

FLUID-V: Elastic-Horizon Diffusion Decoding for Vision-Language Models

Xiangyu Ma¹, Teng Xiao², Zuchao Li^{1*}, Ping Wang³, Hai Zhao⁴

¹School of Artificial Intelligence, Wuhan University

²School of Computer Science, Wuhan University

³School of Cyber Science and Engineering, Wuhan University

⁴AGI Institute, School of Computer Science, Shanghai Jiao Tong University

{maxiangyu, xiaoxiao, zcli-charlie, wangping}@whu.edu.cn

zhaohai@cs.sjtu.edu.cn

Abstract

Autoregressive vision-language models achieve strong multimodal understanding, but their sequential decoding becomes increasingly inefficient as multimodal responses grow longer. Adapting AR-VLMs to diffusion-style block decoding is promising, yet fixed refinement blocks struggle with sharply varying multimodal uncertainty in OCR, chart, document, and reasoning-heavy outputs. Binary mask-or-token states further limit local refinement, as masking discards partial evidence while premature commitment can corrupt the causal prefix. We present FLUID-V, a lightweight AR-to-diffusion adaptation framework for reliable local multimodal refinement. FLUID-V selects refinement spans with elastic horizons and represents unresolved positions as evolving soft semantic mixtures in embedding space. Only confident predictions are committed to the persistent KV cache, preserving causal multimodal conditioning and autoregressive compatibility. Experiments with OpenPangu-VL-7B and 100K self-distilled samples show competitive performance across diverse multimodal benchmarks, with ablations confirming the importance of adaptive horizons and stateful intermediate representations. Our code is available at <https://github.com/Oli-lab-nun/FLUID-V>.

1 Introduction

Autoregressive vision-language models (AR-VLMs) are the standard interface for multimodal generation, providing strong causal grounding by conditioning each token on the image and accepted prefix (Peng et al., 2024). However, this inherently sequential formulation requires L distinct decoding steps for a response of length L , making computational costs prohibitive for long visual descriptions and multi-step reasoning (Shi et al., 2026).

*Corresponding author. This work was supported by the CIPS-SMP-Zhipu Large Model Fund, the National Natural Science Foundation of China (No. 62306216), and the Technology Innovation Program of Hubei Province (No. 2024BAB043).

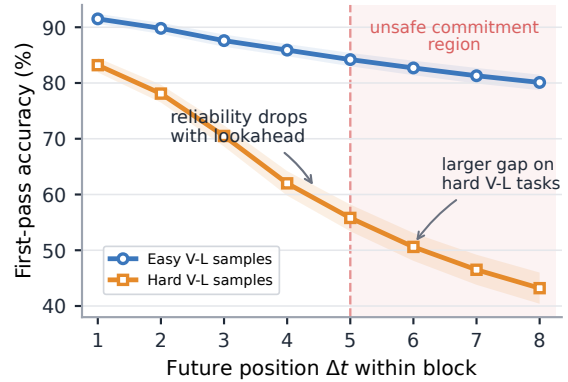


Figure 1: Reliability decay under fixed-horizon refinement. First-pass accuracy decreases as the predicted position moves farther from the committed prefix, such as OCR-, chart-, and reasoning-sensitive samples. The curves are measured from max-horizon probes on representative validation samples, grouped by low- and high-uncertainty visual-language regions.

Diffusion-style block decoding offers a path toward parallel generation by predicting multiple future tokens simultaneously. Yet, directly replacing AR decoding with bidirectional masked denoising discards the causal prefix semantics and KV-cache compatibility central to pretrained AR-VLMs (Li et al., 2025b). While recent AR-to-diffusion methods explore semi-autoregressive generation (Nie et al., 2025; Ye et al., 2025), they predominantly rely on fixed refinement spans and hard mask-to-token transitions, leaving open how to achieve parallel refinement without sacrificing causal reliability.

This pursuit is constrained by the *Entropy-Horizon Dilemma*: the severe misalignment between rigid generation windows and the highly dynamic entropy of multimodal contexts. As illustrated in Figure 1, easy descriptive spans exhibit low entropy and can be reliably predicted several steps ahead, whereas precision-critical segments—such as OCR strings and reasoning transitions—trigger sharp spikes in uncertainty. A fixed block size

thus disrupts the model’s intrinsic semantic rhythm: large blocks in high-entropy regions force premature predictions that degrade quality, while conservative blocks in low-entropy regions fail to exploit parallel decoding potential.

Compounding this structural dilemma is the brittle binary mask-or-token interface used for unresolved future positions. Repeated masking discards partial visual-language evidence accumulated across refinement rounds, while premature token commitment introduces irreversible errors into the causal prefix. Although prior work explores continuous embedding-space generation (Chen et al., 2026b; Hersche et al., 2025; Hu et al., 2026), stateful intermediate representations are particularly important for VLMs, where unresolved positions may already encode useful yet uncertain visual-semantic evidence before they can be safely discretized.

To address these limitations, we propose FLUID-V, a lightweight AR-to-diffusion adaptation framework that decouples temporary refinement from causal commitment. At each decision point, an elastic horizon dynamically selects a refinement span according to the model’s local competence. Within this span, unresolved positions are maintained as evolving intermediate states in continuous embedding space, while only sufficiently confident predictions are committed to the persistent KV cache, thereby preserving AR causal conditioning. Built on OpenPangu-VL-7B with just 100K self-distilled samples, FLUID-V achieves competitive performance across diverse multimodal benchmarks, with ablations showing that efficient VLM decoding requires not merely larger fixed blocks, but quality-aware horizon control together with continuous modeling of unresolved semantics.

2 Related Work

2.1 Diffusion Vision-Language Models

Diffusion VLMs extend denoising-based generation to multimodal instruction following, building on continuous embedding-space or discrete masked-token diffusion (Li et al., 2022; Gong et al., 2022; He et al., 2024; Austin et al., 2021; Sahoo et al., 2024; Nie et al., 2025; Ye et al., 2025). Recent multimodal variants connect diffusion decoders with visual encoders, instruction tuning, or hybrid AR–diffusion components (You et al., 2025; Li et al., 2026; Yu et al., 2025; Yang et al., 2026; Xie et al., 2025; Tong et al., 2025; Ma et al., 2025; Kou et al., 2024). These works establish the viability of

diffusion for multimodal modeling, but often rely on diffusion-native training or bidirectional denoising. FLUID-V instead adapts a causal AR-VLM while preserving prefix semantics and cache-compatible commitment.

