DYNIN-OMNI: OMNIMODAL UNIFIED LARGE DIFFUSION LANGUAGE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Omnimodal unified models aim to understand text, images, video, and speech and generate coordinated text, image, and speech outputs. The challenge is to extend a single backbone to new modalities efficiently while preserving established capabilities and enabling coordinated cross-modal generation. We present Dynin-Omni, an open-source 8B-scale native omnimodal masked-diffusion model that unifies these capabilities within a single token-level backbone under a shared masked-token prediction objective. We propose a three-stage training recipe for effective omnimodal adaptation that first aligns new modalities without replaying the full task mixture, then applies modality-disentangled merging to recover inherited capabilities. Scheduled padding supervision delays termination learning until semantic grounding is established, enabling flexible-length generation. Moreover, Dynin-Omni is the first masked-diffusion model to natively generate interleaved text, image, and speech within a single reverse diffusion process, enabling mutual conditioning and joint refinement without modality-wise orchestration. Across 20 benchmark settings covering text, image, video, speech, and joint omnimodal generation, Dynin-Omni outperforms prior unified systems on a broad range of directly comparable benchmarks while remaining competitive with modality-specific experts. Stage-wise and controlled analyses show capability recovery and improved length control, while native joint decoding approaches the quality of the best evaluated two-stage generation configuration with a lower sampling budget.

1 INTRODUCTION

The pursuit of a *universal interface* has driven the evolution of large-scale foundation models from text-only reasoning toward multimodal understanding and generation. A natural next step is *omnimodal* modeling, where text, vision, and speech serve as both inputs and generative outputs within a single system. This is crucial for building agents that can understand and respond across modalities through a unified interface (Wu et al., 2024; Zhan et al., 2024; Tang et al., 2024). In practice, the value of an omnimodal model lies in both strong performance on individual tasks and the ability to produce coordinated multimodal responses to mixed inputs.

Fully autoregressive (AR) models excel at sequence modeling (Yang et al., 2025a; Zeng et al., 2025), but typically serialize heterogeneous outputs and decode them left-to-right (Lu et al., 2024a; Wang et al., 2024a; Zhan et al., 2024), fixing earlier outputs before later ones can inform them and limiting bidirectional context, parallel decoding, and reciprocal cross-modal refinement (Zhao et al., 2025; Shi et al., 2025). Hybrid models combine AR reasoning with diffusion-based generation or modality-specific experts (Zhou et al., 2024; Xie et al., 2024; 2025; Cui et al., 2025; Deng et al., 2025), but typically use distinct objectives or modality-specific update rules that can restrict joint revision; compositional systems further require external generator handoffs (Wu et al., 2024; Team, 2026). Yet sharing a generative process is insufficient if learning new modalities erodes established capabilities. This raises a central question: *Can a single generative process extend to new modalities while preserving existing capabilities and enabling jointly generated outputs to refine one another?*

Masked diffusion (Sahoo et al., 2024; Nie et al., 2025) offers bidirectional context and parallel refinement without a fixed causal order, providing a natural foundation for interleaved outputs whose modalities mutually condition one another. Recent multimodal diffusion models have extended this paradigm beyond text (Yang et al., 2025b; Xin et al., 2025; Bethune et al., 2026; Li et al., 2026). Yet

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.

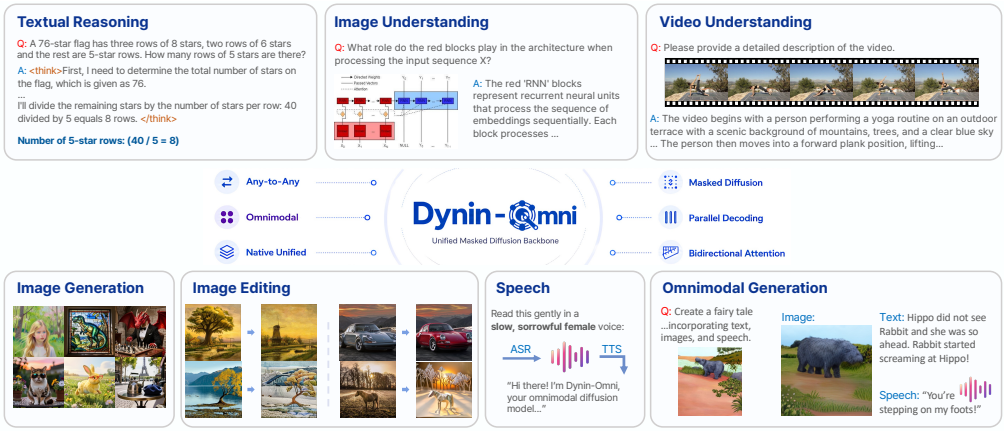


Figure 1: Overview of Dynin-Omni's supported tasks across text, image, video, and speech.

efficiently expanding a shared backbone to new modalities while preserving its existing capabilities remains challenging. Addressing this training problem is central to achieving strong individual-task performance and native interleaved text–image–speech generation in the same model.

We introduce Dynin-Omni, an open-source masked-diffusion model to unify text and image understanding and generation, video understanding, and speech understanding and generation within a single 8B-scale token-level backbone. As shown in Fig. 1, Dynin-Omni represents text, image, video, and speech as discrete tokens in a shared vocabulary and models them under a unified masked-token prediction objective. At inference time, this shared formulation accommodates modality-aware decoding schedules, including block-wise decoding for text and speech and fully parallel decoding for images. The backbone and objective remain unified, while decoding schedules adapt to each modality. Moreover, Dynin-Omni is the first masked-diffusion model to natively generate interleaved text, image, and speech outputs within a single diffusion process, allowing them to mutually condition one another and be jointly refined without external generators or orchestration.

Adding new modality tokens alone is not enough. Training on new modalities can weaken existing capabilities, while learning when to stop before learning what to generate can lead to prematurely short outputs. We therefore adopt a three-stage progressive training recipe. Newly introduced modalities are first aligned through text-centric supervision without replaying the full task mixture, using language as a semantic anchor for video and speech tokens. To preserve inherited capabilities during this expansion, we introduce two core components: *modality-disentangled merging*, which separately handles shared and vocabulary-dependent parameters after vocabulary expansion to retain inherited representations while incorporating new-modality tokens, and *scheduled padding supervision*, which delays termination-token prediction so that semantic alignment is learned before length control. The stages then separate new-modality grounding, consolidation of inherited and newly acquired capabilities, and training for harder individual tasks and joint outputs.

Extensive evaluations show that Dynin-Omni achieves strong and balanced performance across textual reasoning, image understanding, generation and editing, video understanding, speech recognition and synthesis, and joint omnimodal generation. Compared with prior unified models (Xie et al., 2025; Team, 2026; Luo et al., 2025b), it achieves stronger results on a broad range of directly comparable benchmarks while remaining competitive with modality-specific experts (Yang et al., 2025a; Wu et al., 2025a) and perception-centric models (Xu et al., 2025a) on representative benchmarks. Stage-wise evaluations show recovery of inherited text–image abilities, while controlled comparisons isolate the effects of modality-disentangled merging and scheduled padding supervision. On Omni-StoryBench (AIDAS-Lab, 2026), Dynin-Omni surpasses orchestration-based systems without external image or speech generators.

Our contributions are summarized as follows:

- **Native omnimodal masked diffusion.** We introduce Dynin-Omni, the first open-source 8B-scale native omnimodal masked-diffusion model and the first masked-diffusion model to natively generate interleaved text, image, and speech within a single reverse diffusion process.

- **Training recipe for omnimodal expansion and capability preservation.** We propose a three-stage recipe with modality-disentangled merging and scheduled padding supervision to integrate new modalities, preserve inherited capabilities, and stabilize flexible-length generation. Stage-wise evaluations that track capability recovery, while controlled comparisons isolate both components.
- **Broad and balanced performance.** Across 20 benchmark settings, Dynin-Omni outperforms prior unified systems on a broad range of directly comparable benchmarks and orchestration-based systems in joint generation, while remaining competitive with modality-specific experts on representative benchmarks.

2 DYNIN-OMNI

2.1 NATIVE OMNIMODAL TOKEN SPACE

Building on LLaDA (Nie et al., 2025) and MMaDA (Yang et al., 2025b), Dynin-Omni models text, image, video input, and speech input/output as discrete tokens using a single masked-diffusion backbone with shared embeddings and a unified prediction head (Fig. 2). The vocabulary $\mathcal{V} = \mathcal{V}_{\text{text}} \cup \mathcal{V}_{\text{vision}} \cup \mathcal{V}_{\text{speech}}$ extends MMaDA's text–image vocabulary with EMOVA speech tokens (Chen et al., 2024a), while sampled video frames reuse the existing MAGViT-v2 vision tokens (Yang et al., 2025b; Xie et al., 2024). All tokenizers and detokenizers are frozen, with detokenizers used only at inference; training therefore requires no reconstruction loss or modality-specific generative objective. Tokenization details are provided in App. D.

2.2 UNIFIED MASKED DIFFUSION OBJECTIVE

Dynin-Omni formulates omnimodal modeling as masked diffusion over a single discrete token sequence $\mathbf{x}_0 = (x_1, \ldots, x_L)$ drawn from the joint vocabulary $\mathcal{V}$. The sequence may interleave tokens from multiple modalities and is processed by the backbone Transformer with bidirectional attention.

Training objective. Following masked diffusion (App. B) (He et al., 2023; Sahoo et al., 2024; Nie et al., 2025), we sample a masking ratio $t \sim \mathcal{U}(0, 1)$ and independently replace each token in $\mathbf{x}_0$ with a special `[MASK]` token with probability t, yielding a corrupted sequence $\mathbf{x}_t$. The model is trained to reconstruct masked tokens conditioned on the visible context by minimizing the negative log-likelihood over masked positions:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}_0,\, t,\, \mathbf{x}_t} \left[-\frac{1}{t} \sum_{i=1}^{L} \mathbb{I}[x_{t,i} = \texttt{[MASK]}] \log p_\theta(x_{0,i} \mid \mathbf{x}_t) \right]. \tag{1}$$

This objective is shared across all modalities within the unified token space, without modality-specific heads or reconstruction objectives, enabling fully unified omnimodal training under a single objective (Eq. 1). For flexible-length text and speech generation, the model also learns to predict a special `[EOS]` token, which serves as the padding and termination token (Nie et al., 2025).

Tasks as masking patterns. A central consequence of this formulation is that task identity is specified entirely by the observed–masked partition of the sequence. Decomposing $\mathbf{x}_0 = [\mathbf{x}^{\text{cond}}, \mathbf{x}^{\text{tar}}]$ into observed conditioning tokens and masked target tokens, every task reduces to

$$p_\theta(\mathbf{x}^{\text{tar}} \mid \mathbf{x}^{\text{cond}}) = \prod_{i:m_i=1} p_\theta(x_{0,i} \mid \mathbf{x}_{\mathbf{m}}), \tag{2}$$

where $\mathbf{m} \in \{0, 1\}^L$ is a binary task mask ($m_i = 1$ for target, $m_i = 0$ for observed). For example, Visual-Question-Answering (VQA) observes image and question tokens while masking the answer ($\mathbf{x}_0 = [\mathbf{x}^I, \mathbf{x}^Q, \mathbf{x}^A]$); text-to-image generation observes text and masks image tokens ($[\mathbf{x}^T, \mathbf{x}^I]$). Modality transfer, understanding, and generation thus all collapse to a single conditional masked prediction problem under the same backbone and denoising objective (Tab. 6).

2.3 OMNIMODAL INFERENCE

At inference time, Dynin-Omni performs understanding and generation by initializing target tokens as `[MASK]` and iteratively predicting masked positions conditioned on the observed context. Although

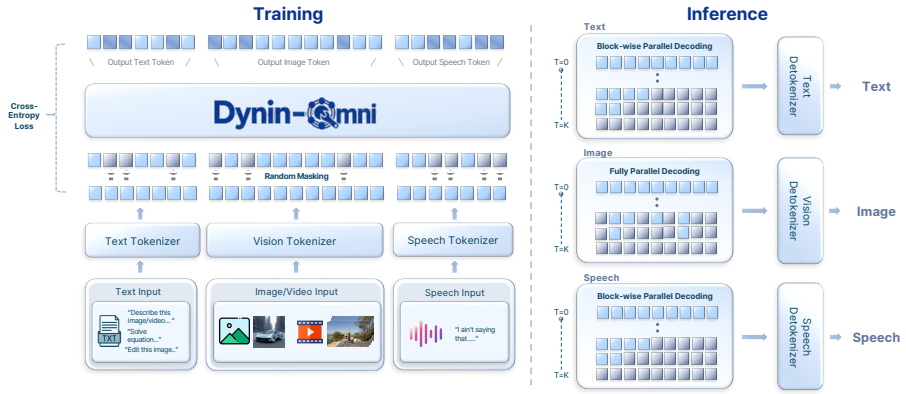


Figure 2: Training and inference pipeline of Dynin-Omni. All modalities are discretized into a shared token space and trained with a unified masked-diffusion objective. At inference, Dynin-Omni applies modality-aware parallel decoding followed by deterministic detokenization.

the backbone and reverse-diffusion process are shared, decoding is modality-aware: text and speech use block-wise refinement to accommodate their sequential structure, while image tokens are refined fully in parallel. Confidence-based remasking revisits uncertain tokens within each schedule (Chang et al., 2022; Nie et al., 2025; Ye et al., 2025c). For faster inference, we combine KV-cache approximation and confidence-based parallel decoding in a compiled implementation (Wu et al., 2026; Ma et al., 2025b), reaching 183.7 tokens per second on GSM8K with a single A100 GPU. Configurations and further efficiency results are provided in App. I.

3 Efficient and Stable Omnimodal Training

3.1 Overall Training Paradigm

To expand the model without replaying the full omnimodal training mixture from the outset, we first focus on aligning newly introduced modalities. This selective training can weaken inherited capabilities, while learning when to terminate before new-modality content is grounded can lead to prematurely short outputs. We address these challenges with a three-stage recipe (Fig. 9). Stage 1 aligns video and speech through text-centric supervision (Chen et al., 2024a), while deferring termination supervision. Stage 2 merges the adapted model with the original backbone and performs omnimodal supervised fine-tuning, restoring termination supervision after new-modality alignment. Stage 3 scales task difficulty, resolution, and sequence length while introducing joint omnimodal generation. Modality-disentangled merging helps retain inherited capabilities, while scheduled padding supervision separates new-modality alignment from flexible-length learning. Data and training details are in Apps. E and F, with stage-wise results in §5.1.

3.2 Modality-Disentangled Model Merging

Challenge. Stage 1 aligns new modalities without replaying the full omnimodal mixture, risking catastrophic forgetting of inherited capabilities (Kirkpatrick et al., 2017). Parameter merging can combine complementary capabilities (Wortsman et al., 2022; Ilharco et al., 2022; Li et al., 2025a), but directly merging the original backbone $\theta^{(0)}$ and Stage 1 model $\theta^{(1)}$ is ill-defined for token embeddings and the output head, whose dimensions change with vocabulary expansion.

Method. At the transition to Stage 2, we apply *modality-disentangled model merging* by separating vocabulary-independent and token-indexed parameters. Let $\theta = \{\theta_{\text{shared}}, \theta_{\text{vocab}}\}$ denote the Transformer parameters and the token embeddings and output head, respectively. We interpolate the shared parameters:

$$\theta^{\text{merge}}_{\text{shared}} = (1-\alpha)\theta^{(0)}_{\text{shared}} + \alpha\theta^{(1)}_{\text{shared}}. \tag{3}$$

For vocabulary-dependent parameters, existing token rows $i \in \mathcal{V}^{(0)}$ come from $\theta^{(0)}_{\text{vocab}}$, while new-token rows $i \in \mathcal{V}^{\text{new}}$ come from $\theta^{(1)}_{\text{vocab}}$. We use $\alpha = 0.4$. The merged initialization lowers Stage 2

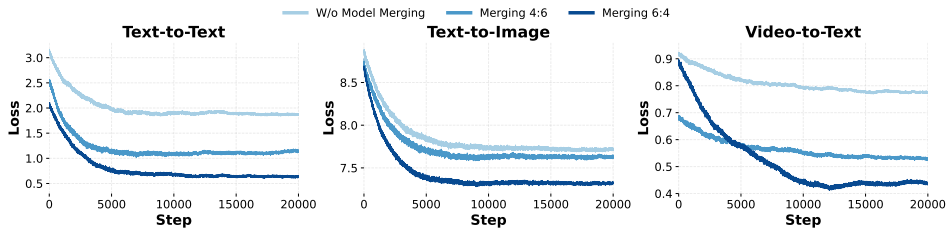
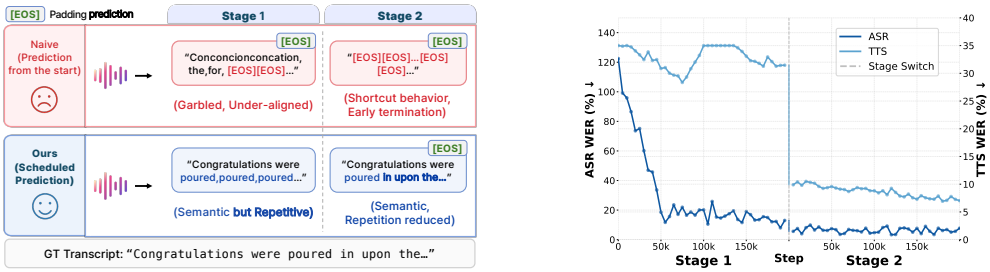


Figure 3: Stage-2 training loss across original (text-to-text and text-to-image) and new (video-to-text) tasks, where modality-disentangled merging accelerates convergence and reduces final loss.



(a) Qualitative effect of scheduled [EOS] supervision.

(b) ASR and TTS WER across training stages.

Figure 4: Effect of scheduled padding prediction. (a) Scheduled padding ([EOS]) supervision avoids early shortcut behavior, and Stage 1 repetition is rapidly stabilized once [EOS] prediction is enabled in Stage 2. (b) Adaptation stabilizes after Stage 2, confirming improved flexible-length generation.

loss (Fig. 3). After matched Stage 2 training, it also improves inherited and new-modality benchmarks (Tab. 4(b)). Alternative merging strategies and interpolation ratios are analyzed in §5.1.

3.3 Scheduled Padding Supervision

Challenge. Flexible-length generation in masked diffusion uses explicit length control (Nie et al., 2025; Bie et al., 2025; You et al., 2025). In our formulation, the padding token is tied to [EOS] and therefore determines output termination. To mitigate padding overprediction, Omni-Diffusion (Li et al., 2026), for example, reduces the masking probability of padding tokens. In contrast, we investigate when termination supervision should be introduced during modality adaptation, using matched runs that hold initialization, training data, and update counts fixed while varying only [EOS] supervision at each stage. These controls reveal a stage-dependent trade-off: enabling termination supervision before new-modality grounding causes premature stopping, whereas withholding it after alignment causes repetition and decoding-cap hits (Fig. 4a). Effective modality expansion therefore benefits from learning semantic alignment before output termination.

