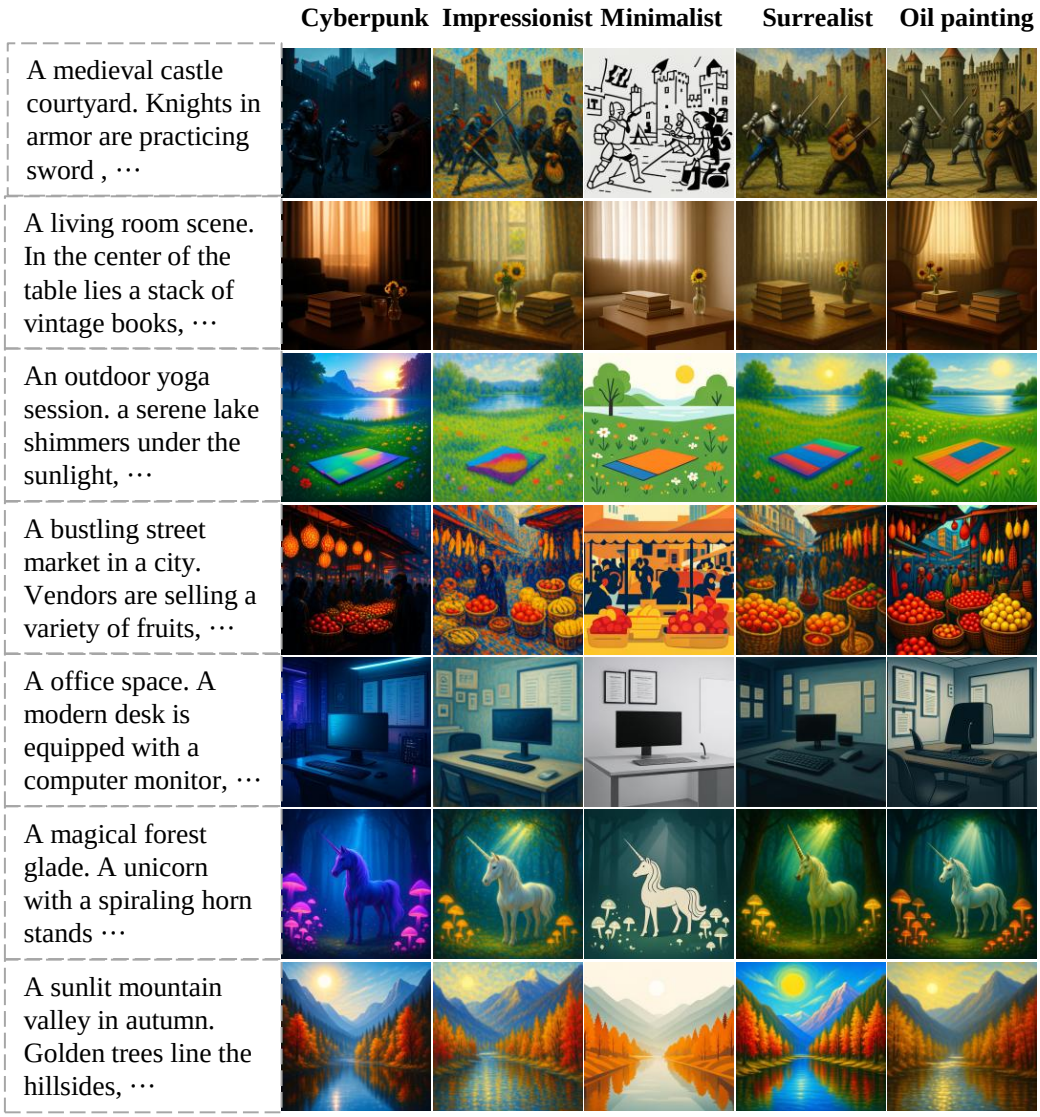


Figure 5: Additional text-to-image generation results from UniOwl. The prompts cover different scenes, themes, and styles. UniOwl consistently produces semantically faithful images across these aspects.

B.3 Unified Training Data (Stage 3)

Stage 3 constructs training samples as quadruplets $S = \{p, x_{\text{img}}, q, a\}$, where p is a generation prompt, x_{img} is the image, and (q, a) is a visual question-answer pair associated with the image. This design ensures that each training example simultaneously provides a generation target and an understanding supervision signal. The data sources are the same as Stage 2, but the samples are reorganized to form these quadruplets. Both the generation loss and the cycle consistency loss are applied during this stage. Table 5 summarizes the dataset statistics across all stages.