2.2 Adapting Autoregressive Models to Diffusion

Recent work adapts AR language models to diffusion-style generation to reuse pretrained priors (Ma et al., 2026; Liu et al., 2025; Wu et al., 2023; Arriola et al., 2025a; Gong et al., 2025; Cheng et al., 2025a; Liu et al., 2025; Wu et al., 2025; Chen et al., 2025). In VLMs, A2D-VL, DiffusionVL, SDAR-VL, Fast-dVLM, and BARD explore annealed windows, block decoding, KV-cache reuse, and stabilized conversion recipes (Arriola et al., 2025b; Cheng et al., 2025b; Wu et al., 2026; Chen et al., 2026a). FLUID-V asks how the block should be chosen under heterogeneous multimodal uncertainty and how unresolved positions should be represented before commitment. This distinction is important for VLMs, where a single response may alternate between easy descriptive spans and precision-sensitive OCR, chart, document, or reasoning tokens.

2.3 Intermediate States for Diffusion Decoding

Standard masked diffusion uses a binary state: a position is masked or discretized. DMax reduces error accumulation through self-revision, SoftMask blends retained masks with top- k prediction feedback, and ELF motivates continuous embedding-space generation before final readout (Chen et al., 2026b; Hersche et al., 2025; Hu et al., 2026). Unlike these methods, FLUID-V couples online competence-boundary prediction with stateful Top- k representations for unresolved causal slots and confidence-gated separation between temporary refinement and persistent causal commitment.

3 Preliminaries

Consider a vision-language dataset $\mathcal{D}$ of triples (I, x, y) , where I is the visual input, x is the instruction, and $y = (y_1, \dots, y_L) \in \mathcal{V}^L$ is the response. A visual encoder and projector g_ψ map I into visual features, yielding the multimodal condition

$$\mathbf{v} = g_\psi(I), \quad c = (\mathbf{v}, x). \quad (1)$$

The goal is to learn $p_\theta(y \mid c)$. In all formulations below, c remains clean conditioning context, and only

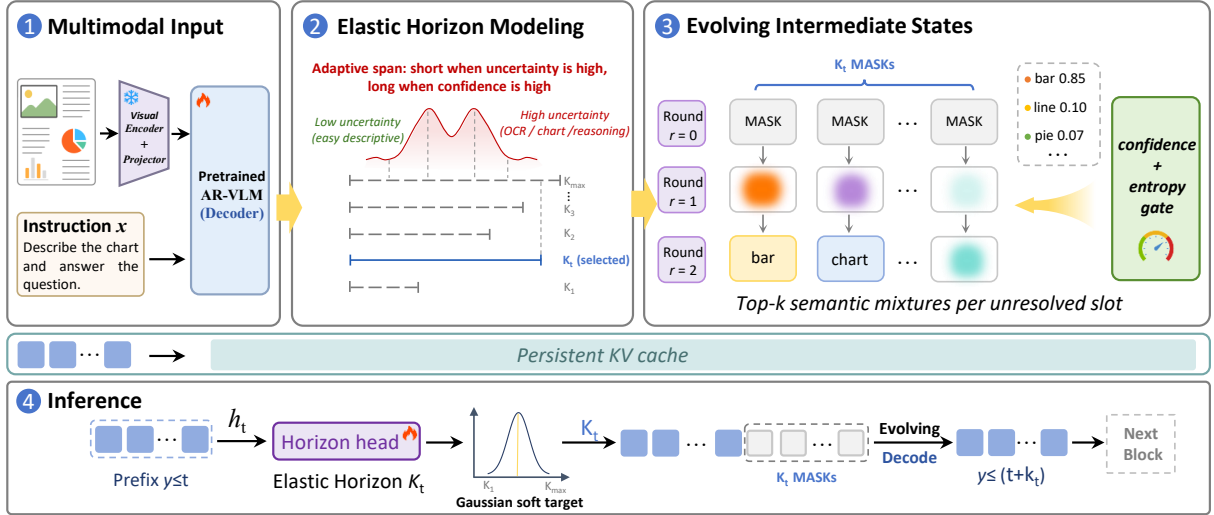


Figure 2: Given the multimodal condition and the committed causal prefix, FLUID-V predicts an elastic horizon that determines how many future slots should be locally refined. Within the selected span, unresolved slots are represented by evolving intermediate states, which accumulate partial visual-language evidence through top- k semantic mixtures and confidence-entropy gated updates. Accepted tokens are committed and replayed into the persistent KV cache, preserving the autoregressive inference interface.

response-side tokens are generated or corrupted.

3.1 Diffusion Vision-Language Modeling

Discrete diffusion replaces autoregressive factorization with a masked corruption-denoising process. Let z_s be the corrupted response at noise level $s \in [0, 1]$, and $\mu = [\text{MASK}]$. The absorbing forward process is:

$$q(z_{s,i} = a \mid y_i) = \alpha_s \mathbf{1}_{a=y_i} + (1 - \alpha_s) \mathbf{1}_{a=\mu}, \quad (2)$$

where α_s governs the noise schedule. The denoiser predicts clean tokens from z_s by minimizing:

$$\mathcal{L}_{\text{diff}} = -\mathbb{E} \left[\lambda(s) \sum_{i \in \Omega_s} \log p_{\theta}(y_i \mid c, z_s) \right], \quad (3)$$

with $\Omega_s = \{i : z_{s,i} = \mu\}$. Standard diffusion VLMs typically employ bidirectional attention over the corrupted sequence.

3.2 Block-Wise Semi-Autoregressive Strategies

Block-wise strategies partition the response y into distinct blocks (Arriola et al., 2025a). Denoting Y_b as the b -th block and $Y_{<b}$ as the committed prefix, the likelihood is factorized as:

$$p_{\text{SAR}}(y \mid c) = \prod_{b=1}^B p_{\theta}(Y_b \mid c, Y_{<b}). \quad (4)$$

Within each Y_b , models generally apply bidirectional diffusion over a *fixed* span. In contrast,

FLUID-V introduces *causal local refinement*: dynamically scoring multiple future slots within an elastic span while strictly preserving the AR commitment interface.

4 FLUID-V

Given a multimodal condition c and an accepted prefix $y_{\leq t}$, the framework dynamically predicts a horizon K_t , instantiates temporary states $z_{t,1:K_t}$, refines them under causal block attention, and commits only a sufficiently confident prefix.