Method. We implement scheduled padding supervision by applying a stage-dependent weight to each token's contribution to the unified masked-diffusion objective in Eq. 1:

$$w_i^{(s)} = \begin{cases} 0, & s = 1 \text{ and } x_{0,i} = \texttt{[EOS]}, \\ 1, & \text{otherwise.} \end{cases} \tag{4}$$

Thus, Stage 1 retains supervision for content tokens while deferring termination targets during new-modality alignment. From Stage 2 onward, [EOS] supervision is restored to learn flexible-length generation under the same backbone and objective. Deferring [EOS] supervision reduces premature termination in Stage 1; restoring it in Stage 2 improves ASR/TTS WER and reduces repetition and decoding-cap hits (Fig. 4b; Tab. 4(c)).

Table 1: Evaluation on reasoning and understanding benchmarks across text, image, and video. Gray indicates unsupported modalities; "–" denotes supported but unreported or officially unused results. * Video generation support. † Native speech input and output. ‡ Reproduced results.

	Text↑				Image Understanding↑				Video Understanding↑			
Model	MMLU	GSM8K	MATH	GPQA	POPE	MME-P	MMMU	MMBench	MVBench	Video-MME	NExT-QA	TempCompass
Modality-specific Experts												
Llama-3-8B (Grattafiori et al., 2024)	66.6	55.3	20.5	25.8								
LLaDA-8B-Base‡ (Nie et al., 2025)	65.9	71.3	30.0	31.5								
LLaDA-8B-Instruct‡ (Nie et al., 2025)	65.4	68.7	29.8	31.8								
Qwen3-8B (Yang et al., 2025a)	76.9	89.8	60.8	44.4								
InternVL3-8B (Zhu et al., 2025)	–	–	–	–	90.4	1748.4	62.2	83.6	73.2	66.0	82.0‡	70.4‡
Ovis2-8B (Lu et al., 2024b)	–	–	–	–	88.6	1628.1	57.4	84.8				
Qwen2.5-VL-7B (Bai et al., 2025)	–	–	–	–	85.9	1698.1	58.0	83.2	69.6	65.1	–	–
LLaDA-V (You et al., 2025)	–	–	–	–	–	1507.0	48.6	82.9	53.1	56.4	–	–
LLaVA-OV (Li et al., 2024a)	–	–	–	–	–	1580.0	48.8	80.8	56.7	58.2	79.4	64.8
InternVL2.5-8B (Chen et al., 2024c)	–	–	–	–	90.6	–	56.0	83.2	72.0	64.2	84.1‡	67.7‡
Perception-centric												
OpenOmni (Luo et al., 2025a)	–	–	–	–	86.4‡	1536.9‡	46.7	76.2				
OmniVinci (Ye et al., 2025a)	–	–	–	–	89.5‡	1651.0‡	49.7	–	70.6	68.2	–	–
Baichuan-Omni-1.5 (Li et al., 2025d)	72.2	–	–	–	87.3‡	1632.1‡	53.9	85.6	63.7	60.1	–	–
MiniCPM-o-4.5 (Yao et al., 2024)	77.0	94.5	77.0	–	–	–	67.6	87.6	–	70.4	–	–
Qwen2.5-Omni-7B (Xu et al., 2025a)	71.8	88.7	71.5	30.8	86.8‡	1481.3‡	59.2	81.8	70.3	64.3	76.4‡	69.3‡
Unified												
Chameleon-7B (Team, 2024)	52.1	41.6	11.5	–	–	–	28.4	35.7				
SEED-X (Ge et al., 2024)	–	–	–	–	84.2	1435.7	35.6	70.1				
Emu3* (Wang et al., 2024a)	–	–	–	–	85.2	–	31.6	58.5	–	–	–	–
MMaDA (Yang et al., 2025b)	68.4	73.4	36.0	28.4	86.1	1410.7	30.2	68.5				
Fudoki (Wang et al., 2025a)	–	–	–	–	86.1	1485.4	34.3	73.9				
Janus-Pro (Chen et al., 2025b)	–	–	–	–	87.4	1567.1	41.0	79.2				
BAGEL (Deng et al., 2025)	–	–	–	–	–	1687.0	55.3	85.0				
Lumina-DiMOO (Xin et al., 2025)	–	–	–	–	87.4	1534.2	58.6	83.1				
Show-o2* (Xie et al., 2025)	70.7	75.3	–	31.5	–	1620.5	48.9	79.3	55.8	57.4	79.0	–
NExT-OMNI*,† (Luo et al., 2025b)	–	–	–	–	87.4	1537.8	43.7	78.9	–	–	–	–
HyperCLOVAX† (Team, 2026)	75.7	87.3	43.4‡	28.6‡	79.8‡	1307.0‡	31.0‡	65.9‡	49.5‡	58.2	–	58.3‡
Dynin-Omni† (Ours)	75.9	83.7	47.4	37.9	88.0	1672.9	52.8	74.7	66.4	63.4	81.3	66.7

4 EVALUATION

We evaluate Dynin-Omni across textual reasoning, image understanding, generation, and editing, video understanding, speech recognition and synthesis (ASR/TTS), and joint omnimodal generation. We include three groups of baselines: (i) modality-specific experts, (ii) perception-centric omnimodal models, and (iii) unified models. Among them, HyperCLOVAX (Team, 2026) and NExT-OMNI (Luo et al., 2025b) are the most direct baselines, as they support unified omnimodal understanding and generation. Baseline comparison is provided in App. C.

4.1 MAIN RESULTS

Tab. 1 and 2 summarize representative benchmarks across reasoning, image understanding and synthesis, video understanding, and ASR/TTS. Detailed inference configurations are provided in App. G. Extended benchmark results, direct comparisons with text-diffusion and diffusion vision–language models, and evaluation uncertainty including perceptual TTS quality are provided in App. I.

Comparison with modality-specific experts. Modality-specific experts achieve the strongest peak performance in their respective domains, including Qwen3-8B (Yang et al., 2025a) for text reasoning, InternVL-series models (Chen et al., 2024c; Zhu et al., 2025) for multimodal understanding, Qwen-Image (Wu et al., 2025a) for image synthesis, and dedicated ASR/TTS systems (Shi et al., 2026; Hu et al., 2026). Nevertheless, Dynin-Omni remains competitive across domains: it approaches Qwen3-8B on MMLU, outperforms Qwen2.5-VL (Bai et al., 2025) on POPE, approaches it on MME-P, matches Qwen-Image on GenEval, and achieves strong ASR and TTS performance. While specialized models retain advantages on mathematical reasoning and SEED-TTS, Dynin-Omni provides balanced performance across these domains within a single backbone.

Comparison with perception-centric omnimodal models. Perception-centric omnimodal models, such as Qwen2.5-Omni (Xu et al., 2025a), MiniCPM-o (Yao et al., 2024), and OmniVinci (Ye et al., 2025a), achieve strong understanding and speech performance, but provide limited visual generation capability within a unified token-level generative process. Dynin-Omni remains competitive with these models on reasoning, image understanding, video understanding, and speech benchmarks,

Table 2: Evaluation on image synthesis and speech benchmarks. Notation follows Tab. 1.

	Image Generation / Editing↑			ASR↓		TTS↓	
Model	GenEval	DPG-Bench	ImgEdit	$\text{LibriSpeech}_{\text{clean}}$	$\text{LibriSpeech}_{\text{other}}$	$\text{LibriSpeech}_{\text{clean}}$	SEED-TTS
Modality-specific Experts							
Seedream 3.0 (Gao et al., 2025)	0.84	88.3					
GPT-Image 1 (OpenAI, 2025)	0.84	85.2	4.28				
Qwen-Image (Wu et al., 2025a)	0.87	88.3	4.33				
Flux.1 Kontext (Labs et al., 2025)	–	–	3.59				
Qwen2-Audio (Chu et al., 2024)				1.6	3.6		
Qwen3-ASR (Shi et al., 2026)				1.6	3.4		
MaskGCT (Wang et al., 2024b)						2.6	2.6
Llasa-8B (Ye et al., 2025e)						1.5	3.0
Qwen3-TTS (Hu et al., 2026)						6.6‡	1.2
Perception-centric							
EMOVA-7B (Chen et al., 2024a)				4.0	–	3.6	–
OpenOmni (Luo et al., 2025a)				3.1	7.0	3.4	–
MiniCPM-o-4.5 (Yao et al., 2024)				1.4	2.8	–	2.4
Qwen2.5-Omni-7B (Xu et al., 2025a)				1.8	3.4	7.0‡	2.3
Unified							
Emu3* (Wang et al., 2024a)	0.66	81.6					
MMaDA (Yang et al., 2025b)	0.63	70.0					
Fudoki (Wang et al., 2025a)	0.77	83.6					
Janus-Pro (Chen et al., 2025b)	0.80	84.2					
Show-o2* (Xie et al., 2025)	0.76	86.1					
Lumina-DiMOO (Xin et al., 2025)	0.88	86.0	–				
BAGEL (Deng et al., 2025)	0.82	85.1	3.30				
UniWorld-V1 (Lin et al., 2025)	0.79	–	3.30				
LaViDa-O (Li et al., 2025b)	0.77	81.8	3.81				
NExT-OMNI*,† (Luo et al., 2025b)	0.85	84.2		3.1	7.0	3.1	–
HyperCLOVAX† (Team, 2026)	0.64	60.8‡	3.83	1.9	4.5‡	7.9‡	7.7‡
Dynin-Omni† (Ours)	0.87	85.5	3.82	2.1	4.8	2.1	4.3

Table 3: Evaluation on omnimodal generation. Overall averages the six metrics, with 0–1 metrics scaled by 10. Notation follows Tab. 1.

		Generator		Text↑		Image↑		Speech↑		
Orchestrator	**Joint Gen.**	Image	Speech	BERTScore	LLM Judge	CLIP Sim.	LLM Judge	Metadata Acc.	LLM Judge	**Overall↑**
LLM										
InternVL3.5-4B (Wang et al., 2025b)	×	Flux.1 K (Labs et al., 2025)	Parler (Lyth & King, 2024)	0.87	7.44	0.76	5.43	0.55	8.06	7.12
Qwen3.5-4B (Qwen Team, 2026)	×	Flux.1 K (Labs et al., 2025)	Parler (Lyth & King, 2024)	0.88	8.24	0.77	6.42	0.52	8.22	7.43
Omnimodal Perception										
EMOVA-7B (Chen et al., 2024a)	×	Flux.1 K (Labs et al., 2025)	–	0.87	7.78	0.78	4.73	0.52	8.23	7.06
Qwen2.5-Omni-7B (Xu et al., 2025a)	×	Flux.1 K (Labs et al., 2025)	–	0.88	7.97	0.76	5.57	0.54	8.07	7.24
Multimodal Unified										
Emu3 (Wang et al., 2024a)	×	–	Parler (Lyth & King, 2024)	0.87	5.48	0.58	1.78	0.51	7.91	5.81
MMaDA (Yang et al., 2025b)	×	–	Parler (Lyth & King, 2024)	0.87	4.79	0.65	1.96	0.49	8.03	5.82
Omnimodal Unified										
AnyGPT (Zhan et al., 2024)	×	–	–	0.80	3.24	0.64	1.66	0.42	5.34	4.81
HyperCLOVAX† (Team, 2026)	×	–	–	0.87	6.43	0.68	2.50	0.53	8.14	6.32
Dynin-Omni† (Ours)	✓	–	–	0.92	6.78	0.86	6.40	0.61	8.70	7.63

while additionally supporting native image generation and editing under the same masked-diffusion objective. This shows that Dynin-Omni extends omnimodal systems beyond perception-centric capability while retaining competitive understanding performance.

Comparison with unified models. Unified models differ in both modality coverage and degree of architectural unification. Compared with image–text unified models such as MMaDA (Yang et al., 2025b) and BAGEL (Deng et al., 2025), Dynin-Omni extends coverage to video understanding and speech input/output while remaining competitive or stronger on shared benchmarks, including MME-P, GenEval, and ImgEdit. Compared with compositional omnimodal systems such as HyperCLOVAX-8B-Omni (Team, 2026), which rely on external generators, Dynin-Omni achieves stronger performance on several shared benchmarks, while using a single native backbone. These results show that native masked diffusion can scale to broader omnimodal modeling.

4.2 Omnimodal Generation

We further evaluate Dynin-Omni on Omni-StoryBench (AIDAS-Lab, 2026), a story-grounded benchmark in which models generate coordinated narration, scene images, and spoken utterances from a given context. Tab. 3 compares different approaches to constructing omnimodal generation systems.

Table 4: Evaluation of the three-stage training recipe: (a) benchmark performance across training stages; (b) performance with and without merging after 10K Stage 2 updates; and (c) the effect of [EOS] supervision within Stages 1 and 2. "new" denotes a capability absent from the backbone.

(a) Benchmark performance across the three training stages.

Checkpoint	GSM8K↑	MME-P↑	GenEval↑	Video-MME↑	ASR WER↓	TTS WER↓
Backbone	56.3	1410.7	0.63	–	–	–
Stage 1	4.2 (-52.1)	413.8 (-996.9)	0.12 (-0.51)	10.0 (new)	12.50 (new)	31.50 (new)
Stage 2	64.1 (+59.9)	1576.5 (+1162.7)	0.63 (+0.51)	52.0 (+42.0)	5.00 (-7.50)	4.17 (-27.33)
Stage 3	**83.7 (+19.6)**	**1672.9 (+96.4)**	**0.87 (+0.24)**	**63.4 (+11.4)**	**2.10 (-2.90)**	**2.10 (-2.07)**

(b) Benchmark performance after 10K Stage 2 updates with and without modality-disentangled merging.

	Original Modality			New Modality	
Initialization	GSM8K↑	POPE↑	MME-P↑	ASR WER↓	TTS WER↓
No merge	18.0	50.0	820.0	10.50	20.00
Modality-disentangled merging	**38.0**	**76.0**	**1150.0**	**6.46**	**10.22**

(c) Speech and termination results after 10K updates with and without [EOS] supervision in each stage. *Non-prem.*: first [EOS] at $\geq 80\%$ of reference length and before the decoding cap; *Prem.*: earlier termination; *Repeat*: three consecutive repeats of the same four-token sequence; *Cap*: no [EOS] before the decoding limit.

Stage	[EOS]	ASR WER↓	TTS WER↓	Non-prem.↑	Prem.↓	Repeat↓	Cap↓
1	Enabled	131.80	55.00	**35%**	60%	**8%**	**5%**
1	Disabled	**100.65**	**35.07**	20%	**3%**	60%	77%
2	Disabled	14.00	24.00	30%	**2%**	55%	68%
2	Enabled	**6.46**	**10.22**	**94%**	3%	**4%**	**3%**

The orchestration baselines employ strong external image and speech specialists, while existing unified models (Yang et al., 2025b; Zhan et al., 2024; Team, 2026) generate different modalities through staged or separate inference calls. In contrast, Dynin-Omni jointly generates interleaved text, image, and speech within a single masked-diffusion process without external generators. Despite this fully native design, Dynin-Omni surpasses the strongest orchestration baseline in overall, with consistent performance after filtering against the omnimodal data used for training (App. I). These results establish native joint decoding as an effective alternative to orchestrated omnimodal generation.

5 Analysis

5.1 Training Recipe

Stage-wise capability evolution. Tab. 4(a) tracks capability evolution across training. Stage 1 focuses adaptation on the new video and speech modalities, enabling efficient capability acquisition but causing forgetting of inherited text–image performance. Stage 2 merges the adapted model with the original backbone, recovering these inherited capabilities while further improving video and speech. Notably, even before the final scaling stage, Stage 2 already surpasses the backbone on GSM8K and MME-P and restores GenEval to its original level, showing that recovery does not come at the expense of the newly added modalities. Stage 3 further improves all domains, with GSM8K, MME-P, and GenEval surpassing the original backbone despite adding video and speech.

Effect of modality-disentangled merging. Tab. 4(b) isolates merging under matched data, optimizer, and update budgets. As shown in Fig. 3, modality-disentangled merging lowers the initial Stage 2 loss and accelerates convergence across original and new-modality tasks. After equal training, it also improves text, image, and speech results over training without merging (Tab. 4(b)). The ASR/TTS gains show that old-task recovery does not come at the expense of new-modality performance in this comparison. The benefit therefore persists beyond initialization, helping Stage 2 consolidate both inherited and newly acquired capabilities.

Comparison of merging strategies. We compare modality-disentangled merging with two alternatives at the Stage 1-to-Stage 2 transition. All three interpolate the shared Transformer parameters. *Shared Merging* interpolates existing-token rows and takes new-token rows from Stage 1, whereas *Stage 1-Only* takes all vocabulary-dependent rows from Stage 1. Complete formulations are provided in App. F. Fig. 5 sweeps α before Stage 2 training, comparing inherited-capability retention (GSM8K, POPE, and MME-P relative to the backbone) with new-modality retention (ASR and TTS relative

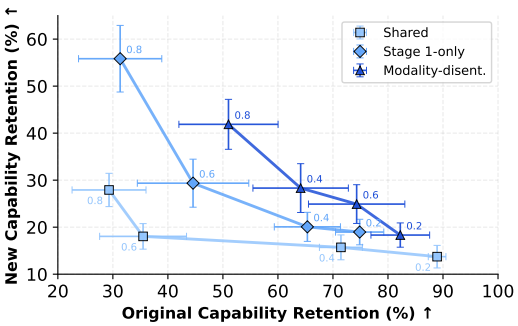


Figure 5: Capability retention across merging strategies and interpolation ratios.

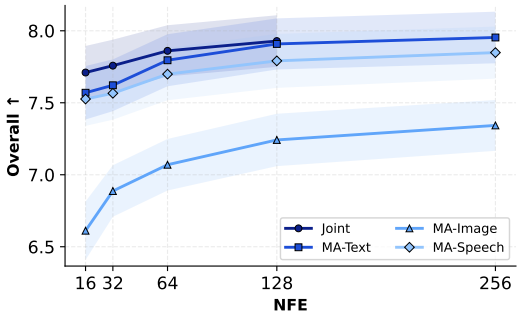


Figure 6: Joint and modality-anchored omni-modal generation across NFE budgets.

to Stage 1). Shared Merging retains neither group well, while Stage 1-Only favors new modalities at the expense of inherited text–image capabilities. Modality-disentangled merging provides the best balance at $\alpha = 0.4$ (backbone-to-Stage 1 ratio of 6:4), showing the value of selecting token rows by their origin. Across the sweep, the alternative token-row rules cannot match the same inherited/new-modality balance through α tuning alone.

Effect of scheduled padding supervision. Tab. 4(c) isolates `[EOS]` supervision under matched data, optimizer, and update budgets; the comparisons also use the same initialization within each stage. Our schedule disables `[EOS]` supervision in Stage 1 and enables it in Stage 2; disabling supervision does not prevent `[EOS]` generation at inference.

In Stage 1, disabling `[EOS]` supervision reduces WER and premature termination, but increases repetition and decoding-cap hits. Enabling it in Stage 2 reduces WER, repetition, and cap hits while increasing non-premature termination. Neither fixed supervision choice resolves both failure modes: early termination targets favor short outputs, whereas continued omission leaves length control undertrained. These results support aligning new-modality content before introducing explicit length control. Expanded merging controls and WER decompositions are provided in App. I.