Question: Describe the scene in the image.

Answer: It is a busy urban street in daytime. There are several storefronts along the road: a cafe with a coffee sign and a bookstore A little child is eating ice cream beside him.



Question: What is the main slogan at the top of the poster?

Answer: Self-discipline gives me freedom, Sculpt yourself, reshape the future.



Question: What is the main message of this eco-friendly poster?

Answer: Protect Our Planet, Save Our Future..



Question: What drink is shown on the advertisement?

Answer: A cool summer drink with lemon, mint and ice.

Figure 6: Qualitative results of UniOwl on visual understanding tasks. UniOwl produces accurate and contextually grounded answers to diverse visual questions, demonstrating effective understanding capability within the unified generation-understanding framework.

C Additional Qualitative Results

Figure 5 shows additional text-to-image generation samples from UniOwl. The prompts cover a wide range of scenes, themes, and artistic styles. UniOwl consistently produces semantically faithful images that align well with the prompts across these diverse aspects, demonstrating the effectiveness of the adaptive conditioning framework in handling complex generation requests.

Figure 6 presents qualitative results of UniOwl on visual understanding tasks. Given an input image and a natural language question, UniOwl leverages its cycle understanding path to produce accurate and contextually grounded answers. The results demonstrate that the model maintains strong understanding capability while jointly supporting image generation, validating the effectiveness of the unified training framework.

Table 6: Efficiency comparison of different conditioning strategies.

Method	Semantic Adapt.	Visual Feedback	Latency ↓ (ms)	Memory ↓ (GB)
Static Condition	×	×	725.9 ± 1.0	3.59
Step-wise LLM	✓	×	1201.0 ± 4.0	3.60
UniOwl-Latent	✓	×	**771.9 ± 9.5**	**3.59**
UniOwl-Lite ($k = 4$)	✓	Partial	1230.4 ± 0.5	4.03
UniOwl-Full	✓	Full	2725.0 ± 24.8	4.03

D Efficiency Analysis

We further analyze the computational cost of different conditioning strategies during inference. All measurements are conducted on a single NVIDIA A100-80G GPU with batch size 1 and 20 diffusion steps. Static Condition uses fixed semantic conditions throughout denoising, while Step-wise LLM recomputes language conditions at each diffusion step. UniOwl-Latent removes the visual conditioning pathway and evaluates the overhead introduced by timestep-aware semantic refinement alone, while UniOwl-Full additionally performs visual conditioning during denoising. UniOwl-Lite reduces the visual update frequency by refreshing visual conditions every $k = 4$ denoising steps. As shown in Table 6, UniOwl-Latent introduces only 6.3% latency overhead compared with static conditioning (771.9 ms vs. 725.9 ms), while avoiding repeated LLM execution during denoising and achieving 1.56× lower latency than step-wise LLM conditioning (771.9 ms vs. 1201.0 ms). These results demonstrate that timestep-aware semantic refinement provides dynamic conditioning with limited computational overhead. The additional cost of UniOwl-Full mainly comes from the visual conditioning pathway, including VAE decoding and visual tokenization, rather than the adaptive semantic refinement module itself. UniOwl-Latent and UniOwl-Lite therefore provide alternative configurations to explore the trade-off between efficiency and visual feedback.

E Limitations and Future Work

While UniOwl demonstrates the effectiveness of diffusion-adaptive conditioning, several limitations remain to be addressed. First, the current framework is evaluated with a 0.5B-scale language model (Qwen2-0.5B), and whether the proposed adaptive conditioning framework scales effectively to larger unified models remains to be explored. Second, the visual conditioning pathway introduces additional computation overhead due to VAE decoding and visual re-tokenization during denoising. Although latent-space and adaptive-refresh strategies can reduce this cost, further improvements in efficient dynamic conditioning are needed. Third, the current system focuses on single-image understanding and generation at 512×512 resolution. Extending UniOwl to higher-resolution generation, video generation, or more complex multi-image scenarios requires additional temporal and spatial modeling capabilities. Finally, the current cycle consistency objective operates at the token-distribution level through KL divergence on LLM output logits, providing a semantic-level consistency signal rather than enforcing pixel-level or structural consistency. More advanced objectives, such as perceptual or structured alignment, may further improve generation-understanding alignment. We also observe that UniOwl can still produce occasional cross-modal hallucinations in the cycle understanding pathway, motivating future work on stronger hallucination mitigation and more reliable unified multimodal training.