4.1 Causal Local Diffusion Blocks

FLUID-V applies diffusion-style supervision exclusively to the response-side text. At a given trigger position t , the model appends $K_t \leq K_{\max}$ temporary slots immediately after the accepted prefix. Rather than operating on a fully corrupted sequence, these slots are anchored to a clean, multimodal prefix. The supervised target pairs for these local slots are defined as:

$$\mathcal{T}_t = \left\{ (z_{t,j}^{(0)}, y_{t+j}) \mid 1 \leq j \leq \min(K_t, L - t) \right\}, \quad (5)$$

where $z_{t,j}^{(0)} = e(\mu)$ initializes the slot with the mask embedding.

To maintain autoregressive compatibility, the expanded sequence utilizes a strict causal block mask. Let m_i indicate whether position i is an

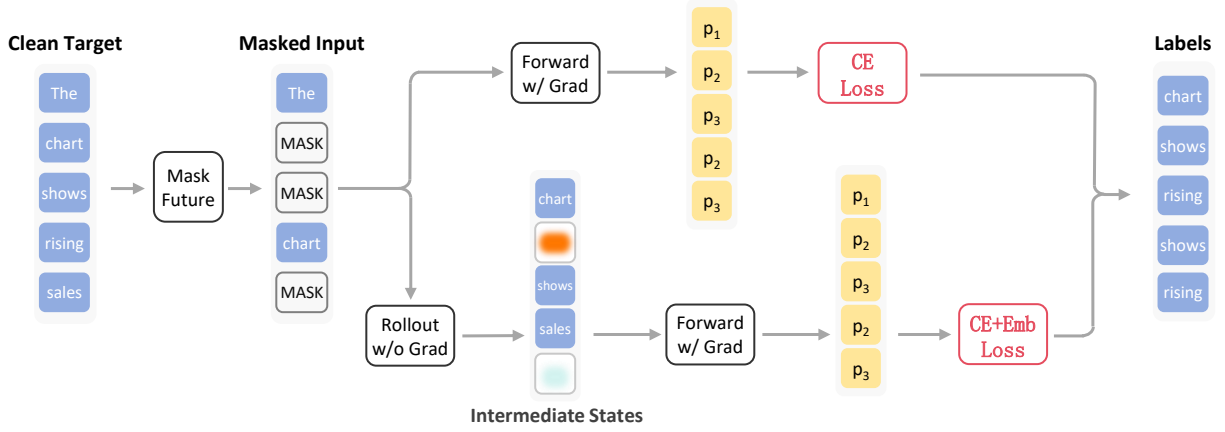


Figure 3: Stage-I training with direct and recurrent views. The direct view supervises masked future slots with distance-decayed cross-entropy. The recurrent view first performs a no-gradient rollout to separate accepted and unresolved slots, then performs a gradient-enabled pass using predicted token embeddings for accepted slots and intermediate states for unresolved slots.

inserted temporary slot, κ_i its original absolute column index, β_i its localized block ID, and r_i its relative rank within the block. The attention visibility $A_{ij} = 1$ is governed by:

$$\begin{aligned} A_{ij} &= C_{ij} \vee S_{ij}, \\ C_{ij} &= \neg m_j \wedge \kappa_j \leq t, \\ S_{ij} &= m_i \wedge m_j \wedge (\beta_i = \beta_j) \wedge (r_j \leq r_i). \end{aligned} \quad (6)$$

This masking strategy ensures that clean tokens only attend to non-future clean tokens (C_{ij}), while temporary slots are permitted to attend to the clean prefix and earlier slots within their shared refinement block (S_{ij}). Consequently, FLUID-V evaluates multiple future slots simultaneously while preserving a unidirectional dependency structure.

4.2 Evolving Intermediate States

To avoid the information loss inherent in binary mask-to-token transitions, FLUID-V models unresolved positions within a continuous embedding space. Each temporary slot maintains an evolving intermediate state $z_j^{(r)}$ at refinement round r .

During each local forward pass, the state $z_j^{(r)}$ yields a predictive distribution $p_j^{(r)}$. Rather than immediately discretizing the output, we construct a soft semantic proposal from the top- k candidates $\mathcal{V}_j$:

$$\tilde{z}_j^{(r)} = \sum_{v \in \mathcal{V}_j} \pi_{j,v}^{(r)} e(v), \quad (7)$$

where $e(v)$ is the token embedding and $\pi_{j,v}^{(r)}$ is the probability normalized over $\mathcal{V}_j$. This mixture pre-

Algorithm 1 Dynamic Causal Inference

Require: Model p_θ , Horizon Head f_ϕ , prompt c , threshold η , limit T

```

1: Prefill  $c$  and initialize persistent KV cache  $\mathcal{C}$ 
2:  $y \leftarrow \emptyset$ 
3: while  $|y| < T$  and EOS is not generated do
4:    $h_t \leftarrow \text{BoundaryState}(y)$ 
5:    $K_t \leftarrow \arg \max_k p_\phi(k | h_t)$   $\triangleright$  plan horizon
6:    $z_{1:K_t}^{(0)} \leftarrow e(\mu); \hat{y} \leftarrow \emptyset$ 
7:   while unresolved temporary states exist do
8:      $P, z'_{1:K_t} \leftarrow \text{Refine}(\mathcal{C}, \hat{y}, z)$ 
9:      $a \leftarrow \text{LongestConfidentPrefix}(P, \eta)$ 
10:    if  $a = 0$  then
11:       $a \leftarrow 1$   $\triangleright$  monotonic progress
12:    end if
13:     $\hat{y} \leftarrow [\hat{y}; \arg \max P_{1:a}]$ 
14:    Remove accepted slots and shift unresolved states
15:     $z \leftarrow z'_{\text{remaining}}$ 
16:  end while
17:  Replay  $\hat{y}$  through decoder to update  $\mathcal{C}$ 
18:   $y \leftarrow [y; \hat{y}]$ 
19: end while
20: return  $y$ 

```

serves multiple plausible semantics before discrete commitment.

The state is then updated via a dynamic gate $\alpha_j^{(r)}$:

$$z_j^{(r+1)} = \text{LN} \left((1 - \alpha_j^{(r)}) z_j^{(r)} + \alpha_j^{(r)} \tilde{z}_j^{(r)} \right), \quad (8)$$

where LN denotes Layer Normalization. The update coefficient $\alpha_j^{(r)}$ scales with the maximum prediction confidence $\max_v \pi_{j,v}^{(r)}$ and is penalized by the top- k entropy $H_j^{(r)}$. Consequently, confident, low-entropy predictions converge rapidly toward discrete embeddings, while uncertain predictions retain their evolving intermediate states to accumulate further context.