5.2 Joint Omnimodal Generation

Fig. 6 compares joint decoding with Modality Anchoring (MA) on Omni-StoryBench using the same checkpoint. Joint decoding refines text, image, and speech in one diffusion pass; MA generates an anchor modality first and conditions the others on it in a second pass. We measure cost by total function evaluations (NFEs) across passes, so two 128-NFE MA passes require 256 NFEs. Both strategies use the same checkpoint and evaluation examples, so this comparison changes the decoding schedule rather than model capacity or external generators.

At every matched budget, joint decoding scores higher than all MA variants on the five-metric Overall measure (Fig. 6). Its margin over the strongest anchor is largest at 16–32 NFEs, when two-stage decoding divides a small budget between passes. At 128 NFEs, joint decoding nearly matches the best two-stage result at 256 NFEs. Joint decoding also leads the best anchor in Omni consistency at these low budgets, indicating that the gain extends beyond the aggregate quality score. These results show that a single joint pass uses a limited sampling budget more effectively than staging generation around a fixed modality anchor. Full modality-wise and cross-modal results are in App. I.

6 Conclusion

We introduced Dynin-Omni, an open-source 8B-scale masked-diffusion model supporting text and image understanding and generation, video understanding, and speech understanding and generation within one backbone. Its three-stage recipe combines modality-disentangled merging and scheduled padding supervision to retain inherited capabilities and learn flexible-length outputs. Dynin-Omni pairs strong individual-task results with native joint text–image–speech generation, outperforming prior unified systems on several shared benchmarks and orchestration systems on joint generation.

AI USE STATEMENT

We used generative AI models for synthetic data augmentation and filtering, as detailed in App. E. We also used AI tools to review and edit the manuscript, including feedback on experimental design and the interpretation of results. The authors reviewed all AI-assisted material and take responsibility for the final paper.

ETHICS STATEMENT

Our model may support applications such as accessible multimodal interaction and creative content production, but its ability to generate images and speech also raises risks of impersonation, misinformation, and other misuse of synthetic media. We document the sources and filtering of training data in App. E and discuss broader impacts in App. A.2. Downstream use should include appropriate safeguards, provenance disclosure for generated media, and consent when a person's likeness or voice is involved.

REPRODUCIBILITY STATEMENT

The model and training procedure are described in Sections 2 and 3. The appendix provides the data sources and construction procedures (App. E), training configurations (App. F), inference and evaluation settings (App. G), and extended results (App. I). We plan to release model weights, code, evaluation scripts, and synthetic-data construction materials as described in App. J.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

AIDAS-Lab. What comes next? Omni-StoryBench for evaluating story-grounded omni-modal generation, 2026. URL `https://huggingface.co/datasets/snu-aidas/Omni-StoryBench`. Dataset: snu-aidas/Omni-StoryBench.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pp. 4218–4222, 2020.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6836–6846, 2021.

Marianne Arriola, Aaron Gokaslan, Justin T Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. *arXiv preprint arXiv:2503.09573*, 2025.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured Denoising Diffusion Models in Discrete State-Spaces. In *Proc. of Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

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. *arXiv preprint arXiv:2502.13923*, 2025.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1728–1738, 2021.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, Ido Shahaf, Oren Tropp, Ehud Karpas, Ran Zilberstein, Jiaqi Zeng, Soumye Singhal, Alexander Bukharin, Yian Zhang, Tugrul Konuk, Gerald Shen, Ameya Sunil Mahabaleshwarkar, Bilal Kartal, Yoshi Suhara, Olivier Delalleau, Zijia Chen, Zhilin Wang, David Mosallanezhad, Adi Renduchintala, Haifeng Qian, Dima Rekesh, Fei Jia, Somshubra Majumdar, Vahid Noroozi, Wasi Uddin Ahmad, Sean Narenthiran, Aleksander Ficek, Mehrzad Samadi, Jocelyn Huang, Siddhartha Jain, Igor Gitman, Ivan Moshkov, Wei Du, Shubham Toshniwal, George Armstrong, Branislav Kisacanin, Matvei Novikov, Daria Gitman, Evelina Bakhturina, Jane Polak Scowcroft, John Kamalu, Dan Su, Kezhi Kong, Markus Kliegl, Rabeeh Karimi, Ying Lin, Sanjeev Satheesh, Jupinder Parmar, Pritam Gundecha, Brandon Norick, Joseph Jennings, Shrimai Prabhumoye, Syeda Nahida Akter, Mostofa Patwary, Abhinav Khattar, Deepak Narayanan, Roger Waleffe, Jimmy Zhang, Bor-Yiing Su, Guyue Huang, Terry Kong, Parth Chadha, Sahil Jain, Christine Harvey, Elad Segal, Jining Huang, Sergey Kashirsky, Robert McQueen, Izzy Putterman, George Lam, Arun Venkatesan, Sherry Wu, Vinh Nguyen, Manoj Kilaru, Andrew Wang, Anna Warno, Abhilash Somasamudramath, Sandip Bhaskar, Maka Dong, Nave Assaf, Shahar Mor, Omer Ullman Argov, Scot Junkin, Oleksandr Romanenko, Pedro Larroy, Monika Katariya, Marco Rovinelli, Viji Balas, Nicholas Edelman, Anahita Bhiwandiwalla, Muthu Subramaniam, Smita Ithape, Karthik Ramamoorthy, Yuting Wu, Suguna Varshini Velury, Omri Almog, Joyjit Daw, Denys Fridman, Erick Galinkin, Michael Evans, Katherine Luna, Leon Derczynski, Nikki Pope, Eileen Long, Seth Schneider, Guillermo Siman, Tomasz Grzegorzek, Pablo Ribalta, Monika Katariya, Joey Conway, Trisha Saar, Ann Guan, Krzysztof Pawelec, Shyamala Prayaga, Oleksii Kuchaiev, Boris Ginsburg, Oluwatobi Olabiyi, Kari Briski, Jonathan Cohen, Bryan Catanzaro, Jonah Alben, Yonatan Geifman, Eric Chung, and Chris Alexiuk. Llama-nemotron: Efficient reasoning models, 2025. URL https://arxiv.org/abs/2505.00949.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proc. of Int'l Conf. on Machine Learning (ICML)*, volume 2, pp. 4, 2021.

Louis Bethune, Victor Turrisi, Bruno Kacper Mlodozeniec, Pau Rodriguez Lopez, Lokesh Boominathan, Nikhil Bhendawade, Amitis Shidani, Joris Pelemans, Theo X. Olausson, Devon Hjelm, Paul Dixon, Joao Monteiro, Pierre Ablin, Vishnu Banna, Arno Blaas, Nick Henderson, Kari Noriy, Dan Busbridge, Josh Susskind, Marco Cuturi, Irina Belousova, Luca Zappella, Russ Webb, and Jason Ramapuram. The design space of tri-modal masked diffusion models, 2026. URL https://arxiv.org/abs/2602.21472.

Tiwei Bie, Maosong Cao, Kun Chen, Lun Du, Mingliang Gong, Zhuochen Gong, Yanmei Gu, Jiaqi Hu, Zenan Huang, Zhenzhong Lan, et al. Llada2.0: Scaling up diffusion language models to 100b. *arXiv preprint arXiv:2512.15745*, 2025.

Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11315–11325, June 2022.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*, 2021.

Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng, Kuo Yang, et al. Emova: Empowering language models to see, hear and speak with vivid emotions. *arXiv preprint arXiv:2409.18042*, 2024a.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024b.

Qian Chen, Yafeng Chen, Yanni Chen, Mengzhe Chen, Yingda Chen, Chong Deng, Zhihao Du, Ruize Gao, Changfeng Gao, Zhifu Gao, et al. Minmo: A multimodal large language model for seamless voice interaction. *arXiv preprint arXiv:2501.06282*, 2025a.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.

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, 2024c.

Shuang Cheng, Yuhua Jiang, Zineng Zhou, Dawei Liu, Wang Tao, Linfeng Zhang, Biqing Qi, and Bowen Zhou. SDAR-VL: Stable and efficient block-wise diffusion for vision-language understanding, 2025. URL https://arxiv.org/abs/2512.14068.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3.5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

Nilaksh Das, Saket Dingliwal, Srikanth Ronanki, Rohit Paturi, Zhaocheng Huang, Prashant Mathur, Jie Yuan, Dhanush Bekal, Xing Niu, Sai Muralidhar Jayanthi, et al. Speechverse: A large-scale generalizable audio language model. *arXiv preprint arXiv:2405.08295*, 2024.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

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.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.

Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL https://github.com/huggingface/open-r1.

Rongyao Fang, Aldrich Yu, Chengqi Duan, Linjiang Huang, Shuai Bai, Yuxuan Cai, Kun Wang, Si Liu, Xihui Liu, and Hongsheng Li. Flux-reason-6m & prism-bench: A million-scale text-to-image reasoning dataset and comprehensive benchmark. *arXiv preprint arXiv:2509.09680*, 2025.

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. *arXiv preprint arXiv:2306.13394*, 2023.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36: 27092–27112, 2023.

Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025.

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. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models. In *Proc. of Int'l Conf. on Learning Representations (ICLR)* , 2023.

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

Zhengfu He, Tianxiang Sun, Qiong Tang, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. DiffusionBERT: Improving Generative Masked Language Models with Diffusion Models. In *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021b.

hexgard. Kokoro-82m. https://github.com/hexgrad/kokoro, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, et al. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 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.

Wenyong Huang, Zhenhe Zhang, Yu Ting Yeung, Xin Jiang, and Qun Liu. SPIRAL: Self-supervised perturbation-invariant representation learning for speech pre-training. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=TBpg4PnXhYH`.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Huggingface. Mixture of thought dataset. https://huggingface.co/datasets/open-r1/Mixture-of-Thoughts, 2025.

Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990*, 2024.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max W.F. Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research*, 2024, 2024. URL https://openreview.net/forum?id=skLtdUVaJa.

Samar Khanna, Siddhant Kharbanda, Shufan Li, Harshit Varma, Eric Wang, Sawyer Birnbaum, Ziyang Luo, Yanis Miraoui, Akash Palrecha, Stefano Ermon, et al. Mercury: Ultra-fast language models based on diffusion. *arXiv e-prints*, pp. arXiv–2506, 2025.

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5530–5540. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/kim21f.html.

Jaeyeon Kim, Kulin Shah, Vasilis Kontonis, Sham Kakade, and Sitan Chen. Train for the worst, plan for the best: Understanding token ordering in masked diffusions. *Proc. of Int'l Conf. on Machine Learning (ICML)*, 2025.

Woojin Kim and Jaeyoung Do. Don't let it fade: Preserving edits in diffusion language models via token timestep allocation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=LOVuux2kvs.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Ariel N Lee, Cole J Hunter, and Nataniel Ruiz. Platypus: Quick, cheap, and powerful refinement of llms. *arXiv preprint arXiv:2308.07317*, 2023.

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. *arXiv preprint arXiv:2408.03326*, 2024a.

Junlin Li, Guodong Du, Jing Li, Sim Kuan Goh, Wenya Wang, Yequan Wang, Fangming Liu, Ho-Kin Tang, Saleh Alharbi, Daojing He, and Min Zhang. Multi-modality expansion and retention for llms through parameter merging and decoupling, 2025a. URL https://arxiv.org/abs/2505.17110.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024b.

Lijiang Li, Zuwei Long, Yunhang Shen, Heting Gao, Haoyu Cao, Xing Sun, Caifeng Shan, Ran He, and Chaoyou Fu. Omni-diffusion: Unified multimodal understanding and generation with masked discrete diffusion, 2026. URL https://arxiv.org/abs/2603.06577. Accepted at ICML 2026.

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. *arXiv preprint arXiv:2509.19244*, 2025b.

Shufan Li, Konstantinos Kallidromitis, Hritik Bansal, Akash Gokul, Yusuke Kato, Kazuki Kozuka, Jason Kuen, Zhe Lin, Kai-Wei Chang, and Aditya Grover. Lavida: A large diffusion language model for multimodal understanding, 2025c. URL https://arxiv.org/abs/2505.16839.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. Diffusion-LM Improves Controllable Text Generation. In *Proc. of Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025d.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

Zhenghao Lin, Yeyun Gong, Yelong Shen, Tong Wu, Zhihao Fan, Chen Lin, Nan Duan, and Weizhu Chen. Text generation with diffusion language models: a pre-training approach with continuous paragraph denoise. In *Proc. of Int'l Conf. on Machine Learning (ICML)*, 2023.

Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026.

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*, 2025a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024.

Zhiyuan Liu, Yicun Yang, Yaojie Zhang, Junjie Chen, Chang Zou, Qingyuan Wei, Shaobo Wang, and Linfeng Zhang. dllm-cache: Accelerating diffusion large language models with adaptive caching, 2025b. URL https://arxiv.org/abs/2506.06295.

Anton Lozhkov, Hynek Kydlíček, Loubna Ben Allal, Guilherme Penedo, Edward Beeching, Quentin Gallouédec, Nathan Habib, Lewis Tunstall, and Leandro von Werra. Openr1-math-220k. https://huggingface.co/datasets/open-r1/OpenR1-Math-220k, 2025.

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

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv:2405.20797*, 2024b.

Run Luo, Ting-En Lin, Haonan Zhang, Yuchuan Wu, Xiong Liu, Min Yang, Yongbin Li, Longze Chen, Jiaming Li, Lei Zhang, et al. Openomni: Advancing open-source omnimodal large language models with progressive multimodal alignment and real-time self-aware emotional speech synthesis. *arXiv preprint arXiv:2501.04561*, 2025a.

Run Luo, Xiaobo Xia, Lu Wang, Longze Chen, Renke Shan, Jing Luo, Min Yang, and Tat-Seng Chua. Next-omni: Towards any-to-any omnimodal foundation models with discrete flow matching, 2025b. URL https://arxiv.org/abs/2510.13721.

Dan Lyth and Simon King. Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *arXiv preprint arXiv:2402.01912*, 2024.

Xinyin Ma, Runpeng Yu, Gongfan Fang, and Xinchao Wang. dKV-cache: The cache for diffusion language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=Gppo2JImHs.

Yuxin Ma, Lun Du, Lanning Wei, Kun Chen, Qian Xu, Kangyu Wang, Guofeng Feng, Guoshan Lu, Lin Liu, Xiaojing Qi, Xinyuan Zhang, Zhen Tao, Haibo Feng, Ziyun Jiang, Ying Xu, Zenan Huang, Yihong Zhuang, Haokai Xu, Jiaqi Hu, Zhenzhong Lan, Junbo Zhao, Jianguo Li, and Da Zheng. dinfer: An efficient inference framework for diffusion language models, 2025b. URL https://arxiv.org/abs/2510.08666.

Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset. *arXiv preprint arXiv:2504.16891*, 2025.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

OpenAI. GPT4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

OpenAI. Gpt-image-1. https://openai.com/index/introducing-4o-image-generation/., 2025.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

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. URL https://arxiv.org/abs/2510.19808.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Subham Sekhar Sahoo, Marianne Arriola, Aaron Gokaslan, Edgar Mariano Marroquin, Alexander M Rush, Yair Schiff, Justin T Chiu, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2024.

Qingyu Shi, Jinbin Bai, Zhuoran Zhao, Wenhao Chai, Kaidong Yu, Jianzong Wu, Shuangyong Song, Yunhai Tong, Xiangtai Li, Xuelong Li, et al. Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model. *arXiv preprint arXiv:2505.23606*, 2025.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-asr technical report. *arXiv preprint*, 2026.

Yuxuan Song, Zheng Zhang, Cheng Luo, Pengyang Gao, Fan Xia, Hao Luo, Zheng Li, Yuehang Yang, Hongli Yu, Xingwei Qu, Yuwei Fu, Jing Su, Ge Zhang, Wenhao Huang, Mingxuan Wang, Lin Yan, Xiaoying Jia, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Yonghui Wu, and Hao Zhou. Seed diffusion: A large-scale diffusion language model with high-speed inference, 2025. URL https://arxiv.org/abs/2508.02193.

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. *Advances in neural information processing systems*, 36:49659–49678, 2023.

Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. Codi-2: In-context interleaved and interactive any-to-any generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27425–27434, June 2024.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024. *URL https://arxiv.org/abs/2405.09818*, 9(8), 2024.

NAVER Cloud HyperCLOVA X Team. Hyperclova x 8b omni. *arXiv preprint arXiv:2601.01792*, 2026.

Z-Image Team. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants, 2023. URL https://huggingface.co/datasets/teknium/OpenHermes-2.5.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri Iyer, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

TrillionLabs. Trida-7b: Block diffusion language model. https://huggingface.co/trillionlabs/Trida-7B, 2026.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

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. *arXiv preprint arXiv:2505.20147*, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

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*, 2024a.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint arXiv:2409.00750*, 2024b.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pp. 23965–23998. PMLR, 2022.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. Fast-dLLM: Training-free acceleration of diffusion LLM by enabling KV cache and parallel decoding. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=3Z3Is6hnOT.

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

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExT-GPT: Any-to-any multimodal LLM. In *Proceedings of the International Conference on Machine Learning*, pp. 53366–53397, 2024.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021.

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. *arXiv preprint arXiv:2408.12528*, 2024.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025.

Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, Jinbin Bai, Qian Yu, Dengyang Jiang, Yuandong Pu, Haoxing Chen, Le Zhuo, Junjun He, Gen Luo, Tianbin Li, Ming Hu, Jin Ye, Shenglong Ye, Bo Zhang, Chang Xu, Wenhai Wang, Hongsheng Li, Guangtao Zhai, Tianfan Xue, Bin Fu, Xiaohong Liu, Yu Qiao, and Yihao Liu. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding, 2025. URL https://arxiv.org/abs/2510.06308.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. MMaDA: Multimodal large diffusion language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=wczmXLuLGd.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv preprint arXiv:2510.15870*, 2025a.

Jiacheng Ye, Shansan Gong, Jiahui Gao, Junming Fan, Shuang Wu, Wei Bi, Haoli Bai, Lifeng Shang, and Lingpeng Kong. Dream-VL & Dream-VLA: Open vision-language and vision-language-action models with diffusion language model backbone, 2025b. URL https://arxiv.org/abs/2512.22615.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b, 2025c. URL https://hkunlp.github.io/blog/2025/dream.

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

Zhen Ye, Xinfa Zhu, Chi-Min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, Yizhu Jin, Zheqi Dai, et al. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *arXiv preprint arXiv:2502.04128*, 2025e.

Peiqi Yin, Jiangyun Zhu, Han Gao, Chenguang Zheng, Yongxiang Huang, Taichang Zhou, Ruirui Yang, Weizhi Liu, Weiqing Chen, Canlin Guo, et al. vllm-omni: Fully disaggregated serving for any-to-any multimodal models. *arXiv preprint arXiv:2602.02204*, 2026.

Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. Llada-v: Large language diffusion models with visual instruction tuning, 2025. URL https://arxiv.org/abs/2505.16933.

Zebin You, Xiaolu Zhang, Jun Zhou, Chongxuan Li, and Ji-Rong Wen. LLaDA-o: An effective and length-adaptive omni diffusion model, 2026. URL https://arxiv.org/abs/2603.01068.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

Runpeng Yu, Xinyin Ma, and Xinchao Wang. Dimple: Discrete diffusion multimodal large language model with parallel decoding, 2025. URL https://arxiv.org/abs/2505.16990.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 9127–9134, 2019.