4.3 Elastic Horizon Calibration

The horizon K_t dictates the breadth of temporary states opened for local refinement. A structurally sound horizon must terminate before the decoder ventures into high-uncertainty regions. To derive a training target, we conduct a max-horizon probe using the frozen adapted decoder, recording the cross-entropy loss ℓ_{t+i} at each future slot i . Let $\bar{\ell}_{t,k}$ denote the average loss over the first k probed slots. The oracle horizon is established at the competence boundary:

$$K_t^* = \max(1, \max\{k : \bar{\ell}_{t,k} < \tau\}), \quad (9)$$

Rather than employing a rigid one-hot target, we smooth the horizon label using a Gaussian distribution centered at K_t^* :

$$q_t(k) \propto \exp\left(-\frac{(k - K_t^*)^2}{2\sigma^2}\right), \quad (10)$$

where the distribution is subsequently normalized such that $\sum_{k=1}^{K_{\max}} q_t(k) = 1$. This soft target respects the ordinal nature of spatial horizons, ensuring that adjacent span predictions are penalized proportionally. The horizon head is optimized via cross-entropy: $\mathcal{L}_K = -\sum q_t(k) \log p_\phi(k | h_t)$, where h_t is the boundary hidden state.

4.4 Training

Training operates in two decoupled stages.

Stage I: Causal state refinement. As illustrated in Figure 3, Stage I alternates between a direct view and a recurrent view. The direct view optimizes localized prediction using a distance-decayed cross-entropy loss, where $w_j = \exp(-(j-1)/\gamma)$ downweights slots further from the prefix:

$$\mathcal{L}_{\text{dir}} = \frac{\sum_{\mathcal{M}} w_j \text{CE}(p_{t,j}, y_{t+j})}{\sum_{\mathcal{M}} w_j}. \quad (11)$$

The recurrent view simulates partially refined local contexts by executing a no-gradient rollout, partitioning slots into an accepted set $\mathcal{A}_t = \{j : p_{t,j}^{\max} \geq \eta_{\text{tr}}\}$ and an unresolved set $\mathcal{U}_t$. It then constructs a mixed recurrent context:

$$\bar{e}_{t,j} = \begin{cases} e(\hat{y}_{t+j}), & j \in \mathcal{A}_t, \\ z_j^{(1)}, & j \in \mathcal{U}_t. \end{cases} \quad (12)$$

A subsequent gradient-enabled pass optimizes $\mathcal{L}_{\text{rec}} = (1 - \lambda_{\text{emb}})\mathcal{L}_{\text{CE}} + \lambda_{\text{emb}}\mathcal{L}_{\text{emb}}$, where $\mathcal{L}_{\text{emb}} = |\mathcal{U}_t|^{-1} \sum_{\mathcal{U}_t} \|z_j^{(1)} - e(y_{t+j})\|_2^2$ applies mean squared error regression exclusively on unresolved states.

Stage II: horizon head calibration. Stage II freezes the Stage I decoder and optimizes only $\mathcal{L}_K$. Freezing makes the target horizon reflect the learned decoder’s competence boundary instead of drifting with the state model.

4.5 Dynamic Causal Inference

Algorithm 1 summarizes the generation pipeline. Following the standard multimodal prefill, the horizon head predicts K_t at each decision boundary, opening an elastic temporary refinement block. Post-refinement, FLUID-V extracts the longest monotonic confident prefix:

$$a_t = \max \left\{ a \mid \min_{1 \leq j \leq a} p_{t,j}^{\max} \geq \eta \right\}. \quad (13)$$

Crucially, temporary KV entries corresponding to intermediate states are discarded after local refinement. The dynamically accepted a_t tokens are subsequently replayed through the causal model to update the persistent KV cache. Thus, FLUID-V fully encapsulates intermediate states within the local refinement loop, maintaining compatibility with standard AR cache structure.

5 Experiments

5.1 Experimental Setup

Backbone and adaptation data. FLUID-V is initialized from OpenPangu-VL-7B (OpenPangu Team, 2025) and reuses its visual encoder, projector, tokenizer, and language decoder as the AR-VLM backbone. We construct a balanced self-distillation corpus from multiple multimodal instruction sources, sampling 20k instruction-response pairs from each source dataset (including **Infinity-Instruct-7M** (Li et al., 2025a), **deepctrl-sft-data** (zh) (DeepCtrl, 2024), **moss-003-sft** (no-tools) (Sun et al., 2024), **UltraChat** (Ding et al., 2023), and **llava-next** (Li et al., 2024)). All responses are normalized into the standard direct-answer format of the backbone model, enabling controlled AR-to-diffusion adaptation under a compact data budget.

Training protocol. Training proceeds in two stages. Stage I trains the causal local refinement module and evolving intermediate states with LoRA finetuning. We use LoRA rank 16, LoRA alpha 32, maximum horizon $K_{\max} = 16$, learning rate 2×10^{-4} , cosine decay, warmup ratio 0.1, bf16 precision, and maximum text length 1024. Stage I

Model	Size	Data	MMMU [val]	MMMU-Pro [vision]	AI2D [test]	MMBench [en-dev]	RWQA [test]	MMStar [test]
<i>Autoregressive vision-language models</i>								
Cambrian-1	8B	–	42.7	–	73.0	75.9	64.2	–
LLaVA-OV	7B	7.8M	48.8	–	81.4	80.8	66.3	61.7
LLaMA3-V	8B	16.5M	45.4	14.5	81.1	79.8	66.0	56.5
InternVL2	8B	–	47.9	–	83.7	82.7	64.7	61.2
InternVL2.5	8B	–	53.4	25.3	84.5	84.1	69.4	62.6
Qwen2.5-VL	7B	> 9M	51.1	33.4	83.9	83.5	68.5	63.9
<i>Diffusion vision-language models</i>								
LaViDA-L	8B	1.6M	43.3	–	70.0	70.5	–	–
LaViDA-O	8B	–	45.1	–	76.7	76.4	–	–
Dimple	7B	1.3M	45.2	–	74.4	–	–	–
LLaDA-V	8B	16.5M	48.6	18.6	77.8	82.9	63.2	60.1
DiffusionVL	7B	738K	49.3	25.0	82.2	83.5	68.0	63.2
SDAR-VL	8B	> 16M	53.0	–	79.9	82.2	66.5	59.9
FLUID-V	7B	100k	55.8	35.4	82.3	86.1	71.4	64.1
Model	Size	Data	ChartQA [test]	InfoVQA [val]	DocVQA [val]	OCRBench [test]	MathVista [testmini]	MathVerse [testmini]
<i>Autoregressive vision-language models</i>								
Cambrian-1	8B	–	73.3	–	77.8	62.4	49.0	–
LLaVA-OV	7B	7.8M	80.0	70.7	90.2	–	63.2	26.2
LLaMA3-V	8B	16.5M	77.8	58.9	86.2	–	62.1	29.0
InternVL2	8B	–	82.2	73.0	91.0	–	58.2	24.2
InternVL2.5	8B	–	82.8	75.5	92.0	82.2	63.2	27.2
Qwen2.5-VL	7B	> 9M	60.3	79.3	91.6	80.2	67.3	–
<i>Diffusion vision-language models</i>								
LaViDA-L	8B	1.6M	64.6	34.2	59.0	–	44.8	27.2
LaViDA-O	8B	–	80.0	44.6	73.7	–	56.9	36.9
Dimple	7B	1.3M	63.4	–	–	–	42.3	–
LLaDA-V	8B	16.5M	78.3	66.3	83.9	57.3	59.7	28.5
DiffusionVL	7B	738K	84.2	69.9	90.8	76.2	61.5	–
SDAR-VL	8B	> 16M	82.7	73.2	88.3	59.7	62.5	36.5
FLUID-V	7B	100k	85.3	74.4	91.9	78.8	63.7	31.0