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. *arXiv preprint arXiv:2311.16502*, 2023.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yu-Gang Jiang, and Xipeng Qiu. AnyGPT: Unified multimodal LLM with discrete sequence modeling. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9637–9662, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.521. URL https://aclanthology.org/2024.acl-long.521/.

Bowen Zhang, Congchao Guo, Geng Yang, Hang Yu, Haozhe Zhang, Heidi Lei, Jialong Mai, Junjie Yan, Kaiyue Yang, Mingqi Yang, et al. Minimax-speech: Intrinsic zero-shot text-to-speech with a learnable speaker encoder. *arXiv preprint arXiv:2505.07916*, 2025.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024a.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024b.

Haozhe Zhao, Xiaojian Shawn 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. *Advances in Neural Information Processing Systems*, 37:3058–3093, 2024.

Shanshan Zhao, Xinjie Zhang, Jintao Guo, Jiakui Hu, Lunhao Duan, Minghao Fu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, et al. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*, 2025.

Lin Zheng, Jianbo Yuan, Lei Yu, and Lingpeng Kong. A Reparameterized Discrete Diffusion Model for Text Generation. *arXiv preprint arXiv:2302.05737*, 2024.

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

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Appendix

CONTENTS

A Limitations and Discussions **23**
A.1 Limitations . . . 23
A.2 Broader Impacts . . . 23

B Discrete Diffusion Modeling **24**
B.1 Forward and Reverse Processes . . . 24
B.2 Training . . . 24
B.3 Inference . . . 25

C Related Work **25**
C.1 Diffusion Language Models . . . 25
C.2 Perception-centric and Unified Models . . . 26

D Architecture **27**
D.1 Backbone . . . 27
D.2 Omnimodal Tokenization . . . 27

E Data Composition **29**

F Training Recipes **31**
F.1 Dataset Curation . . . 31
F.2 Training Details . . . 31
F.3 Formulation of Merging Strategies . . . 33

G Inference Details **34**
G.1 Inference Configurations . . . 34
G.2 Efficiency Experiment Details . . . 34
G.3 Statistical Details . . . 35

H Serving Infrastructure **35**

I Additional Results **36**
I.1 Training Recipe Controls . . . 36
I.2 Text-Diffusion and Diffusion-MLLM Comparisons . . . 36
I.3 Statistical Uncertainty and Perceptual TTS Quality . . . 37
I.4 Omni-StoryBench Decontamination . . . 37

I.5 Inference Efficiency and Sampling Budget 37
I.6 Additional Quantitative and Qualitative Results 38

J License and Safeguards **39**

A Limitations and Discussions

A.1 Limitations

Scope of native generation. Dynin-Omni represents text, image, video, and speech within a shared token space, but our native joint generation experiments focus on interleaved text, image, and speech outputs. Video is currently evaluated primarily as an input modality for understanding rather than as dense video output generation. Extending the same formulation to temporally coherent video generation is an important direction for future work. We view this as a scope limitation rather than an architectural one: video tokens are already processed by the same backbone and objective, while native video generation mainly requires longer sequence budgets, stronger video tokenization/detokenization, and large-scale temporal generation data.

Generalist-specialist trade-off. Dynin-Omni is not intended to replace the strongest modality-specific experts on every individual benchmark. Models specialized for text reasoning, visual understanding, image synthesis, ASR, or TTS can allocate their full capacity, training data, and objectives to a single domain. In contrast, Dynin-Omni uses a single 8B-scale backbone and a unified masked-diffusion objective across modalities, without relying on external modality-specific generators. Therefore, remaining gaps to expert models on some tasks should be interpreted as the expected cost of native unification. The main result is that a single native backbone can achieve strong and balanced performance across modalities while supporting joint omnimodal generation.

Dependence on discrete tokenizers. Our formulation relies on pretrained discrete tokenizers and deterministic detokenizers for non-text modalities. These modules define the discrete representation space and map generated tokens back to pixels or waveforms, but they do not perform conditional generation. All generative decisions are made by the unified masked-diffusion backbone, which predicts tokens in the shared vocabulary under a single objective, without modality-specific generators, prediction heads, or reconstruction losses. Nevertheless, final output fidelity can still be bounded by the quality of the underlying tokenizers and detokenizers. For example, fine visual details, speech prosody, or long-range temporal consistency may benefit from stronger tokenization schemes. This limitation is orthogonal to the proposed backbone: improved tokenizers can be substituted while preserving the same unified masked-diffusion formulation.

Training-scale ablations. A complete ablation of all training factors is difficult at the scale of Dynin-Omni. Since the model is an 8B-scale omnimodal foundation model, retraining it across many combinations of model sizes, data mixtures, stage schedules, and merging ratios would require substantial compute beyond the scope of this work. This is a practical constraint shared by large-scale unified and omnimodal foundation-model studies, where resource constraints often limit training and validation to a fixed large-scale setting (Luo et al., 2025b). Accordingly, we provide targeted analyses of the components most relevant to our training challenges, rather than exhaustive training sweeps. For modality-disentangled merging, we compare alternative merging strategies and interpolation ratios to analyze capability preservation under vocabulary expansion. For scheduled padding supervision, we provide training diagnostics, including qualitative examples of early `[EOS]` shortcut behavior and stage-wise ASR/TTS WER changes after enabling `[EOS]` supervision. These analyses support the effectiveness of the proposed components, while broader scaling-law, data-mixture, and stage-schedule studies remain valuable directions for future work.

Backbone generality. Our training-recipe controls evaluate the extension from MMaDA to Dynin-Omni and therefore establish effectiveness for this pretrained backbone and modality-expansion setting. Applying the same recipe to other masked-diffusion or autoregressive backbones would require additional full training runs and remains an important direction for testing cross-backbone generality.

A.2 Broader Impacts

Dynin-Omni may benefit applications that require unified interaction across text, vision, and speech, such as accessible assistants, educational tools, creative storytelling, and multimodal content production. At the same time, stronger omnimodal generation can increase risks of synthetic media misuse,

impersonation, misinformation, biased outputs, and inappropriate generated content. We therefore encourage responsible deployment with provenance tracking, content filtering, human oversight in high-stakes settings, and clear disclosure when generated media is used.

B Discrete Diffusion Modeling

In this subsection, we describe the masked diffusion framework underlying Dynin-Omni, providing formal definitions of the forward and reverse processes, the training objective, and the inference procedure.

B.1 Forward and Reverse Processes

Given a discrete token sequence $\mathbf{x}_0 = (x_1, \ldots, x_L)$ drawn from the joint vocabulary $\mathcal{V}$, the forward process independently corrupts each token by replacing it with a special `[MASK]` token m. At masking ratio $t \in [0, 1]$, the marginal distribution over the corrupted sequence $\mathbf{x}_t$ factorizes across positions as:

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \prod_{i=1}^{L} q(x_{t,i} \mid x_{0,i}), \qquad q(x_{t,i} \mid x_{0,i}) = \begin{cases} 1 - t, & x_{t,i} = x_{0,i}, \\ t, & x_{t,i} = m, \end{cases} \tag{5}$$

so each token is independently masked with probability t and left unchanged with probability $1 - t$. At $t = 1$ all tokens are masked; at $t = 0$ the sequence is fully observed.

The reverse process learns to recover $\mathbf{x}_0$ from $\mathbf{x}_t$. For $0 \le s < t \le 1$, the true reverse posterior factorizes across positions and is given by:

$$q(x_{s,i} \mid x_{t,i}, x_{0,i}) = \begin{cases} \mathrm{Cat}(x_{s,i};\ x_{t,i}), & x_{t,i} \neq m, \\ \mathrm{Cat}\left(x_{s,i};\ \dfrac{(1 - \alpha_s)\, m + (\alpha_s - \alpha_t)\, x_{0,i}}{1 - \alpha_t}\right), & x_{t,i} = m, \end{cases} \tag{6}$$

where $\alpha_t = 1 - t$ under the linear schedule. For an unmasked token the posterior is deterministic ($x_{s,i} = x_{t,i}$); for a masked token it interpolates between remaining masked (weight $(1-\alpha_s)/(1-\alpha_t)$) and recovering the clean token (weight $(\alpha_s - \alpha_t)/(1 - \alpha_t)$). Since $x_{0,i}$ is unknown at inference, the parameterized reverse model replaces it with the model's prediction:

$$p_\theta(x_{s,i} \mid \mathbf{x}_t) = \begin{cases} \mathrm{Cat}(x_{s,i};\ x_{t,i}), & x_{t,i} \neq m, \\ \mathrm{Cat}\left(x_{s,i};\ \dfrac{(1 - \alpha_s)\, m + (\alpha_s - \alpha_t)\, p_\theta(x_{0,i} \mid \mathbf{x}_t)}{1 - \alpha_t}\right), & x_{t,i} = m, \end{cases} \tag{7}$$

where $p_\theta(x_{0,i} \mid \mathbf{x}_t)$ is the predicted distribution over $\mathcal{V}$ for position i conditioned on the full partially masked sequence $\mathbf{x}_t$ (Sahoo et al., 2024; Nie et al., 2025).

B.2 Training

The backbone Transformer p_θ is trained as a mask predictor: given a corrupted sequence $\mathbf{x}_t$ sampled from the forward process, it predicts the original token $x_{0,i}$ at every masked position simultaneously. Following prior work (Sahoo et al., 2024; Nie et al., 2025), the training objective is a cross-entropy loss computed only over masked positions, derived as a variational upper bound on the negative log-likelihood:

$$\mathcal{L}(\theta) = -\mathbb{E}_{t,\, \mathbf{x}_0,\, \mathbf{x}_t} \left[\frac{1}{t} \sum_{i=1}^{L} \mathbb{I}[x_{t,i} = m] \log p_\theta(x_{0,i} \mid \mathbf{x}_t) \right], \tag{8}$$

where $t \sim \mathcal{U}(0, 1)$, $\mathbf{x}_0$ is sampled from the training data, and $\mathbf{x}_t \sim q(\mathbf{x}_t \mid \mathbf{x}_0)$ from Eq. (5). The weighting factor $1/t$ arises from the continuous-time variational lower bound (Sahoo et al., 2024), and it has been shown that $\mathcal{L}(\theta)$ upper-bounds the negative log-likelihood of the model distribution:

$$-\mathbb{E}_{\mathbf{x}_0 \sim p_{\mathrm{data}}}[\log p_\theta(\mathbf{x}_0)] \le \mathcal{L}(\theta). \tag{9}$$

Unmasked tokens do not contribute to the loss, as their predictions are trivially correct due to the carry-over property of the reverse process (Eq. (7)). This objective is shared uniformly across all modalities in the joint vocabulary $\mathcal{V}$.

Table 5: Supported modality capabilities and architectural characteristics of representative perception-centric and unified perception–generation models. *Diff.* denotes diffusion-based generation and *FM* denotes flow-matching-based generation. Shaded cells indicate omnimodal capability (supporting text, vision, and speech concurrently) within each task group (Understanding or Generation). Entries written in parentheses, such as (+ Diff.), indicate auxiliary external generation modules attached to the primary architecture. For Omni-Diffusion, the image-editing entry denotes reported image inpainting rather than instruction-driven editing.

		Understanding Tasks				Generation Tasks		
Methods	Modeling	Textual Reasoning	Image	Video	Speech	Image	Image Editing	Speech
Perception-centric								
EMOVA-7B (Chen et al., 2024a)	AR		✓		✓			✓
VITA-1.5 (Fu et al., 2025)	AR		✓	✓	✓			✓
MiniCPM-o-2.6 (Yao et al., 2024)	AR		✓	✓	✓			✓
Baichuan-Omni-1.5 (Li et al., 2025d)	AR	✓	✓	✓	✓			✓
OpenOmni (Luo et al., 2025a)	AR		✓	✓	✓			✓
OmniVinci (Ye et al., 2025a)	AR		✓	✓	✓			✓
Qwen2.5-Omni-7B (Xu et al., 2025a)	AR	✓	✓	✓	✓			✓
Compositional Unified								
NextGPT (Wu et al., 2024)	AR (+ Diff.)		✓	✓	✓	✓		✓
CoDi-2 (Tang et al., 2024)	AR (+ Diff.)		✓	✓	✓	✓	✓	
DreamLLM (Dong et al., 2023)	AR (+ Diff.)		✓			✓		
SEED-X (Ge et al., 2024)	AR (+ Diff.)		✓			✓	✓	
UniWorld-V1 (Lin et al., 2025)	AR (+ Diff.)		✓			✓	✓	
HyperCLOVAX-8B-Omni (Team, 2026)	AR (+ Diff.)	✓	✓	✓	✓	✓	✓	✓
Native Unified								
AnyGPT (Zhan et al., 2024)	AR		✓		✓	✓		✓
Chameleon-7B (Team, 2024)	AR	✓	✓			✓		
Emu3 (Wang et al., 2024a)	AR		✓	✓		✓		
Janus-Pro (Chen et al., 2025b)	AR		✓			✓		
BAGEL (Deng et al., 2025)	AR + FM		✓			✓	✓	
Show-o2 (Xie et al., 2025)	AR + FM	✓	✓	✓		✓		
MMaDA (Yang et al., 2025b)	Discrete Diff.	✓	✓			✓		
Omni-Diffusion (Li et al., 2026)	Discrete Diff.		✓		✓	✓	✓	✓
Fudoki (Wang et al., 2025a)	Discrete FM		✓			✓		
NExT-OMNI (Luo et al., 2025b)	Discrete FM		✓	✓	✓	✓	✓	✓
LaViDa-O (Li et al., 2025b)	Discrete Diff.		✓			✓	✓	
Lumina-DiMOO (Xin et al., 2025)	Discrete Diff.		✓			✓	✓	
Dynin-Omni (Ours)	Discrete Diff.	✓	✓	✓	✓	✓	✓	✓

B.3 Inference

At inference time, generation begins from a fully masked target sequence $\mathbf{x}_1 = (m, \ldots, m)$ of a pre-specified length, and the reverse process is discretized into T steps from $t = 1$ to $t = 0$. At each step from time t to $s = t - 1/T$, the backbone predicts a categorical distribution $p_\theta(x_{0,i} \mid \mathbf{x}_t)$ over $\mathcal{V}$ for all masked positions in parallel, and a subset of tokens are committed while the rest are remasked for the next step. In principle, a fraction s/t of predicted tokens should be remasked at random to align the reverse process with the forward process (Nie et al., 2025); in practice, *low-confidence remasking* (Chang et al., 2022; Nie et al., 2025) is adopted, where the $\lfloor L(1-s) \rfloor$ currently masked tokens with the lowest predicted probability are remasked while the remaining high-confidence predictions are committed and held fixed. To further respect sequential structure, *block-wise decoding* (Arriola et al., 2025) divides the target sequence into contiguous blocks that are decoded left to right, with masked tokens within each block refined in parallel over multiple diffusion steps conditioned on all previously committed blocks.

C Related Work

C.1 Diffusion Language Models

Diffusion language models (DLMs) have recently emerged as an alternative to autoregressive (AR) language models by replacing left-to-right decoding with iterative denoising over token sequences (Austin et al., 2021; Li et al., 2022; Gong et al., 2023; Lin et al., 2023). By relaxing strict causal ordering, DLMs enable any-order generation, flexible intermediate editing, and improved global coherence (Kim et al., 2025; Khanna et al., 2025; Song et al., 2025), while also supporting controllable generation and parallel token updates for enhanced efficiency (Li et al., 2022; Kim &

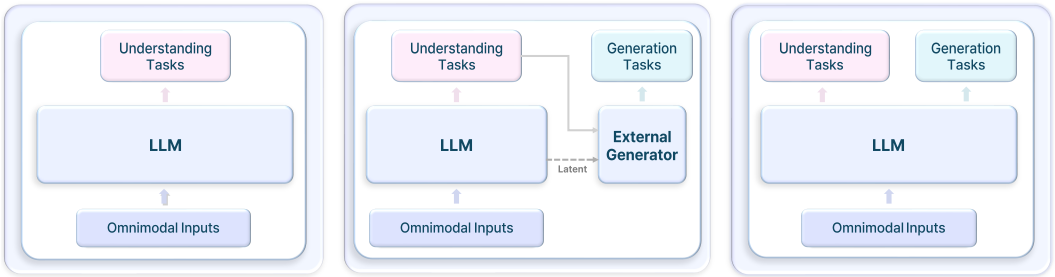


Figure 7: Architectural comparison of omnimodal modeling paradigms. (a) Perception-centric models focus on multimodal understanding with limited generation. (b) Compositional unified models rely on external modality-specific generators. (c) Dynin-Omni adopts a native unified formulation with shared token-level modeling.

Do, 2025; Wu et al., 2026; Liu et al., 2025b; Ma et al., 2025a). Among various formulations, discrete diffusion models instantiate DLMs directly over categorical token spaces, avoiding continuous relaxations while remaining compatible with standard language modeling objectives (He et al., 2023; Zheng et al., 2024; Sahoo et al., 2024). In particular, masked (absorbing-state) diffusion models formulate generation as iterative mask prediction, where subsets of tokens are replaced with `[MASK]` and reconstructed conditioned on the remaining context (He et al., 2023; Sahoo et al., 2024; Nie et al., 2025). This discrete denoising paradigm naturally supports bidirectional context modeling and parallel token refinement, making it well-suited for flexible and globally consistent generation. At scale, LLaDA (Nie et al., 2025) and Dream (Ye et al., 2025c) demonstrate that masked diffusion can be trained into a strong general-purpose language model competitive with AR baselines. Building on masked diffusion foundations, recent works extend diffusion-based modeling to multimodal settings. For example, LLaDA-V (You et al., 2025), Dimple (Yu et al., 2025), and LaViDa (Li et al., 2025c) explore discrete diffusion for vision-language modeling. MMaDA (Yang et al., 2025b) and Lumina-DiMOO (Xin et al., 2025) propose multimodal large diffusion language models for unified understanding and generation. Tri-modal masked-diffusion work includes text–image–audio modeling studies (Bethune et al., 2026) and Omni-Diffusion (Li et al., 2026), which supports text, image, and speech tasks under one masked-token objective and paired text–speech responses. Our focus is capability-preserving expansion to video understanding and broad reasoning, vision, and speech tasks, together with native interleaved text–image–speech output.

C.2 Perception-centric and Unified Models

We review three lines of related work: (1) perception-centric models, (2) compositional unified models, and (3) native unified models. Detailed model-wise comparisons are provided in Tab. 5.

Perception-centric models. Beyond traditional vision–language models (Chen et al., 2024c; Lu et al., 2024b; Li et al., 2024a) designed for visual perception tasks such as visual question answering (VQA) (Yue et al., 2023; Li et al., 2023), recent systems—including EMOVA (Chen et al., 2024a), VITA-1.5 (Fu et al., 2025), MiniCPM-o (Yao et al., 2024), Baichuan-Omni (Li et al., 2025d), OpenOmni (Luo et al., 2025a), OmniVinci (Ye et al., 2025a), and Qwen-Omni (Xu et al., 2025a;b)—extend multimodal perception beyond vision and language to additional modalities such as video, speech, and audio (Fig. 7(a)). These models support omnimodal inputs while also generating speech or audio responses. However, their architectures remain largely optimized for perception and understanding. In most cases, speech generation is implemented through auxiliary decoders or modality-specific pipelines (*e.g.* codec or TTS-style modules) rather than through a unified generative modeling framework. Consequently, these systems primarily function as omnimodal understanding models rather than fully unified omnimodal models.

Compositional unified models. Compositional unified models integrate multimodal perception and generation by coordinating a shared backbone with modality-specific generators (Fig. 7(b)). Rather than modeling all modalities within a single generative process, these systems delegate generation to auxiliary modules specialized for each modality. Early compositional unified models such as DreamLLM (Dong et al., 2023), SEED-X (Ge et al., 2024), and UniWorld-V1 (Lin et al., 2025) combine a multimodal backbone with external modality-specific generators, typically focusing on visual generation tasks such as image synthesis or editing. Another line of work extends this compositional design toward broader omnimodal capabilities. For example, NeXT-GPT (Wu et al., 2024) and CoDi-2 (Tang et al., 2024) support both visual and speech generation by coupling a multimodal language model with multiple diffusion-based generators. More recently, HyperCLOVAX-8B-Omni (Team, 2026) further advances this paradigm by enabling text, vision, and speech as both inputs and outputs, forming an omnimodal system. However, its generation still relies on heavy external diffusion generators such as FLUX (Labs, 2024) rather than being natively modeled within the backbone, which increases system complexity and can introduce orchestration failures during inference.

Native unified models. Native unified models aim to directly model multimodal perception and generation within a single backbone without relying on external generators (Fig. 7(c)). Early native unified models primarily focus on image–text modeling using autoregressive formulations. For example, AnyGPT (Zhan et al., 2024), Chameleon (Team, 2024), Emu3 (Wang et al., 2024a), and Janus-Pro (Chen et al., 2025b) extend language modeling to multimodal tokens, allowing the backbone to autoregressively generate text and visual representations within a shared token space. Subsequent works explore non-autoregressive generative paradigms to improve efficiency and scalability. Models such as BAGEL (Deng et al., 2025) and Show-o2 (Xie et al., 2025) combine autoregressive reasoning with diffusion or flow-matching-based generation, while MMaDA (Yang et al., 2025b), Fudoki (Wang et al., 2025a), LaViDa-O (Li et al., 2025b), and Lumina-DiMOO (Xin et al., 2025) adopt fully discrete diffusion or flow-based formulations to directly model multimodal token generation. Omni-Diffusion (Li et al., 2026) also models text, image, and speech with a unified masked-token objective. While recent work such as NExT-OMNI (Luo et al., 2025b) moves toward unified omnimodal perception and generation across multiple modalities, its training objective combines flow-matching generation with additional modality-specific reconstruction losses, introducing heterogeneous objectives that partially weaken the simplicity of fully unified modeling. Like Omni-Diffusion, Dynin-Omni models modality-specific discrete codes with a single masked-token objective and backbone. This unifies sequence-level modeling and generation across modalities while leaving their low-level tokenization spaces modality-specific. Our contribution additionally examines retention of inherited reasoning and visual capabilities during expansion to video and speech, and native interleaved text–image–speech output.

D Architecture

D.1 Backbone

Dynin-Omni builds upon the masked-diffusion paradigm introduced in LLaDA (Nie et al., 2025) and adopts MMaDA (Yang et al., 2025b) as its architectural foundation, extending it toward full omnimodal capability. Beyond the original vision–language modeling scope of MMaDA, we incorporate video understanding, speech understanding and generation, and introduce image editing, all within a unified masked token prediction framework. To support this expansion, we extend the discrete vocabulary to incorporate speech tokens and adapt the embedding lookup table and language modeling heads accordingly, while retaining a unified Transformer backbone without introducing modality-specific subnetworks. This architectural continuity enables scalable training through model merging and stage-wise optimization (§3) without backbone fragmentation across modalities.

D.2 Omnimodal Tokenization

Dynin-Omni supports textual, visual, and speech inputs within a unified token-based formulation. Specifically, images, videos, and speech signals are mapped into a shared discrete token space using modality-specific pretrained tokenizers. This enables all modalities to be processed under a single masked diffusion backbone, sharing the same embedding lookup table and Transformer

Table 6: Input–output template families across training stages. Red tokens denote masked target responses that are supervised throughout the stages. Blue tokens indicate the global termination token <EOS>, which also serves as padding token in masked diffusion language models, and is supervised only from Stage 2 onward.

Template Family	S1	S2	S3	Input-Output Template
Video → Text	✓	✓	✓	<image> ...<image> (video frames) <|user|> {prompt} <|assistant|> {response} <EOS>
Speech → Text	✓	✓	✓	<|startofspeech|> {speech} <|endofspeech|> <|startoftext|> {text} <EOS>
Text → Speech	✓	✓	✓	<|startoftext|> {text} <|endoftext|> <|startofspeech|> {speech} <|endofspeech|> <EOS>
Text Chat		✓	✓	<|user|> {prompt} <|assistant|> {response} <EOS>
Image → Text		✓	✓	<image> <|user|> {prompt} <|assistant|>; {response} <EOS>
Text → Image		✓	✓	<|startoftext|> {instruction} <|endoftext|> {image}
Image → Image		✓	✓	<image> <|startoftext|> {instruction} <|endoftext|> {image}
Omni Gen.			✓	<image> <|startoftext|> {instruction} <|endoftext|> {image} <|startoftext|>{Text}<EOS> <|startofspeech|>{Speech}<|endofspeech|><EOS>
Thinking-mode			✓	<|user|> {prompt} \think | \no_think <|assistant|> {response} <EOS>

architecture. Throughout training, both tokenizers and detokenizers remain frozen to preserve stable representations. Importantly, the detokenizer is a lightweight reconstruction module that deterministically maps discrete tokens back to modality space, without performing iterative generation or introducing additional learnable generative stages. By representing all modalities in the same token space, omnimodal interactions are modeled directly at the token level under a unified objective. This streamlined design reduces the need for separate generation pipelines or complex modality-specific objectives often found in existing omnimodal models (Luo et al., 2025b; Team, 2026).

Text. Built upon the LLaDA (Nie et al., 2025) architecture, Dynin-Omni retains its original text tokenizer (vocabulary size 126,080) to preserve the backbone's tokenization scheme, while introducing a small set of task-oriented special tokens (*e.g.* <think> and modality delimiters) to support structured reasoning and multimodal interaction. Beyond modality-specific tokenization, the central design of Dynin-Omni lies in mapping all modality tokens into a shared discrete vocabulary and processing them through a single embedding matrix and Transformer backbone. This unified token space enables multimodal interactions to occur directly at the token level under a single masked diffusion objective, integrating cross-modal reasoning and generation without relying on late-stage fusion or modality-specific orchestration.

Image. For image inputs, we follow Show-o (Xie et al., 2024) and MMaDA (Yang et al., 2025b) and adopt a MAGVIT-v2 (Yu et al., 2023)-style vector-quantized image tokenizer pretrained jointly with Show-o (Xie et al., 2024). The tokenizer consists of a convolutional encoder, a discrete codebook, and a corresponding decoder, and is trained to reconstruct images from discrete latent representations. Given an input image $x^I \in \mathbb{R}^{H\times W\times C}$, the encoder downsamples the image by a fixed factor f (using strided convolutions) and maps it into a latent feature map of spatial resolution $\frac{H}{f} \times \frac{W}{f}$. Each latent vector is then quantized to its nearest entry in a shared visual codebook of size $|\mathcal{V}_{\text{vision}}| = 8192$, resulting in a grid of discrete visual tokens. This grid is flattened in raster-scan order to form a 1D token sequence of length $\frac{H}{f} \cdot \frac{W}{f}$, which is subsequently processed by the backbone language model together with tokens from other modalities. For image generation and editing, a pretrained reconstruction module from MAGVIT-v2 maps the predicted discrete visual tokens back to pixel space to produce the output image. Importantly, this module is frozen and serves solely as a token-to-image reconstruction component; all sequence-level generative modeling is performed within the masked diffusion backbone.

Video. To maintain a unified visual token space across images and videos and to avoid introducing modality-specific tokenization branches, we design the video tokenizer by reusing the pretrained image tokenizer. Rather than learning a separate video tokenizer, we treat a video as a temporally ordered sequence of images while preserving the same vision token space $\mathcal{V}_{\text{vision}}$. Following previous works (Bertasius et al., 2021; Arnab et al., 2021), we apply the pretrained image tokenizer independently to sampled video frames and aggregate the resulting tokens into a single sequence.

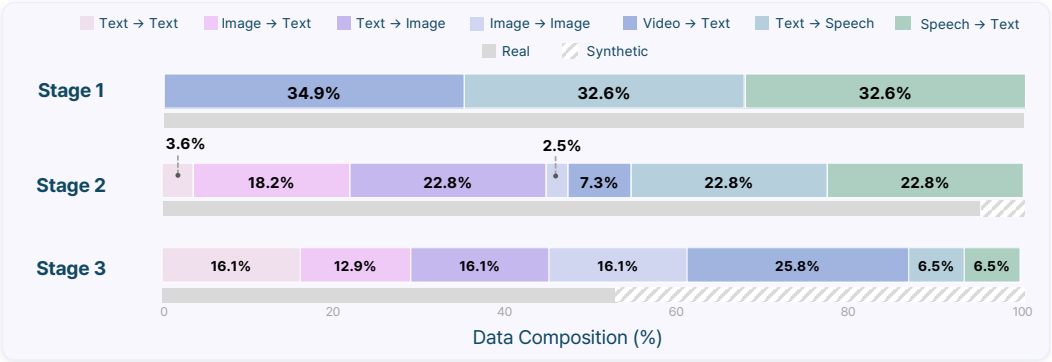


Figure 8: Distribution of training data across stages and tasks. Each stage is represented by a pair of horizontal bars: the upper bar shows the composition of training tasks (*e.g.* Text→Text, Image→Text), while the lower bar depicts the proportion of real and synthetic data used in that stage. Percentages indicate the relative contribution of each task within a stage.

Given a video, we uniformly sample T frames $\{x_t^I\}_{t=1}^T$, where each frame $x_t^I \in \mathbb{R}^{H\times W\times C}$. Each frame is tokenized independently using the same visual codebook as for images, producing a discrete token sequence of length $\frac{H}{f}\cdot\frac{W}{f}$. The per-frame token sequences are then concatenated in temporal order to form a video token sequence of length $T\cdot\frac{H}{f}\cdot\frac{W}{f}$. This design yields a unified discrete representation for both images and videos within a shared latent space. Empirically, we observe that improvements in image understanding consistently translate to enhanced video understanding performance, suggesting effective knowledge transfer enabled by the shared visual tokenization.

Speech. For speech, we adopt the EMOVA speech-to-unit (S2U) tokenizer (Chen et al., 2024a), which encodes an input waveform into discrete speech units. Concretely, EMOVA uses a SPIRAL-based speech encoder (Huang et al., 2022) to capture phonetic and tonal cues, followed by a finite scalar quantizer (FSQ) that discretizes the semantic content into a codebook of size $|\mathcal{V}_{\text{speech}}| = 4096$ at a rate of 25 tokens per second (*i.e.* one unit every $\sim$40 ms). We integrate speech into our token-based backbone by concatenating the speech codebook indices to the original vocabulary. Speech tokenizer further adopts a semantic–acoustic disentanglement design, where only the semantic embedding is quantized into speech units, while acoustic style is modeled separately. For detokenization, we use EMOVA's unit-to-speech (U2S) detokenizer built on VITS (Kim et al., 2021), which conditions a speech decoder on (i) unit embeddings derived from generated speech units and (ii) a style embedding extracted from a reference style prototype. In particular, the tokenizer constructs 24 style prototypes by combining two genders (male/female), four emotions (neutral/happy/sad/angry), and three pitches (normal/high/low), enabling controllable speech generation across these styles. Similar to the image modality, the speech modality also relies on a pretrained and frozen reconstruction module, with all sequence-level generative modeling performed within the masked diffusion backbone.

E Data Composition

Fig. 8 illustrates the data allocation across the three training stages, and Tab. 7 summarizes the dataset statistics. Our data composition follows a progressive scaling strategy: Stage 1 prioritizes modality-specific datasets to align newly introduced modalities with the backbone; Stage 2 substantially increases cross-modal instruction data to enable balanced joint omnimodal supervision; and Stage 3 emphasizes high-quality reasoning traces and curated multimodal samples for capability refinement. This staged allocation ensures stable modality integration in early training, broad omnimodal coverage in intermediate training, and targeted capability enhancement in later stages.

Training Stage 1 (Modality Alignment). Stage 1 aims to align newly introduced modalities with text as the semantic anchor. For video-to-text alignment, we leverage WebVid-10M (Bain et al., 2021), a large-scale video captioning corpus containing 10.7M video–text pairs covering approximately 52K hours of video. For speech alignment (used for both ASR and TTS), we combine multiple

large-scale speech datasets, including GigaSpeech (10K hours) (Chen et al., 2021), LibriSpeech (360 hours) (Panayotov et al., 2015), and CommonVoice (3K hours) (Ardila et al., 2020), resulting in approximately 14K hours of speech data. The same speech datasets are consistently used for both ASR and TTS to ensure stable semantic grounding across modalities.

Training Stage 2 (Omnimodal SFT). Stage 2 performs joint supervised fine-tuning across text, image, video, and speech to establish unified cross-modal capability. For text, we train on large-scale human–assistant dialogue data (*e.g.* Evol-Instruct (Xu et al., 2023) and Magpie-Pro (Xu et al., 2024)) to improve conversational ability and general knowledge, and additionally include mathematical reasoning datasets (*e.g.* Open-Platypus (Lee et al., 2023), OpenR1-Math (Face, 2025), OpenHermes-2.5 (Teknium, 2023)), where supervision is applied only to concise final answers without explicit chain-of-thought traces. For image understanding, we use Cambrian10M (Tong et al., 2024) to cover diverse and complex visual formats. Image generation and editing are trained using a mixture of high-quality datasets, including JourneyDB (Sun et al., 2023), FLUX-Reason-6M (Fang et al., 2025), PickaPic (Kirstain et al., 2023), UltraEdit (Zhao et al., 2024), HQEdit (Hui et al., 2024), and Pico-Banana-400K (Qian et al., 2025).

For Stage 2 video understanding, we combine LLaVA-Video-178K (Zhang et al., 2024b) and Open-Vid1M (Nan et al., 2024) with approximately 2M synthetic video–text pairs. We construct generation prompts from captions and instructions in the public video datasets, remove near-duplicate and low-information prompts, and generate the corresponding clips using Wan2.2 (Wan et al., 2025). The generated clips are filtered for decoding failures, insufficient motion, poor visual quality, and low agreement between the conditioning text and video content using InternVL-based consistency checks (Chen et al., 2024c). Near-duplicate clips and samples flagged by the benchmark-overlap checks are also removed. The retained clips are paired with their originating descriptions and converted into video captioning and instruction-following examples. Together with the public datasets, this produces the 3.2M video-to-text examples used in Stage 2. For speech, we reuse the ASR and TTS datasets employed in Stage 1, ensuring continuity between modality alignment and joint training.

Training Stage 3 (Capability Scaling). Stage 3 focuses on enhancing advanced reasoning, long-form generation quality, and multimodal robustness. For textual reasoning, we introduce explicit chain-of-thought (CoT) supervision with thinking-mode control tokens. We leverage publicly available reasoning datasets, including Llama-Nemotron Post-Training (Bercovich et al., 2025), OpenR1-Math-220K (Lozhkov et al., 2025), Mixture-of-Thoughts (Huggingface, 2025), and OpenMathReasoning (Moshkov et al., 2025).

For textual reasoning, we supplement the public reasoning datasets with traces generated by GLM-4.5-Air (Zeng et al., 2025), GPT-OSS-120B (Agarwal et al., 2025), Qwen3-Next-80B-A3B-Thinking (Yang et al., 2025a), and DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI, 2025). Starting from approximately 10M candidate traces, we verify final answers using task-specific checkers where available, remove traces whose reasoning is inconsistent with the final answer, and discard incomplete, repetitive, or truncated generations. We additionally balance the retained examples across subjects and difficulty levels, resulting in 200K synthetic reasoning examples.

For image understanding, we augment Cambrian10M (Tong et al., 2024) with synthetic captioning, question-answering, and multi-turn instruction-following examples. InternVL (Chen et al., 2024c) generates questions and answers grounded in the source images, after which a separate verification pass checks whether each answer is supported by visible image content. We remove examples with ungrounded answers, ambiguous questions, OCR corruption, or low image–text consistency. The resulting public and synthetic mixture contains 0.4M image-understanding examples.

For text-to-image generation, we construct prompts from the public image–text mixture and expand them with variations in composition, attributes, spatial relations, and rendering style. We generate candidate images using FLUX (Labs, 2024), FLUX2 (Labs, 2025), Z-Image (Team, 2025), and Qwen-Image (Wu et al., 2025a). The candidates are filtered using InternVL (Chen et al., 2024c) for prompt–image consistency and instruction satisfaction, together with rule-based checks for resolution, malformed outputs, duplicated images, embedded text, and low visual quality. After filtering and balancing prompt categories, we retain 0.5M text-to-image pairs.

For image editing, we construct source–instruction–target triplets using FLUX-Kontext (Labs et al., 2025) and Qwen-Image (Wu et al., 2025a). Candidate targets are retained only when InternVL

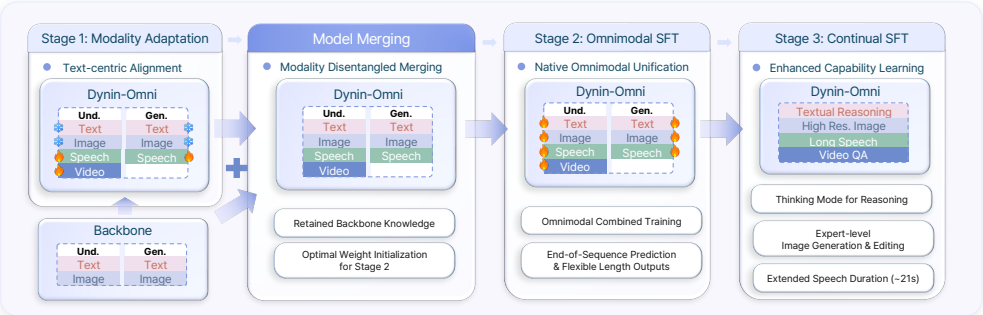


Figure 9: Three-stage training recipe for stable omnimodal extension of Dynin-Omni.

confirms that the requested modification is present while unrelated source content is preserved. We additionally remove pairs with excessive global changes, weak source–target correspondence, or ambiguous editing instructions. The final image-editing mixture contains 0.5M examples.

For video understanding, we combine LLaVA-Video-178K (Zhang et al., 2024b) and ShareGPT4Video (Chen et al., 2024b) with additional question–answer pairs constructed from the training videos. Qwen2.5-VL (Bai et al., 2025) generates questions requiring object, action, temporal-order, and multi-frame reasoning together with answers grounded in the corresponding clips. We discard questions that can be answered without observing the video, answers unsupported by the sampled frames, and examples for which independent teacher predictions disagree. The resulting video-understanding mixture contains 0.8M examples.