Table 1: Main results across multimodal benchmarks. Data denotes the number of training samples. Bracketed labels indicate evaluation splits or settings. RWQA denotes RealWorldQA. Best results in diffusion methods are **bolded**, and our results are highlighted in .

runs for 32k steps. Stage II freezes the adapted decoder and trains only the horizon head for 4k steps with learning rate 1×10^{-4} . The horizon labels are generated using competence-boundary threshold $\tau = 2.8$, and $K_{\max} = 16$ probes. All experiments are conducted on 4 NVIDIA A40 48GB GPUs.

Evaluation. To comprehensively evaluate FLUID-V’s performance, we considered a diverse suite of multimodal benchmarks: *Knowledge and Reasoning*: MMMU (Yue et al., 2024), MMMU-Pro (Yue et al., 2025), MMStar (Chen et al., 2024), MMBench (Liu et al., 2024a), and MathVista (Lu et al., 2023). *Real-world Perception*: RealWorldQA (xAI, 2024) and AI2D (Kembhavi et al., 2016). *Chart and Document Understanding*: ChartQA (Masry et al., 2022), InfoVQA (Mathew et al., 2022), and DocVQA (Mathew et al.,

2021). *OCR and Text-centric Understanding*: OCRBench (Liu et al., 2024b).

We follow the official evaluation protocol and metric for each benchmark. Unless otherwise specified, inference uses greedy decoding with dynamic KV-cache reuse, maximum horizon $K_{\max} = 16$, and confidence threshold $\eta = 0.85$.

5.2 Main Results

Table 1 provides a contextual comparison across AR and diffusion-style VLMs. Among the compared diffusion-style VLMs, FLUID-V obtains the best reported scores on several benchmarks under a substantially smaller adaptation budget, although these comparisons are not fully controlled due to differences in backbone and training data. These results indicate that adapting an AR-VLM to diffusion-style decoding does not require sacrificing multimodal

generation quality.

Compared with large AR-VLMs, FLUID-V remains competitive despite using substantially fewer adaptation samples. The gains are especially clear on reasoning-oriented and real-world perception benchmarks, suggesting that causal prefix preservation and local refinement are both important for stable multimodal generation. Overall, the results show that FLUID-V provides an effective quality-efficiency trade-off for AR-to-diffusion VLM adaptation.

Components		Benchmarks		
Elastic	Evolving	MMMU	RWQA	OCRBench
✗	✗	54.6	70.2	76.3
✗	✓	55.2	71.0	77.9
✓	✗	54.9	70.8	76.4
✓	✓	55.8	71.4	78.8

Table 2: Component ablation of FLUID-V. Elastic denotes adaptive horizon prediction, and Evolving denotes entropy-gated Top- k intermediate-state updates.

5.3 Ablation Study

We conduct ablations to verify three central design choices of FLUID-V: whether the elastic horizon improves decoding efficiency, whether evolving intermediate states improve refinement quality, and whether the horizon head predicts meaningful local decoding spans.

Intermediate	MMMU	RWQA	OCRBench
Binary mask	53.9	70.1	76.4
Top-1 embedding	54.5	70.4	77.6
Top- k mixture	54.8	70.6	78.3
Top-k + gate	55.8	71.4	78.8

Table 3: Impact of intermediate-state design. Binary mask reinitializes unresolved slots, Top-1 uses the most likely token embedding, Top- k mixture keeps multiple semantic candidates, and Top- k + gate further modulates the update by uncertainty.

5.3.1 Benefit of Elastic Horizon

Table 2 and Figure 4(a) evaluate the effect of elastic horizon prediction. FLUID-V improves average throughput by $2\times$ over LLaDA-V decoding while selecting shorter horizons on OCRBench than on more descriptive benchmarks. In contrast, removing the elastic horizon forces a fixed refinement span, weakening the balance between acceleration and reliability. By expanding in confident continuations and contracting around OCR-, chart-, and

reasoning-sensitive regions, FLUID-V improves throughput without overextending refinement into high-uncertainty segments.

The fixed-horizon comparison in Table 2 further shows that adaptive span selection also improves accuracy, especially on tasks requiring fine-grained grounding. This indicates that the horizon head is not merely an efficiency module, but also a quality-control mechanism that regulates how far local refinement should proceed before commitment.

5.3.2 Impact of Intermediate States

Table 2 verifies the benefit of evolving intermediate states. Without them, unresolved positions are repeatedly reset to static masks or premature token embeddings, limiting the partial visual-language evidence carried across refinement rounds. Evolving states improve performance by maintaining soft semantic representations until predictions become reliable enough to commit.

Table 3 further compares different intermediate-state designs. Binary masking discards accumulated evidence, while Top-1 feedback is vulnerable to transient errors. Top- k mixtures preserve multiple plausible semantics before commitment, and the entropy gate further stabilizes the update by reducing interpolation when the candidate distribution remains diffuse. This design is especially useful in OCR-, chart-, and reasoning-sensitive regions, so we use entropy-gated Top- k mixtures as the default representation.

5.3.3 Verification of Horizon Head Accuracy

Figure 4(b) evaluates whether the horizon head predicts meaningful local decoding spans. We compare the predicted horizon with the target horizon derived from uncertainty-based competence boundaries. FLUID-V achieves low prediction error and high tolerance-based accuracy across benchmarks, indicating that the horizon head is not simply learning an average block length.