We construct a candidate pool of 300K long-form speech examples using Kokoro (hexgard, 2025). The input texts are sampled from public transcripts and the filtered text-instruction mixture and are synthesized with diverse voices and speaking styles, covering utterances up to 21 seconds. We remove clips containing excessive silence, clipping, decoding failures, or durations outside the training range. We further transcribe each generated clip with Qwen3-ASR (Shi et al., 2026) and retain only samples whose transcription agrees with the conditioning text below a 10% word-error threshold. After filtering, 0.2M speech examples are used for Stage 3 ASR and TTS training.

To train native joint generation, we construct 10K text–image-to-text–image–speech examples from public text–image instructional data and synthetic procedural examples. For the public portion, consecutive instructions and their accompanying images form the context and continuation. For the synthetic portion, GLM-4.5-Air (Zeng et al., 2025) generates procedural contexts, narration, scene descriptions, and spoken utterances. We create or edit candidate scene images using the image-generation and editing models described above—FLUX (Labs, 2024), FLUX2 (Labs, 2025), Z-Image (Team, 2025), Qwen-Image (Wu et al., 2025a), and FLUX-Kontext (Labs et al., 2025)—and synthesize speech with Kokoro (hexgard, 2025). We retain examples that pass contextual-consistency, image–text alignment, ASR-agreement, output-length, and duplicate-content checks. Each retained sequence supervises interleaved text, image, and speech generation.

Reproducibility. We will release the prompts, generator configurations, filtering code, and manifests used to construct the synthetic datasets above. For source media that cannot be redistributed, we will provide source identifiers, hashes, and reconstruction scripts.

F Training Recipes

F.1 Dataset Curation

We report the datasets used at each stage of Dynin-Omni, including their sources and sizes, in Tab. 7.

F.2 Training Details

Tab. 8 summarizes the stage-wise training configurations of Dynin-Omni, including optimization hyperparameters, compute budget, modality-specific resolutions, and speech duration. All stages

Table 7: Detailed data composition across training stages for Dynin-Omni. Tasks are described using unified input–output directions (e.g., V→T denotes video-to-text).

Stage	Task	Dataset Source	Scale (# pair)
Stage 1	V→T	WebVid-10M (Bain et al., 2021)	10.7M
	S→T	GigaSpeech (Chen et al., 2021); LibriSpeech (Panayotov et al., 2015); Common Voice (Ardila et al., 2020)	10M
	T→S	GigaSpeech (Chen et al., 2021); LibriSpeech (Panayotov et al., 2015); Common Voice (Ardila et al., 2020)	
Stage 2	T→T	Evol-Instruct (Xu et al., 2023); Magpie-Pro (Xu et al., 2024); Open-Platypus (Lee et al., 2023); OpenR1-Math (Lozhkov et al., 2025) ; OpenHermes-2.5 (Teknium, 2023)	1.6M
	I→T	Mantis-Instruct (Jiang et al., 2024); Cambrian10M (Tong et al., 2024)	8M
	T→I	JourneyDB (Sun et al., 2023); FLUX-Reason-6M (Fang et al., 2025); PickaPic (Kirstain et al., 2023)	10M
	I→I	UltraEdit (Zhao et al., 2024); HQEdit (Hui et al., 2024); Pico-Banana-400K (Qian et al., 2025)	1.1M
	V→T	LLaVA-Video-178K (Zhang et al., 2024b); OpenVid1M (Nan et al., 2024); Synthetic Inst. Data	3.2M
	S→T	GigaSpeech (Chen et al., 2021); LibriSpeech (Panayotov et al., 2015); Common Voice (Ardila et al., 2020)	10M
	T→S	GigaSpeech (Chen et al., 2021); LibriSpeech (Panayotov et al., 2015); Common Voice (Ardila et al., 2020)	
Stage 3	T→T	Llama-Nemotron Post-Training (Bercovich et al., 2025); OpenR1-Math-220K (Lozhkov et al., 2025); Mixture-of-Thoughts (Huggingface, 2025); OpenMathReasoning (Moshkov et al., 2025); Synthetic Data	0.5M
	I→T	Cambrian10M (Tong et al., 2024) ; Synthetic Data	0.4M
	T→I	FLUX-Reason-6M (Fang et al., 2025) ; Synthetic Data	0.5M
	I→I	Pico-Banana-400K (Qian et al., 2025) ; Synthetic Data	0.5M
	V→T	LLaVA-Video-178K (Zhang et al., 2024b); ShareGPT4Video (Chen et al., 2024b); Synthetic QA Data	0.8M
	S→T	LibriSpeech (Panayotov et al., 2015); Synthetic Data	0.2M
	T→S	LibriSpeech (Panayotov et al., 2015); Synthetic Data	
	TI→TIS	Public + Synthetic Omnimodal Data	0.01M

use AdamW with cosine learning-rate scheduling, weight decay 0.1, gradient norm clipping 1.0, and batch size 128. Stage 1 uses a learning rate of 1.0×10^{-5} on 32 H100 GPUs, while Stages 2 and 3 use 2.0×10^{-5} on 32 and 40 H100 GPUs, respectively. The complete adaptation cost after initialization from the released MMaDA checkpoint is 63,425 H100-hours: 5,200 for Stage 1, 34,560 for Stage 2, and 23,665 for Stage 3. This accounting covers all three adaptation stages and excludes the inherited MMaDA pretraining cost.

Vocabulary extension. Starting from the MMaDA backbone, Dynin-Omni extends the original text–vision vocabulary with a dedicated speech vocabulary. The unified vocabulary is organized as

$$\mathcal{V} = \mathcal{V}_{\text{text}} \cup \mathcal{V}_{\text{vision}} \cup \mathcal{V}_{\text{speech}},$$

where $|\mathcal{V}_{\text{text}}| = 126080$, $|\mathcal{V}_{\text{vision}}| = 8192$, and $|\mathcal{V}_{\text{speech}}| = 4096$. Images and videos share the same visual vocabulary, with videos represented by tokenizing sampled frames using the image tokenizer. Speech tokens are added as newly introduced vocabulary entries and processed by the same shared embedding table and masked-diffusion Transformer.

Stage 1: new-modality adaptation. Stage 1 adapts newly introduced modalities through text-centric alignment tasks, including video captioning, automatic speech recognition (ASR), and text-to-speech (TTS). As shown in Tab. 8, we use 8 video frames at 128×128 resolution and speech segments up to approximately 10 seconds. This low-resolution, higher-frame setting emphasizes temporal alignment during early adaptation. During this stage, `[EOS]` tokens are excluded from supervision, allowing the model to prioritize semantic alignment before learning termination behavior.

Stage 2: omnimodal supervised fine-tuning. Stage 2 jointly trains text, image, video, and speech tasks under the unified objective. Image understanding and image editing use 256×256 resolution, while image generation uses 336×336 resolution. For video, we use 5 frames at 224×224 resolution. For speech, utterances are limited to approximately 15 seconds. From Stage 2 onward, `[EOS]` supervision is enabled so that the model learns flexible-length termination after semantic alignment has been established.

Stage 3: capability scaling. Stage 3 further improves reasoning, high-resolution visual generation, image editing, video understanding, and long-form speech modeling. We increase image understanding resolution to 480×480, image generation resolution to 512×512, and image editing resolution

Table 8: Detailed hyperparameters and configurations of the training recipe across different stages.

Hyperparameters	**Stage 1** (New Modality Adapt.)	**Stage 2** (Omni. SFT)	**Stage 3** (Cont. SFT)
Learning Rate	1.0×10^{-5}	2.0×10^{-5}	2.0×10^{-5}
LR Scheduler	Cosine	Cosine	Cosine
Weight Decay	0.1	0.1	0.1
Gradient Norm Clip	1.0	1.0	1.0
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999$)		
Batch Size	128	128	128
Training GPUs	32×H100	32×H100	40×H100
GPU Hours	5,200	34,560	23,665
Image Und. Resolution	–	256	480
Image Gen. Resolution	–	336	512
Image Edit Resolution	–	256	336
Video Resolution / # Frame	128 / 8	224 / 5	224 / 5
Speech Duration (sec)	∼10	∼15	∼21

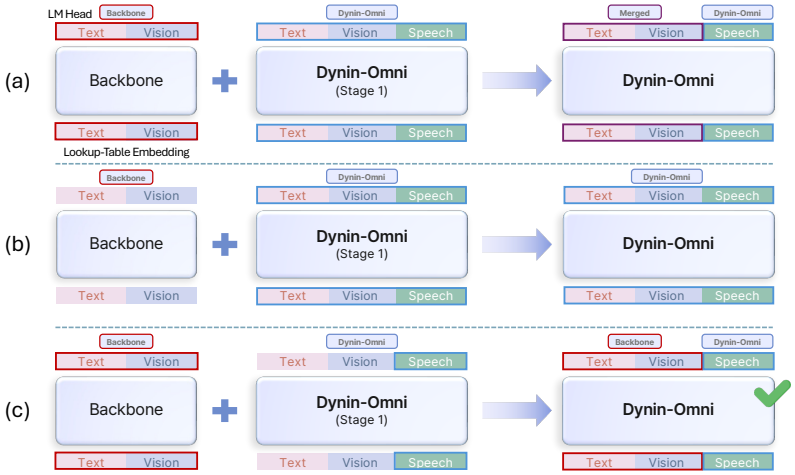


Figure 10: Comparison of model merging strategies under vocabulary extension. (a) Shared Merging: Original vocabulary parameters are interpolated between the backbone and Stage 1, while new speech embeddings come from Stage 1. (b) Stage 1-Only Merging: All vocabulary-related parameters are taken from Stage 1. (c) Modality-Disentangled Merging (Ours): Original vocabulary parameters remain from the backbone, while new speech-specific dimensions come from Stage 1. Remaining backbone weights are interpolated.

to 336×336. Video inputs remain at 5 frames with 224×224 resolution. For speech, Stage 3 extends the maximum duration to approximately 21 seconds.

For the joint omnimodal examples described in App. E, images are generated at 336×336 resolution and speech outputs are capped at 5 seconds.

F.3 Formulation of Merging Strategies

Fig. 10 illustrates the three vocabulary-merging strategies considered under vocabulary expansion. We provide formal definitions of these strategies, which are compared in §5.1. Let $\theta^{(0)}$ denote the original backbone and $\theta^{(1)}$ denote the Stage 1-adapted model. Stage 1 extends the vocabulary from $\mathcal{V}^{(0)}$ to $\mathcal{V} = \mathcal{V}^{(0)} \cup \mathcal{V}^{\text{new}}$, where $\mathcal{V}^{\text{new}}$ contains newly introduced modality tokens. We decompose the model parameters into $\theta = \{\theta_{\text{shared}}, \theta_{\text{vocab}}\}$, where θ_{shared} denotes vocabulary-independent parameters and θ_{vocab} denotes vocabulary-dependent parameters, such as token embeddings and output heads.

For parameters with matching dimensionality, we define the interpolation operator as

$$\mathcal{M}(\theta^{(0)}, \theta^{(1)}; \alpha) = (1 - \alpha)\theta^{(0)} + \alpha\theta^{(1)}. \tag{10}$$

Here, α denotes the interpolation weight assigned to the Stage 1-adapted model. All strategies apply this interpolation to vocabulary-independent parameters:

$$\theta^{\text{merge}}_{\text{shared}} = \mathcal{M}(\theta^{(0)}_{\text{shared}}, \theta^{(1)}_{\text{shared}}; \alpha) = (1-\alpha)\theta^{(0)}_{\text{shared}} + \alpha\theta^{(1)}_{\text{shared}}. \tag{11}$$

The strategies differ only in how they handle vocabulary-dependent parameters.

Shared Merging. For parameters corresponding to the original vocabulary, Shared Merging interpolates the original and Stage 1 parameters. Newly introduced token parameters are inherited from the Stage 1 model:

$$\theta^{\text{merge}}_{\text{vocab}}(i) = \begin{cases} \mathcal{M}\Big(\theta^{(0)}_{\text{vocab}}(i), \theta^{(1)}_{\text{vocab}}(i); \alpha\Big), & i \in \mathcal{V}^{(0)}, \\ \theta^{(1)}_{\text{vocab}}(i), & i \in \mathcal{V}^{\text{new}}. \end{cases} \tag{12}$$

This strategy follows standard parameter interpolation on the overlapping vocabulary, but may disrupt token-level representations learned by either model.

Stage 1-Only Merging. Stage 1-Only Merging inherits all vocabulary-dependent parameters from the Stage 1-adapted model:

$$\theta^{\text{merge}}_{\text{vocab}}(i) = \theta^{(1)}_{\text{vocab}}(i), \quad i \in \mathcal{V}. \tag{13}$$

This maximally preserves newly acquired modality-specific representations, but may retain catastrophic forgetting of original text–image token representations.

Modality-Disentangled Merging. Our modality-disentangled strategy preserves original vocabulary parameters from the backbone and inherits only newly introduced token parameters from the Stage 1 model:

$$\theta^{\text{merge}}_{\text{vocab}}(i) = \begin{cases} \theta^{(0)}_{\text{vocab}}(i), & i \in \mathcal{V}^{(0)}, \\ \theta^{(1)}_{\text{vocab}}(i), & i \in \mathcal{V}^{\text{new}}. \end{cases} \tag{14}$$

This separates backbone preservation from new-modality incorporation, allowing the merged model to retain existing text–image representations while adding newly learned modality tokens.

Discussion. The key difficulty in omnimodal modality expansion is that vocabulary extension introduces new rows in the embedding table and output head. As a result, naive parameter interpolation is ill-defined for the full vocabulary-dependent parameters. The above strategies instantiate different choices for resolving this mismatch. Shared merging interpolates overlapping rows, stage 1-only favors newly adapted representations, and modality-disentangled merging explicitly preserves original-token representations while importing new-token representations. As shown in §5.1, this disentangled treatment provides the most stable initialization for Stage 2 omnimodal supervised fine-tuning.

G Inference Details

G.1 Inference Configurations

We use block-wise decoding with block size 16 and 1024 steps for text reasoning, and block size 16 with 128 steps for image/video understanding and ASR. For TTS, we use block-wise decoding with block size 128 and 512 steps. For image synthesis, we use fully parallel decoding with 16 steps and classifier-free guidance (CFG) (Ho & Salimans, 2022) scale 3.5 for generation, and 128 steps with CFG scale 2.5 for editing. For all tasks, we use confidence-based remasking (Chang et al., 2022; Nie et al., 2025; Ye et al., 2025c) during iterative decoding.

G.2 Efficiency Experiment Details

Accelerated inference setup. We evaluate text-to-text and text-to-speech inference speed on a single NVIDIA A100 80GB SXM GPU. Following Fast-dLLM (Wu et al., 2026), we apply KV-cache approximation and confidence-based parallel decoding. The optimized inference path is implemented in a customized version of dInfer (Ma et al., 2025b). For both GSM8K and LibriSpeech-clean TTS, we use DualCache with block size 64 and threshold 0.7. For the matched MMaDA comparison, we port the same dInfer/DualCache path to MMaDA and hold batch size, maximum generation length, confidence threshold, block size, `torch.compile`, and CUDA Graph fixed.

Baseline configurations. For text-to-text evaluation, Qwen3-8B is evaluated in a 4-shot setting with exact autoregressive decoding using vLLM. MMaDA is evaluated in a 5-shot setting using its official repository, with generation length 256, fixed-step parallel decoding with 128 steps, and block length 64; KV caching is not supported in this setting. The standard released-checkpoint evaluation, the paper-reported final UniGRPO result, and the accelerated dInfer throughput run use distinct checkpoints or inference configurations. For Dynin-Omni, the accelerated run scores 84.23 on GSM8K, while the standard benchmark result is 83.7; the corresponding accelerated scores and throughput are reported in Tab. 15 and Fig. 11a. Qwen2.5-Omni-7B is evaluated in a zero-shot setting with exact autoregressive decoding using vLLM. For text-to-speech evaluation, Qwen3-TTS-1.7B is evaluated using vLLM-Omni, while Qwen2.5-Omni-7B is evaluated using vLLM.

G.3 Statistical Details

For the large-scale benchmark tables, we evaluate a single trained checkpoint using the official evaluation splits and protocols. We use item-bootstrap intervals for GSM8K and MATH-500 and four-decoding-seed intervals for GenEval, ImgEdit, TTS WER, and TTS UTMOS; diagnostic plots show benchmark-level dispersion.

Merging-strategy analysis. For Fig. 5, each point corresponds to one merged checkpoint under a specific merging strategy and interpolation ratio α. We compute original-capability retention over GSM8K, POPE, and MME-P, and new-modality retention over $\text{LibriSpeech}_{\text{clean}}$ ASR and TTS WER. For higher-is-better metrics, retention is computed by normalizing the score of the merged checkpoint by the corresponding reference checkpoint. For lower-is-better WER metrics, retention is computed using the inverse normalization so that larger values consistently indicate better retention. Each normalized retention score is clipped to $[0, 1]$ before averaging. The plotted coordinate is the mean retention over the corresponding benchmark group. Error bars denote standard deviation across the normalized benchmark-level retention scores within each group.

Omnimodal-generation analysis. For Fig. 6, each method and NFE budget is evaluated on the same Omni-StoryBench (AIDAS-Lab, 2026) examples. The main-table Overall (Tab. 3) equally averages text, image, and speech scores, each the mean of its two displayed metrics after scaling 0–1 metrics to 0–10. For this NFE analysis, Overall averages five metrics: BERTScore, text judge, CLIP similarity, image judge, and speech judge, with BERTScore and CLIP similarity similarly scaled. We omit speech metadata accuracy because it varies little across the tested budgets; the Omni judge is excluded from both averages and reported separately in the appendix tables. To visualize how evenly each method performs across evaluation categories, we compute the standard deviation across category-level scores at each method and NFE budget. Shaded regions denote ± 0.2 times this cross-category standard deviation around the NFE-specific Overall score.

H Serving Infrastructure

For serving, we use a vLLM-Omni (Yin et al., 2026)-compatible implementation for text-to-image generation, text-to-speech synthesis, and image editing. It follows the framework's stage-based execution interface and supports block-wise parallel decoding with iterative remasking for diffusion-based outputs. To further optimize inference efficiency, we incorporate Dynin-Omni with a unified inference framework supporting both autoregressive baselines and masked-diffusion generation. The masked-diffusion inference path is implemented on top of a customized version of dInfer (Ma et al., 2025b), extended for modality-aware iterative decoding with block-wise refinement for sequential modalities. To reduce overhead in iterative decoding, we retain dInfer's compiled execution path using `torch.compile`, CUDA Graph replay, and loop unrolling across diffusion refinement steps. For efficient long-context inference, we additionally integrate KV-cache approximation following Fast-dLLM (Wu et al., 2026), reusing KV states for unchanged context and committed regions across refinement steps.

I Additional Results

I.1 Training Recipe Controls

We provide the full controls underlying the summarized results in Tab. 4. For merging, $K{=}0$ denotes the checkpoint immediately after initialization and before Stage 2; $K{=}10$K denotes exactly 10,000 Stage 2 updates with identical data and compute. Original-capability retention averages clipped GSM8K, POPE, and MME-P ratios to the released backbone; new-modality retention averages clipped inverse-WER ratios to the Stage 1 checkpoint; and H-mean is their harmonic mean (Tab. 9). For the `[EOS]` controls, disabling supervision removes `[EOS]` from the training targets while still permitting its generation at inference. The termination metrics are computed over the pooled ASR and TTS samples. Non-premature termination means that the first `[EOS]` occurs at or after 80% of the reference length and before the decoding cap; premature termination occurs earlier, and a cap hit means that no `[EOS]` appears before the maximum decoding length. Repetition denotes a four-token sequence repeated consecutively and can overlap a cap hit (Tab. 10).

Table 9: Full merging controls at initialization and after 10K Stage 2 updates. Ratios denote backbone-to-Stage 1 interpolation weights.

K	Initialization	ASR/TTS WER↓	GSM8K↑	POPE↑	MME-P↑	H-mean↑
0	Stage 1 / no merge	12.50 / 31.50	4.2	30.7	413.8	38.9
0	Shared, 6:4	120.00 / 150.00	35.0	70.0	1000.0	25.8
0	Stage 1-only, 6:4	90.00 / 120.00	28.0	68.0	950.0	30.7
0	Modality-disent., 6:4	69.70 / 81.40	23.4	72.1	946.3	**39.3**
10K	No merge	10.50 / 20.00	18.0	50.0	820.0	66.1
10K	Modality-disent., 4:6	9.00 / 17.00	27.0	62.0	950.0	76.9
10K	Modality-disent., 6:4	**6.46 / 10.22**	**38.0**	**76.0**	**1150.0**	**88.3**

Table 10: Detailed controls for scheduled `[EOS]` supervision.

Stage	`[EOS]` supervision	ASR WER (S/D/I)↓	TTS WER (S/D/I)↓	Non-prem.↑	Prem.↓	Repeat↓	Cap↓
1	Enabled	131.80 (9.20/88.60/34.00)	55.00 (5.00/42.00/8.00)	**35%**	60%	**8%**	**5%**
1	Disabled	**100.65 (20.00/3.00/77.65)**	**35.07 (5.00/2.00/28.07)**	20%	**3%**	60%	77%
2	Disabled	14.00 (3.00/1.50/9.50)	24.00 (4.00/2.00/18.00)	30%	**2%**	55%	68%
2	Enabled	**6.46 (2.50/1.50/2.46)**	**10.22 (3.00/1.22/6.00)**	**94%**	3%	**4%**	**3%**

I.2 Text-Diffusion and Diffusion-MLLM Comparisons

We evaluate LLaDA-8B-Base and LLaDA-8B-Instruct from their released checkpoints; Trida-7B and Dream-7B use their reported results (Tab. 11).

Table 11: Comparison with text-diffusion baselines.

Model	MMLU↑	GSM8K↑	MATH-500↑	GPQA↑
LLaDA-Base†	65.9±0.8 (5-shot)	71.3±2.8 (5-shot)	30.0±5.1 (4-shot)	31.5±4.9 (5-shot)
LLaDA-Instruct†	65.4±1.5 (5-shot)	68.7±2.4 (4-shot)	29.8±4.2 (0-shot)	31.8±6.0 (5-shot)
Trida-7B (TrillionLabs, 2026)	67.2 (5-shot)	65.1 (0-shot CoT)	33.6 (0-shot CoT)	–
Dream-7B-Base (Ye et al., 2025c)	69.5 (5-shot)	77.2 (8-shot)	39.6 (4-shot)	36.6 (5-shot)
Dream-7B-Instruct (Ye et al., 2025c)	67.0 (4-shot)	81.0 (0-shot)	39.2 (0-shot)	33.0 (5-shot)
Dynin-Omni	**75.9** (0-shot)	**83.7** (0-shot)	**47.4** (0-shot)	**37.9** (0-shot)

We also evaluate LaViDa-LLaDA (Li et al., 2025c), LLaDA-o (You et al., 2026), SDAR-VL-8B-Instruct (Cheng et al., 2025), and Dream-VL (Ye et al., 2025b) on the same POPE, MME-P, and MMMU-val splits (Tab. 12).

Table 12: Comparison with recent diffusion MLLMs. Baseline entries report the published result followed by our evaluation in parentheses. N/R indicates that the corresponding result was not reported.

Benchmark	LaViDa-LLaDA	LLaDA-o	SDAR-VL-8B-Inst.	Dream-VL	Dynin-Omni
POPE (%)	N/R (83.4)	N/R (86.2)	N/R (84.3)	N/R (84.0)	**88.0**
MME-P	1365.6 (1349.4)	1412.0 (1454.0)	1632.0 (1465.0)	N/R (1546.9)	**1672.9**
MMMU-val (%)	43.3 (44.4)	44.9 (44.3)	53.0 (47.9)	52.2 (51.7)	52.8

Table 13: Evaluation uncertainty. Generation and speech metrics are aggregated over four decoding seeds, while reasoning metrics use item-level bootstrap estimates from fixed benchmark predictions. The reasoning estimates use five bootstrap resamples.

Metric	Mean	95% CI
GenEval	0.894	[0.871, 0.916]
ImgEdit	3.745	[3.710, 3.780]
GSM8K	82.34%	[80.29, 84.38]%
MATH-500	47.40%	[43.00, 51.80]%
TTS WER	2.69%	[1.94, 3.44]%
TTS UTMOS	4.049	[3.984, 4.114]

I.3 Statistical Uncertainty and Perceptual TTS Quality

We report decoding-seed intervals for generation and speech metrics and bootstrap intervals for reasoning benchmarks; UTMOS complements WER with perceptual speech quality (Tab. 13). The main tables report the fixed-run point estimates used for model comparison. To quantify evaluation uncertainty, we repeat image generation, image editing, and speech decoding with four random seeds and report the mean and 95% confidence interval across seeds. For GSM8K and MATH-500, we compute item-level bootstrap confidence intervals from the fixed benchmark predictions. The generation and speech means can differ from the main-table point estimates because they aggregate repeated decoding runs. The reasoning estimates use five bootstrap resamples; their sample means can also differ from the fixed-run point estimates.

I.4 Omni-StoryBench Decontamination

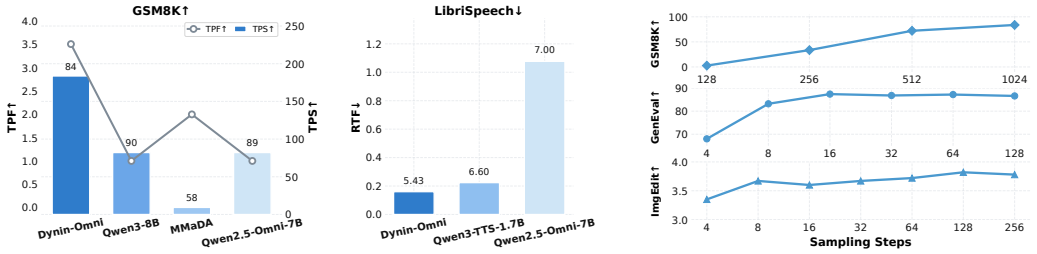
We flag a test example if it is identified by any of four overlap checks: an exact match to one of three sampled 50-character spans (GPT-4-style OpenAI (2023)), at least 20% token coverage at $L = 40$ or $L = 50$ (LLaMA-style Grattafiori et al. (2024)), or visual similarity between its target image and a training image. For the visual checks, we extract SSCD (Self-Supervised Copy Detection) ResNet-50 descriptors and compute cosine similarity; an example is flagged if its maximum similarity to any training image is at least 0.95, or at least 0.75 under the DataComp-style setting Gadre et al. (2023). The union flags 188 of the 900 examples as potential overlaps. After excluding them, we re-evaluate all nine systems on the same 712-item subset. The ranking is unchanged, and Dynin-Omni retains the highest Overall score, with 7.55 on the clean subset versus 7.63 on the full set (Tab. 14).

I.5 Inference Efficiency and Sampling Budget

With the accelerated dInfer/DualCache path, Dynin-Omni reaches 183.68 TPS on GSM8K, compared with 150.70 TPS for MMaDA under the same implementation settings (Tab. 15). The larger comparisons with Qwen baselines in Fig. 11a are implementation-level measurements across different decoding backends and should not be interpreted as an architecture-only speedup. For TTS, the accelerated path in Fig. 11a yields 5.43 WER, whereas the 2.1 WER in Tab. 2 uses the standard benchmark decoding configuration. Across sampling budgets, reasoning improves as the refinement budget increases, while image generation and editing reach their strongest or near-saturated performance at moderate step counts (Fig. 11b).

Table 14: Omni-StoryBench results on the full set and the 712-item clean subset. All systems are evaluated on the same clean examples. Overall averages the six displayed metrics after scaling BERTScore, CLIP similarity, and metadata accuracy to $[0, 10]$. Δ is Clean minus Full, computed from unrounded scores.

	Generator		Text↑		Image↑		Speech↑		Overall↑		
Model	Image	Speech	BERTScore	LLM Judge	CLIP Sim.	LLM Judge	Metadata Acc.	LLM Judge	Full	Clean	Δ
InternVL3.5-4B	FLUX.1 Kontext	Parler	0.87	7.46	0.75	5.46	0.55	8.06	7.12	7.12	+0.01
Qwen3.5-4B	FLUX.1 Kontext	Parler	0.88	8.19	0.77	6.42	0.52	8.19	7.44	7.42	-0.01
EMOVA-7B	FLUX.1 Kontext	Native	0.87	7.79	0.78	4.73	0.53	8.27	7.06	7.08	+0.02
Qwen2.5-Omni-7B	FLUX.1 Kontext	Native	0.88	7.96	0.76	5.55	0.54	8.14	7.24	7.24	+0.00
Emu3	Native	Parler	0.87	5.52	0.58	1.79	0.51	7.94	5.81	5.82	+0.02
MMaDA	Native	Parler	0.87	4.78	0.65	1.99	0.50	8.10	5.82	5.84	+0.02
AnyGPT	Native	Native	0.80	3.28	0.64	1.68	0.43	5.52	4.81	4.85	+0.04
HyperCLOVAX	Native	Native	0.87	6.43	0.68	2.53	0.54	8.20	6.32	6.33	+0.01
Dynin-Omni (Ours)	Native	Native	0.91	6.57	0.85	6.26	0.61	8.61	**7.63**	**7.55**	-0.08



(a) Accelerated inference on reasoning and TTS.

(b) Effect of diffusion sampling steps.

Figure 11: Inference efficiency analysis. (a) For reasoning on GSM8K, bars show tokens-per-second (TPS), the line shows tokens-per-forward (TPF), and labels show accuracy. For TTS on LibriSpeech$_{\text{clean}}$, bars show real-time-factor (RTF) and labels show WER. (b) Reasoning scales with larger refinement budgets, while image generation and editing saturate with moderate sampling steps.

I.6 Additional Quantitative and Qualitative Results

Textual reasoning Tab. 16 reports textual reasoning performance across representative language benchmarks, and Fig. 12 provides qualitative examples for textual reasoning, image understanding, and video understanding. Dynin-Omni remains competitive with prior unified and omnimodal models on the reported reasoning benchmarks, while text-specialized LLMs retain an advantage on several tasks.

Multimodal understanding Tab. 17 and Tab. 18 summarize image and video understanding benchmarks, with qualitative examples shown in Fig. 12. Dynin-Omni remains competitive with perception-centric omnimodal systems despite additionally supporting native image generation, editing, and speech synthesis within the same backbone. In particular, Dynin-Omni achieves strong MME-P and POPE performance while remaining competitive on MVBench, NExT-QA, and Video-MME.

Image generation Tab. 19 reports image generation results, and Fig. 13 provides qualitative generation examples. Dynin-Omni is competitive with prior unified models on GenEval and DPG-Bench and matches Qwen-Image on GenEval while also supporting video and speech.

Image editing Tab. 20 summarizes image editing performance across diverse editing categories, with qualitative examples shown in Fig. 14. Dynin-Omni demonstrates balanced editing capability across object addition, replacement, extraction, removal, and style transfer tasks. Its Overall editing score is close to the strongest unified baselines while the same backbone also supports speech input and output.

Table 15: Matched dInfer/DualCache comparison under shared implementation settings.

Model	Shots	GSM8K↑	TPS↑	TPF↑
MMaDA-MixCoT	5	47.60	150.70	4.09
Dynin-Omni	0	**84.23**	**183.68**	3.51

Speech recognition and synthesis Tab. 21 reports ASR and TTS performance. Dynin-Omni achieves competitive speech recognition and synthesis quality while using the same masked-diffusion backbone employed for text, image, and video modeling. Compared with the reported unified omni-modal baselines, Dynin-Omni obtains lower TTS WER on LibriSpeech test-clean while maintaining competitive ASR WER, without an external speech generator.

Omnimodal generation Tab. 22 and Tab. 23 provide modality-wise analysis of joint omnimodal generation and Modality Anchoring (MA) under non-block and block-wise decoding. Non-block decoding refines all target tokens in parallel, whereas block-wise decoding generates sequential blocks and refines tokens in parallel within each block, using a fixed block size of 8. We compare *Joint*, which generates text, image, and speech in a single masked-diffusion process, with *MA-Text*, *MA-Image*, and *MA-Speech*, which first generate an anchor modality and then condition the remaining modalities on it. We also evaluate omnimodal consistency using Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b) as a multimodal judge; the *Omni* score measures whether generated text, image, and speech remain semantically aligned and narratively coherent.

At 128 NFEs, Joint scores 7.93 Overall/6.97 Omni and the strongest single anchor scores 7.91/6.93; with 256 NFEs, the best two-stage anchor reaches 7.95/7.05. Thus, joint decoding achieves comparable cross-modal quality with a lower sampling budget. Performance improves with more NFEs and gradually saturates. Among MA variants, text is the strongest anchor, while MA-Image substantially degrades text-related metrics; for example, under block-wise decoding at 64 NFEs, MA-Image obtains 3.38 on text LLM evaluation compared with 6.34 for Joint generation, which also lowers Omni consistency. Speech scores remain relatively stable across strategies, but strong speech-specific performance alone does not imply higher global omnimodal coherence. Finally, non-block decoding generally achieves stronger Omni and Overall scores at larger NFEs, e.g., Joint improves from 6.69/7.82 under block-wise decoding at 128 NFEs to 6.97/7.93 under non-block decoding, although block-wise decoding remains competitive for sequential modalities such as speech.

J LICENSE AND SAFEGUARDS

Dynin-Omni model weights will be released under the CC BY 4.0 license, together with code and evaluation scripts for research use. We cite the datasets, tokenizers, pretrained models, benchmarks, and synthetic-data generation models used in training and evaluation throughout the paper, and will provide release documentation summarizing their sources, scales, licenses, and filtering procedures. The data-construction prompts, configurations, filters, and manifests described in App. E will also be released; restricted source media will be represented by source identifiers, hashes, and reconstruction scripts. For responsible release, we will provide model documentation describing intended uses, limitations, and misuse risks, along with usage guidelines that prohibit harmful applications such as impersonation, deception, harassment, and generation of unsafe or illegal content. We also recommend downstream deployments to use additional safety classifiers, provenance or watermarking mechanisms, and human review for sensitive use cases.

Table 16: Evaluation on textual reasoning benchmarks (higher is better for all metrics). Notation follows Tab. 1.

Model	MMLU (Hendrycks et al., 2021a)	ARC-C (Clark et al., 2018)	GSM8K (Cobbe et al., 2021)	MATH (Hendrycks et al., 2021b)	GPQA (Rein et al., 2024)
Text-Experts					
Llama-3-8B (Grattafiori et al., 2024)	66.6	-	55.3	20.5	25.8
Ministral-3-8B (Liu et al., 2026)	76.1	88.0	-	62.6	-
Qwen3-8B (Yang et al., 2025a)	76.9	-	89.8	60.8	44.4
LLaDA-8B-Base‡ (Nie et al., 2025)	65.9	–	71.3	30.0	31.5
LLaDA-8B-Instruct‡ (Nie et al., 2025)	65.4	–	68.7	29.8	31.8
Trida-7B (TrillionLabs, 2026)	67.2	-	65.1	33.6	-
Perception-centric					
Qwen2.5-Omni-7B (Xu et al., 2025a)	71.8	-	88.7	71.5	30.8
Baichuan-Omni-1.5 (Li et al., 2025d)	72.2	-	-	-	-
Unified					
Chameleon-7B (Team, 2024)	52.1	46.5	41.6	11.5	-
MMaDA (Yang et al., 2025b)	68.4	57.4	73.4	36.0	28.4
Show-o2* (Xie et al., 2025)	70.7	–	75.3	–	31.5
HyperCLOVAX-8B-Omni† (Team, 2026)	75.7	85.8‡	87.3	43.4‡	28.6‡
Dynin-Omni† (Ours)	75.9	68.5	83.7	47.4	37.9

Table 17: Evaluation on multimodal understanding (MMU) benchmarks (*image*). Higher is better for all metrics. Notation follows Tab. 1.

Model	POPE (Li et al., 2023)	MME-P (Fu et al., 2023)	GQA (Hudson & Manning, 2019)	MMMU (Yue et al., 2023)	MMBench (Liu et al., 2023)
MMU-Experts					
InternVL3-8B (Zhu et al., 2025)	90.4	1748.4	–	62.2	83.6
Ovis2-8B (Lu et al., 2024b)	88.6	1628.1	–	57.4	84.8
Qwen2.5-VL-7B (Bai et al., 2025)	85.9	1698.1	–	58.0	83.2
LLaDA-V (You et al., 2025)	–	1507.0	–	48.6	82.9
Perception-centric					
OpenOmni (Luo et al., 2025a)	86.4‡	1536.9‡	62.2‡	46.7	76.2
Baichuan-Omni-1.5 (Li et al., 2025d)	87.3‡	1632.1‡	58.6‡	53.9	85.6
Qwen2.5-Omni-7B (Xu et al., 2025a)	86.8‡	1481.3‡	49.1‡	59.2	81.8
OmniVinci (Ye et al., 2025a)	89.5‡	1651.0‡	65.7‡	49.7	-
Unified					
SEED-X (Ge et al., 2024)	84.2	1435.7	47.9	35.6	70.1
Janus-Pro (Chen et al., 2025b)	87.4	1567.1	62.0	41.0	79.2
BAGEL (Deng et al., 2025)	–	1687.0	–	55.3	85.0
Emu3* (Wang et al., 2024a)	85.2	–	60.3	31.6	58.5
MMaDA (Yang et al., 2025b)	86.1	1410.7	61.3	30.2	68.5
Fudoki (Wang et al., 2025a)	86.1	1485.4	57.6	34.3	73.9
Lumina-DiMOO (Xin et al., 2025)	87.4	1534.2	–	58.6	83.1
Show-o2* (Xie et al., 2025)	–	1620.5	63.1	48.9	79.3
NExT-OMNI*,† (Luo et al., 2025b)	87.4	1537.8	62.7	43.7	78.9
HyperCLOVAX-8B-Omni† (Team, 2026)	79.8‡	1307.0‡	45.5‡	31.0‡	65.9‡
Dynin-Omni† (Ours)	88.0	1672.9	57.0	52.8	74.7

Table 18: Evaluation on multimodal understanding (MMU) benchmarks (*Video*). Higher is better for all metrics. Notation follows Tab. 1.