Additionally, when the future block contains confident continuations, the model selects a longer horizon; when the block contains visually grounded or reasoning-sensitive tokens, it selects a shorter horizon. This behavior supports the role of the horizon head as an uncertainty-aware controller for dynamic causal inference.

5.4 Controlled Efficiency and Generalization

Table 4 reports controlled backbone evaluations and end-to-end efficiency under the No-Thinking

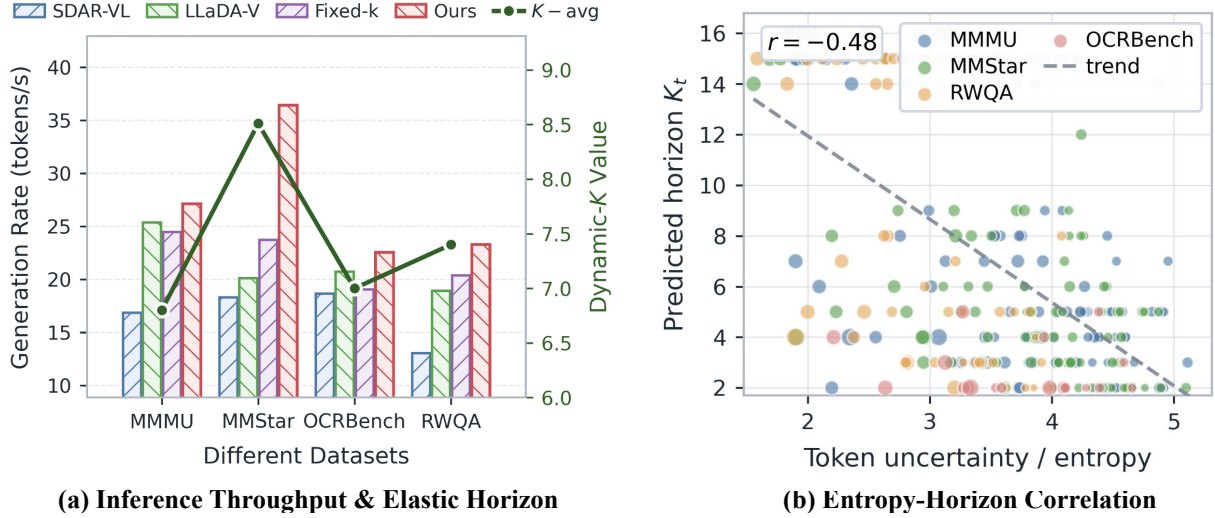


Figure 4: Elastic horizon analysis. (a) Generation throughput and average selected horizon across representative benchmarks. Bars show decoding speed, and the line reports the average predicted horizon. (b) Relationship between token uncertainty and predicted horizon. FLUID-V assigns shorter horizons to high-uncertainty regions and longer horizons to confident local continuations.

Model	Type	Data	MMMU	RWQA	OCRBench	Latency (s) ↓	Throughput (Tok/s) ↑	Memory (GB) ↓
GLM-4.6V-Flash-10B	AR	0	54.1	64.5	81.2	7.5	15.8	20.3
GLM-4.6V-Flash-10B + SFT	AR	100K	51.6	62.8	78.7	7.3	15.5	20.0
OpenPangu-VL-7B	AR	0	64.7	72.4	89.7	3.7	19.9	16.9
OpenPangu-VL-7B + SFT	AR	100K	63.3	71.9	89.3	4.0	18.5	16.9
SDAR-VL-Instruct-8B	Diff.	> 16M	53.0	66.5	59.7	6.9	6.4	22.4
LLaDA-V-8B	Diff.	16.5M	48.6	63.2	57.3	3.6	16.7	21.7
FLUID-V (Qwen2.5-VL-7B)	Diff.	100K	52.6	68.8	73.5	3.0	21.7	16.2
FLUID-V (GLM-4.6V-Flash-10B)	Diff.	100K	44.8	56.6	69.3	6.0	16.8	19.7
FLUID-V (OpenPangu-VL-7B)	Diff.	100K	55.8	71.4	78.8	3.1	23.8	17.1

Table 4: Controlled comparison of multimodal performance, backbone generalization, and end-to-end decoding efficiency. Latency, throughput, and peak GPU memory are averaged over MMMU, RealWorldQA (RWQA), and OCRBench. Data denotes the additional adaptation data used in our experiments; SDAR-VL and LLaDA-V are evaluated from their released checkpoints.

Output length	Tok/s ↑	Memory (GB) ↓
0–256	20.88	17.14
256–512	24.47	16.76
512–1024	24.90	17.08
1024–2048	28.77	16.97

Table 5: Length-bucketed decoding behavior of FLUID-V on MMMU.

setting for fair comparison with existing diffusion VLMs. On OpenPangu-VL-7B, FLUID-V reaches 23.8 Tok/s with 3.1 s latency, versus 19.9 Tok/s and 3.7 s for the AR backbone, and provides $1.42\times$ and $3.74\times$ higher throughput than LLaDA-V and SDAR-VL, respectively. Peak memory remains comparable to the original AR backbone and substantially lower than the evaluated diffusion baselines. This

confirms that uncertainty-aware refinement retains practical parallel decoding gains with only modest memory overhead.

Same-backbone controls further show that ordinary SFT with the same 100K samples does not account for the observed behavior. Beyond OpenPangu-VL-7B, FLUID-V is also instantiated on Qwen2.5-VL-7B and GLM-4.6V-Flash-10B, extending the evaluation across model families and scales. The results on both additional backbones support the transferability of the framework beyond its primary OpenPangu setting. On GLM-4.6V-Flash-10B, FLUID-V reduces latency from 7.5 s to 6.0 s and increases throughput from 15.8 to 16.8 Tok/s, while the larger accuracy drop indicates that capability retention remains backbone-dependent.

This variation further suggests that the efficiency–capability trade-off is influenced by the properties of the underlying AR backbone rather than by the decoding framework alone.

Scaling with response length. As shown in Table 5, FLUID-V throughput on MMMU increases from 20.88 Tok/s for outputs below 256 tokens to 28.77 Tok/s for 1024–2048 tokens, while peak memory remains within 16.76–17.14 GB. This trend is consistent with elastic decoding: longer responses provide more opportunities for multi-token refinement and better amortize fixed decoding costs. It therefore suggests that FLUID-V’s parallelism advantage becomes more pronounced with response length, without positioning the method as a dedicated long-form generation model.