Model	ActivityNet (Yu et al., 2019)	MVBench (Li et al., 2024b)	NExT-QA (Xiao et al., 2021)	TempCompass (Liu et al., 2024)	Video-MME (Fu et al., 2024)
MMU-Experts					
InternVL3-8B (Zhu et al., 2025)	63.9‡	73.2	82.0‡	70.4‡	66.0
InternVL2.5-8B (Chen et al., 2024c)	62.8‡	70.5	84.1‡	67.7‡	63.7
IXC-2.5 (Zhang et al., 2024a)	52.8	69.1	55.8	–	62.8
LLaVA-OV (Li et al., 2024a)	56.6	56.7	79.4	64.8	58.2
LLaDA-V (You et al., 2025)	–	53.1	–	–	56.4
Perception-centric					
MiniCPM-o-2.6 (Yao et al., 2024)	63.0	58.6	-	57.0	63.4
VITA-1.5 (Fu et al., 2025)	59.6	55.5	-	–	57.6
Baichuan-Omni-1.5 (Li et al., 2025d)	62.0	63.7	–	–	60.1
Qwen2.5-Omni-7B (Xu et al., 2025a)	62.7‡	70.3	76.4‡	69.3‡	64.3
Unified					
Show-o2* (Xie et al., 2025)	56.4	55.8	79.0	–	57.4
HyperCLOVAX-8B-Omni† (Team, 2026)	-	49.5‡	-	58.3‡	58.2
Dynin-Omni† (Ours)	56.8	66.4	81.3	66.7	63.4

Table 19: Evaluation on image generation benchmarks (higher is better for all metrics). Notation follows Tab. 1.

Model	GenEval (Ghosh et al., 2023)							DPG-Bench (Hu et al., 2024)
	Single Obj.	Two Obj.	Count.	Colors	Pos.	Color Attr.	Overall	
Generation-Experts								
HiDream-I1-Full (Cai et al., 2025)	1.00	0.98	0.79	0.91	0.60	0.72	0.83	85.9
FLUX.1 [Dev] (Labs, 2024)	0.98	0.81	0.74	0.79	0.22	0.45	0.66	83.8
Qwen-Image (Wu et al., 2025a)	0.99	0.92	0.89	0.88	0.76	0.77	0.87	88.3
GPT-Image 1 [High] (OpenAI, 2025)	0.99	0.92	0.85	0.92	0.75	0.61	0.84	85.2
Seedream 3.0 (Gao et al., 2025)	0.99	0.96	0.91	0.93	0.47	0.80	0.84	88.3
Unified								
Emu3* (Wang et al., 2024a)	0.99	0.81	0.42	0.80	0.49	0.45	0.66	81.6
Janus-Pro (Chen et al., 2025b)	0.99	0.89	0.59	0.90	0.79	0.66	0.80	84.2
BAGEL (Deng et al., 2025)	0.99	0.94	0.81	0.88	0.64	0.63	0.82	85.1
MMaDA (Yang et al., 2025b)	0.99	0.76	0.61	0.84	0.20	0.37	0.63	70.0
Fudoki (Wang et al., 2025a)	0.96	0.85	0.56	0.88	0.68	0.67	0.77	83.6
Lumina-DiMOO (Xin et al., 2025)	1.00	0.94	0.85	0.89	0.85	0.76	0.88	86.0
Show-o2* (Xie et al., 2025)	1.00	0.87	0.58	0.92	0.52	0.62	0.76	86.1
NExT-OMNI*,† (Luo et al., 2025b)	0.99	0.92	0.79	0.85	0.78	0.74	0.85	84.2
HyperCLOVAX-8B-Omni† (Team, 2026)	–	–	–	–	–	–	0.64	60.8‡
Dynin-Omni† (Ours)	0.98	0.93	0.82	0.90	0.86	0.74	0.87	85.5

Table 20: Evaluation on the ImgEdit benchmark (Ye et al., 2025d). Higher is better. All metrics are evaluated using GPT-4.1. *Overall* denotes the average score across all tasks. Notation follows Tab. 1.

Model	Add	Adjust	Extract	Replace	Remove	Background	Style	Action	Overall
Editing-Experts									
Step1X-Edit (Liu et al., 2025a)	3.88	3.14	1.76	3.40	2.41	3.16	4.63	2.52	3.11
OmniGen2 (Wu et al., 2025b)	3.57	3.06	1.77	3.74	3.20	3.57	4.81	4.68	3.55
Flux.1 Kontext [Pro] (Labs et al., 2025)	3.76	3.45	2.15	3.98	2.94	3.78	4.38	4.26	3.59
GPT-Image 1 [High] (OpenAI, 2025)	4.61	4.33	2.90	4.35	3.66	4.57	4.93	4.89	4.28
Qwen-Image (Wu et al., 2025a)	4.38	4.16	3.43	4.66	4.14	4.38	4.81	4.69	4.33
Unified									
UniWorld-V1 (Lin et al., 2025)	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.74	3.30
BAGEL (Deng et al., 2025)	3.56	3.31	1.70	3.30	2.62	3.24	4.49	4.17	3.30
LaViDa-O (Li et al., 2025b)	4.04	3.62	2.01	4.39	3.98	4.06	4.82	3.54	3.81
Lumina-DiMOO (Xin et al., 2025)	3.82	–	–	3.83	2.76	–	4.18	–	–
HyperCLOVAX-8B-Omni† (Team, 2026)	–	–	–	–	–	–	–	-	3.83
Dynin-Omni† (Ours)	4.15	3.45	3.30	4.29	3.98	3.74	4.10	3.54	3.82

Table 21: Evaluation on automatic speech recognition (ASR) and text-to-speech (TTS) benchmarks. Lower is better. All metrics report word error rate (WER) and are evaluated using Whisper-large-v3 (Radford et al., 2022). Notation follows Tab. 1.

(a) ASR Benchmarks (WER↓)

Model	LibriSpeech (Panayotov et al., 2015)	
	Clean	Other
ASR-Experts		
SpeechVerse (Das et al., 2024)	2.1	4.4
MinMo (Chen et al., 2025a)	1.7	3.9
Qwen2-Audio (Chu et al., 2024)	1.6	3.6
Qwen3-ASR-1.7B (Shi et al., 2026)	1.6	3.4
Perception-centric		
EMOVA-7B (Chen et al., 2024a)	4.0	–
OpenOmni (Luo et al., 2025a)	3.1	7.0
VITA-1.5 (Fu et al., 2025)	3.0	7.2
Qwen2.5-Omni-7B (Xu et al., 2025a)	1.8	3.4
MiniCPM-o-2.6 (Yao et al., 2024)	1.7	–
Unified		
HyperCLOVAX-8B-Omni (Team, 2026)	1.9	4.5‡
NExT-OMNI* (Luo et al., 2025b)	3.1	7.0
Dynin-Omni (Ours)	2.1	4.8

(b) TTS Benchmarks (WER↓)

Model	LibriSpeech	SEED-TTS (Anastassiou et al., 2024)
TTS-Experts		
MaskGCT (Wang et al., 2024b)	2.6	2.6
Llasa-8B (Ye et al., 2025e)	1.5	3.0
MiniMax-Speech (Zhang et al., 2025)	–	1.7
Qwen3-TTS-1.7B (Hu et al., 2026)	6.6‡	1.2
Perception-centric		
EMOVA-7B (Chen et al., 2024a)	3.6	–
OpenOmni (Luo et al., 2025a)	3.4	–
Qwen2.5-Omni-7B (Xu et al., 2025a)	7.0‡	2.3
Unified		
HyperCLOVAX-8B-Omni (Team, 2026)	7.9‡	7.7‡
NExT-OMNI* (Luo et al., 2025b)	3.1	–
Dynin-Omni (Ours)	2.1	4.3

Table 22: Omni-StoryBench results under non-block decoding.

NFE	Method	Text↑		Image↑		Speech↑	Omni	Overall
		BERT	LLM Judge	CLIP	LLM Judge	LLM Judge		
16	Joint	0.91	6.17	0.85	6.29	8.45	6.32	7.71
	MA-Text	0.91	5.88	0.84	6.03	8.45	6.13	7.57
	MA-Image	0.85	2.27	0.84	5.44	8.44	3.31	6.61
	MA-Speech	0.91	5.60	0.84	6.07	8.50	5.92	7.53
32	Joint	0.91	6.03	0.86	6.39	8.68	6.51	7.76
	MA-Text	0.91	5.50	0.85	6.29	8.75	6.18	7.62
	MA-Image	0.85	2.54	0.85	6.30	8.55	4.20	6.89
	MA-Speech	0.91	5.37	0.85	6.30	8.59	6.02	7.57
64	Joint	0.92	6.51	0.86	6.37	8.69	6.83	7.86
	MA-Text	0.91	6.22	0.85	6.33	8.73	6.68	7.80
	MA-Image	0.87	3.20	0.85	6.35	8.58	4.82	7.07
	MA-Speech	0.91	5.86	0.85	6.38	8.58	6.43	7.70
128	Joint	0.92	6.78	0.86	6.40	8.70	6.97	7.93
	MA-Text	0.92	6.66	0.86	6.40	8.74	6.93	7.91
	MA-Image	0.88	3.81	0.86	6.38	8.65	5.24	7.24
	MA-Speech	0.92	6.29	0.86	6.39	8.55	6.57	7.79
256	MA-Text	0.92	6.89	0.86	6.36	8.74	7.05	7.95
	MA-Image	0.89	4.18	0.86	6.41	8.69	5.51	7.34
	MA-Speech	0.92	6.52	0.86	6.37	8.61	6.72	7.85

Table 23: Omni-StoryBench results under block-wise decoding.

NFE	Method	Text↑		Image↑		Speech↑	Omni	Overall
		BERT	LLM Judge	CLIP	LLM Judge	LLM Judge		
16	Joint	0.90	5.05	0.85	6.25	8.32	5.60	7.43
	MA-Text	0.90	4.62	0.84	6.05	8.33	5.24	7.27
	MA-Image	0.83	1.82	0.84	6.06	8.26	3.02	6.56
	MA-Speech	0.90	4.53	0.84	6.08	8.44	5.06	7.28
32	Joint	0.91	5.92	0.85	6.37	8.35	6.28	7.66
	MA-Text	0.90	5.33	0.85	6.27	8.35	5.90	7.50
	MA-Image	0.85	2.54	0.85	6.28	8.29	3.83	6.83
	MA-Speech	0.90	5.05	0.85	6.28	8.48	5.67	7.46
64	Joint	0.91	6.34	0.86	6.43	8.43	6.60	7.78
	MA-Text	0.91	6.23	0.86	6.39	8.42	6.54	7.74
	MA-Image	0.87	3.38	0.85	6.35	8.39	4.76	7.08
	MA-Speech	0.91	5.81	0.86	6.37	8.48	6.26	7.66
128	Joint	0.91	6.53	0.86	6.37	8.50	6.69	7.82
	MA-Text	0.91	6.51	0.86	6.33	8.45	6.66	7.80
	MA-Image	0.88	3.97	0.86	6.39	8.43	5.26	7.23
	MA-Speech	0.91	6.14	0.86	6.40	8.57	6.49	7.76
256	MA-Text	0.91	6.53	0.86	6.35	8.42	6.75	7.81
	MA-Image	0.88	4.01	0.86	6.40	8.40	5.27	7.24
	MA-Speech	0.89	6.10	0.84	6.22	8.44	6.43	7.62

Question

John drives for 3 hours at a speed of 60 mph and then turns around because he realizes he forgot something very important at home. He tries to get home in 4 hours but spends the first 2 hours in standstill traffic. He spends the next half-hour driving at a speed of 30mph, before being able to drive the remaining time of the 4 hours going at 80 mph. How far is he from home at the end of those 4 hours?

Answer

<think>
First, calculate the distance John initially traveled by multiplying his speed by the time he spent driving: 60 mph * 3 hours = 180 miles.

Next, determine the distance he covered during his return trip. The first 2 hours were in standstill, so he traveled 0 miles. The next half-hour of driving at 30 mph is 15 miles. For the remaining 1.5 hours, he drove at 80 mph, which is 120 miles. Adding these distances together: 0 + 15 + 120 = 135 miles.

Finally, subtract the distance he covered during his return trip from the distance he initially traveled to find out how far he is from home: 180 - 135 = 45 miles.
</think>

John drives for 3 hours at 60 mph, covering $60 \times 3 = 180$ miles.
On his return trip:

- First 2 hours in standstill: 0 miles
- Next 0.5hours at 30 mph: $30 \times 0.5 = 15$miles
- Remaining 1.5hours at 80 mph: $80 \times 1.5 = 120$ miles

Total return distance: $0 + 15 + 120 = 135$ miles.
Distance from home: $180 - 135 = 45$

Question

Explain the visual content of the image in great detail.

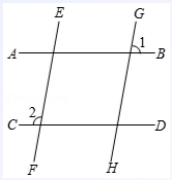


Answer

In the image, a person's hand is seen stirring a bowl of mashed potatoes. The hand, appearing on the right side of the frame, is slightly blurred, leaving the potatoes to be the central focus. The bowl, which is green in color, is placed on a gray surface. The surface beneath the bowl and the hand is not visible in the image. The potatoes in the bowl are a pale yellow color, indicating that they are being mashed. The relative position of the bowl and the hand suggests that the person is in the process of stirring the mashed potatoes.

Question

In the provided diagram, let's denote angle 1 as variable u where u = 80°. If AB is parallel to CD and EF is parallel to



GH, what is the degree measure of angle 2, expressed as variable v in relation to u?

Answer

Let angle 3 be the angle corresponding to angle 1 on line CD, and let angle 4 be the angle corresponding to angle 2 on line GH. To determine the value of v, the measure of angle 2, we apply the geometric principle stating that corresponding angles formed by parallel lines are congruent. Since AB is parallel to CD, angle 1 is equal to angle 3. Additionally, since EF is parallel to GH, angle 2 is equal to angle 4. Therefore, we can express v as 180° - u. Substituting the value of u, we have v = 180° - 80° = 100°. Hence, the degree measure of angle 2 is 100°.

Question

Please describe this video in detail.



Answer

The video captures the vibrant and dynamic atmosphere of a running track during a race. The scene is set on a red running track, with several people running along the sides, dressed in various athletic outfits. The background features tall buildings and trees, adding to the urban setting. The track is lined with colorful banners and flags, enhancing the festive atmosphere. The video is filled with excitement and energy as the runners sprint along the track, with some wearing orange shirts and others in different colors. The video captures the essence of a community event, highlighting the joy and determination of the participants.

Question

Please describe this video in detail.



Answer

The video captures a vibrant and bustling urban street in a city, likely in Korea, characterized by a mix of traditional and modern architecture. The street is filled with colorful neon signs and advertisements on both sides, creating a lively and dynamic atmosphere. Various shops and stalls are visible along the street, contributing to the energetic environment. The street is relatively empty, with a few people walking along the sidewalk, adding a sense of calm to the otherwise busy scene. The sky is overcast, providing a soft light that enhances the vibrant colors of the signs and advertisements. Throughout the video, the scene consistently captures the essence of a lively urban setting.

Figure 12: Qualitative examples of Dynin-Omni for textual reasoning (top-left), image understanding (top-right), and video understanding (bottom).



Figure 13: Qualitative examples of Dynin-Omni for image generation task.

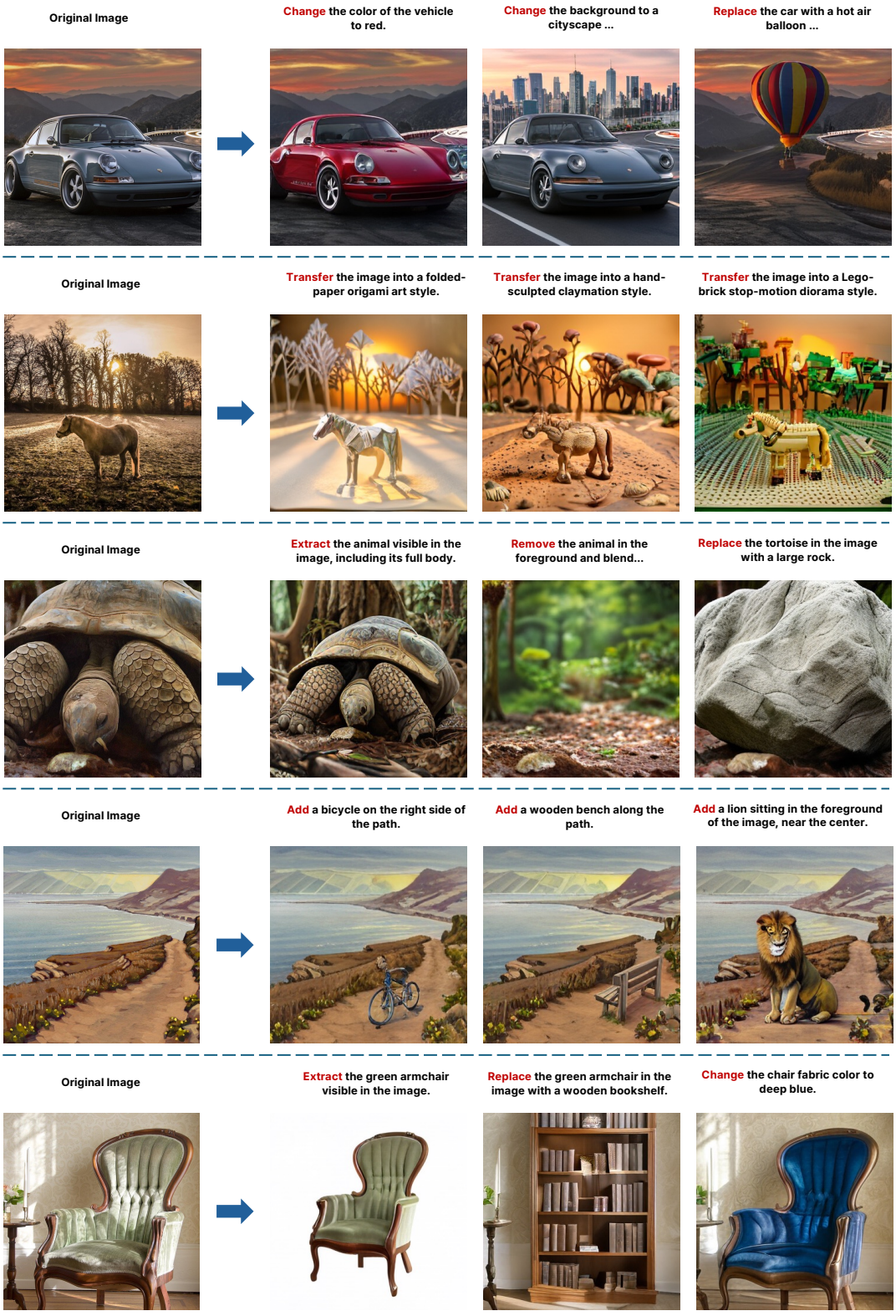


Figure 14: Qualitative examples of Dynin-Omni for image editing task.