6 Conclusion

In this work, we presented FLUID-V, a lightweight framework for adapting autoregressive vision-language models to diffusion-style local refinement. FLUID-V improves fixed-block decoding by introducing elastic horizon prediction to control refinement spans and evolving Top- k intermediate states to preserve partial semantic evidence before discrete commitment. By replaying only confident predictions into the persistent KV cache, FLUID-V preserves causal prefix semantics and autoregressive compatibility while adapting to the uncertainty structure of multimodal generation. Ultimately, our findings show that effective AR-to-diffusion VLM adaptation requires more than predicting multiple tokens at once; it requires aligning causal consistency with uncertainty-aware refinement flexibility.

Limitations

FLUID-V studies efficient AR-to-diffusion adaptation for three distinct AR-VLM backbones. Its performance and speed may vary with the strength of the source model, the quality of self-distilled adaptation data, and inference implementation details. In addition, the current study focuses primarily on text-side diffusion adaptation for multimodal generation, and extending the framework to broader multimodal settings remains an open direction. Finally, FLUID-V models unresolved states in the language embedding space; richer multimodal intermediate representations may provide further opportunities for improvement.

References

- Marianne Arriola, Aaron Gokaslan, Justin Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sahoo, and Volodymyr Kuleshov. 2025a. Block diffusion: Interpolating between autoregressive and diffusion language models. In *International Conference on Learning Representations*, volume 2025, pages 50726–50753.
- Marianne Arriola, Naveen Venkat, Jonathan Granskog, and Anastasis Germanidis. 2025b. Autoregressive-to-diffusion vision language models. <https://runwayml.com/research/autoregressive-to-diffusion-vlms>.
- Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993.
- Baoyou Chen, Hanchen Xia, Peng Tu, Haojun Shi, Shan Mu, Weihao Yuan, and Siyu Zhu. 2026a. Bard: Bridging autoregressive and diffusion vision-language models via highly efficient progressive block merging and stage-wise distillation. *arXiv preprint arXiv:2604.16514*.
- Hanting Chen, Yasheng Wang, Kai Han, Dong Li, Lin Li, Zhenni Bi, Jinpeng Li, Haoyu Wang, Fei Mi, Mingjian Zhu, and 1 others. 2025. Pangu embedded: An efficient dual-system llm reasoner with metacognition. *arXiv preprint arXiv:2505.22375*.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.
- Zigeng Chen, Gongfan Fang, Xinyin Ma, Ruonan Yu, and Xinchao Wang. 2026b. Dmax: Aggressive parallel decoding for dllms. *arXiv preprint arXiv:2604.08302*.
- Shuang Cheng, Yihan Bian, Dawei Liu, Linfeng Zhang, Qian Yao, Zhongbo Tian, Wenhai Wang, Qipeng Guo, Kai Chen, Biqing Qi, and 1 others. 2025a. Sdar: A synergistic diffusion-autoregression paradigm for scalable sequence generation. *arXiv preprint arXiv:2510.06303*.
- Shuang Cheng, Yuhua Jiang, Zineng Zhou, Dawei Liu, Wang Tao, Linfeng Zhang, Biqing Qi, and Bowen Zhou. 2025b. Sdar-vl: Stable and efficient block-wise diffusion for vision-language understanding. *arXiv preprint arXiv:2512.14068*.
- DeepCtrl. 2024. deepctrl-sft-data. <https://www.modelscope.cn/datasets/deepctrl/deepctrl-sft-data>.
- Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou.

2023. [Enhancing chat language models by scaling high-quality instructional conversations](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore. Association for Computational Linguistics.
- Shansan Gong, Shivam Agarwal, Yizhe Zhang, Jiacheng Ye, Lin Zheng, Mukai Li, Chenxin An, Peilin Zhao, Wei Bi, Jiawei Han, and 1 others. 2025. Scaling diffusion language models via adaptation from autoregressive models. In *International Conference on Learning Representations*, volume 2025, pages 5046–5073.
- Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. 2022. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*.
- Liqi He, Zuchao Li, Xiantao Cai, and Ping Wang. 2024. Multi-modal latent space learning for chain-of-thought reasoning in language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18180–18187.
- Michael Hersche, Samuel Moor-Smith, Thomas Hofmann, and Abbas Rahimi. 2025. Soft-masked diffusion language models. *arXiv preprint arXiv:2510.17206*.
- Keya Hu, Linlu Qiu, Yiyang Lu, Hanhong Zhao, Tianhong Li, Yoon Kim, Jacob Andreas, and Kaiming He. 2026. Elf: Embedded language flows. *arXiv preprint arXiv:2605.10938*.
- Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer.
- Siqi Kou, Jiachun Jin, Zhihong Liu, Chang Liu, Ye Ma, Jian Jia, Quan Chen, Peng Jiang, and Zhijie Deng. 2024. Orthus: Autoregressive interleaved image-text generation with modality-specific heads. *arXiv preprint arXiv:2412.00127*.
- Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024. [Llava-next: Stronger llms supercharge multimodal capabilities in the wild](#).
- Jijie Li, Li Du, Hanyu Zhao, Bo wen Zhang, Liangdong Wang, Boyan Gao, Guang Liu, and Yonghua Lin. 2025a. [Infinity instruct: Scaling instruction selection and synthesis to enhance language models](#). *Preprint*, arXiv:2506.11116.
- Shufan Li, Konstantinos Kallidromitis, Hritik Bansal, Akash Gokul, Yusuke Kato, Kazuki Kozuka, Jason Kuen, Zhe Lin, Kai-Wei Chang, and Aditya Grover. 2026. Lavidia: A large diffusion language model for multimodal understanding. *Advances in Neural Information Processing Systems*, 38:105101–105134.
- Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35:4328–4343.
- Zuchao Li, Yonghua Hei, Qiwei Li, Lefei Zhang, Ping Wang, Baoyuan Qi, Liu Guoming, and 1 others. 2025b. What limits bidirectional model’s generative capabilities? a uni-bi-directional mixture-of-expert method for bidirectional fine-tuning. In *Forty-second International Conference on Machine Learning*.
- Yangzhou Liu, Yue Cao, Hao Li, Gen Luo, Zhe Chen, Weiyun Wang, Xiaobo Liang, Biqing Qi, Lijun Wu, Changyao Tian, and 1 others. 2025. Sequential diffusion language models. *arXiv preprint arXiv:2509.24007*.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024a. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.
- Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024b. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *CoRR*, abs/2310.02255, 5:15.
- Xiangyu Ma, Teng Xiao, Zuchao Li, and Lefei Zhang. 2026. From ar to diffusion: Efficiently adapting large language models with strictly causal and elastic horizons. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20914–20927.
- Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, and 1 others. 2025. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7739–7751.
- Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279.
- Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

- Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.
- Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan LI. 2025. *Large language diffusion models*. In *Advances in Neural Information Processing Systems*, volume 38, pages 50608–50646. Curran Associates, Inc.
- Huawei OpenPangu Team. 2025. Openpangu-vl-7b: A multi-modal large language model designed and optimized for ascend npus. <https://ai.gitcode.com/ascend-tribe/openPangu-VL-7B>.
- Tianshuo Peng, Zuchao Li, Lefei Zhang, Hai Zhao, Ping Wang, and Bo Du. 2024. Multi-modal auto-regressive modeling via visual tokens. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10735–10744.
- Subham S Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. 2024. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184.
- Luohe Shi, Zuchao Li, Lefei Zhang, Baoyuan Qi, Guoming Liu, and Hai Zhao. 2026. Scaling llm speculative decoding: Non-autoregressive forecasting in large-batch scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 32947–32955.
- Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Xiangyang Liu, Hang Yan, Yunfan Shao, Qiong Tang, Shiduo Zhang, and 1 others. 2024. Moss: An open conversational large language model. *Machine Intelligence Research*, 21(5):888–905.
- Shengbang Tong, David Fan, Jiachen Li, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. 2025. Metamorph: Multimodal understanding and generation via instruction tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17001–17012.
- Chengyue Wu, Shiyi Lan, Yonggan Fu, Sensen Gao, Jin Wang, Jincheng Yu, Jose M Alvarez, Pavlo Molchanov, Ping Luo, Song Han, and 1 others. 2026. Fast-dvlm: Efficient block-diffusion vlm via direct conversion from autoregressive vlm. *arXiv preprint arXiv:2604.06832*.
- Chengyue Wu, Hao Zhang, Shuchen Xue, Shizhe Diao, Yonggan Fu, Zhijian Liu, Pavlo Molchanov, Ping Luo, Song Han, and Enze Xie. 2025. Fast-dllm v2: Efficient block-diffusion llm. *arXiv preprint arXiv:2509.26328*.
- Tong Wu, Zhihao Fan, Xiao Liu, Hai-Tao Zheng, Yeyun Gong, Jian Jiao, Juntao Li, Jian Guo, Nan Duan, Weizhu Chen, and 1 others. 2023. Ar-diffusion: Auto-regressive diffusion model for text generation. *Advances in Neural Information Processing Systems*, 36:39957–39974.
- xAI. 2024. Grok-1.5 vision preview: Connecting the digital and physical worlds with our first multimodal model. <https://x.ai/news/grok-1.5v>.
- Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2025. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pages 28240–28264.
- Ling Yang, Ye Tian, Bowen Li, Xincheng Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. 2026. Mmada: Multimodal large diffusion language models. *Advances in Neural Information Processing Systems*, 38:138867–138907.
- Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*.
- Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. 2025. Llada-v: Large language diffusion models with visual instruction tuning. *arXiv preprint arXiv:2505.16933*.
- Runpeng Yu, Xinyin Ma, and Xinchao Wang. 2025. Dimple: Discrete diffusion multimodal large language model with parallel decoding. *arXiv preprint arXiv:2505.16990*.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, and 1 others. 2025. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186.

A Empirical Findings on Intermediate States

This appendix provides trajectory-level evidence for the evolving intermediate states used in FLUID-V. The main paper argues that unresolved future positions should not be treated as binary mask-or-token variables; here we make this claim empirical by tracing refinement trajectories on MMMU.

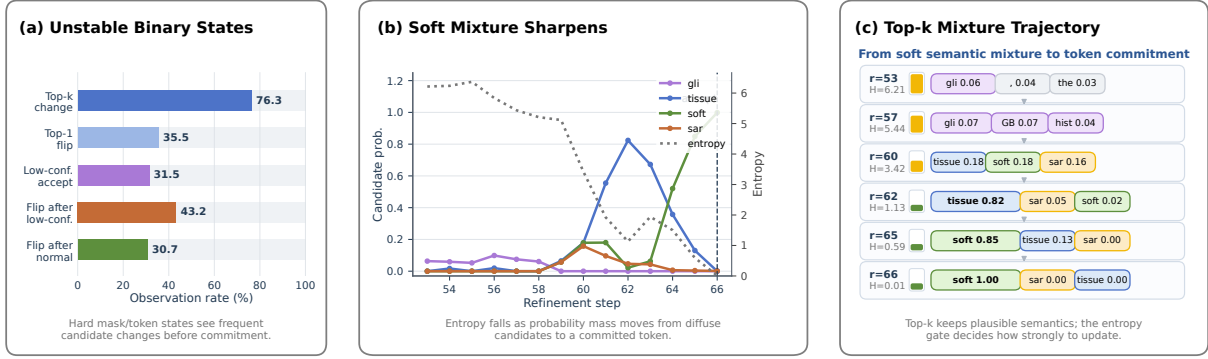


Figure 5: Empirical findings behind evolving intermediate states. (a) Across 50 MMMU traces, unresolved slots frequently change their Top- k set and Top-1 prediction; low-confidence acceptance also increases the next-step flip rate. (b) A representative slot evolves from diffuse histology-related candidates to a sharp final token. (c) Selected Top- k snapshots show that the distribution becomes progressively less ambiguous before commitment.

A.1 Setup.

We sample 50 MMMU validation traces and record statistics over all unresolved slot-round observations, using the same Top- k , confidence threshold, and entropy gate as in the main experiments. The analysis focuses on MMMU because it contains visually grounded and reasoning-sensitive continuations where premature commitment is especially harmful.

A.2 Instability of Hard Binary Decisions

Figure 5(a) shows that unresolved positions are not stable placeholders. The Top- k candidate set changes in 76.3% of slot observations, and the Top-1 prediction flips in 35.5%. Moreover, slots accepted under low confidence are followed by a higher next-step Top-1 flip rate (43.2%) than normal accepted slots (30.7%). This suggests that early hard decisions can become a source of later refinement instability, especially for OCR-, chart-, and reasoning-sensitive spans. A hard-token state therefore risks polluting the causal prefix before the local semantics have stabilized.

A.3 Top- k Candidates Contain Useful Partial Evidence

The candidate distribution is nevertheless informative. In the 49 multiple-choice traces (one trace is excluded because it does not have a valid letter-form answer), the gold answer letter enters the Top- k set in 47 cases. Among 29 wrong cases, the gold letter still enters Top- k in 26 cases and becomes Top-1 at least once in 14 cases. Thus, many failures are not caused by a complete absence of the correct semantic evidence, but by premature ranking or commitment. A binary mask discards this partial

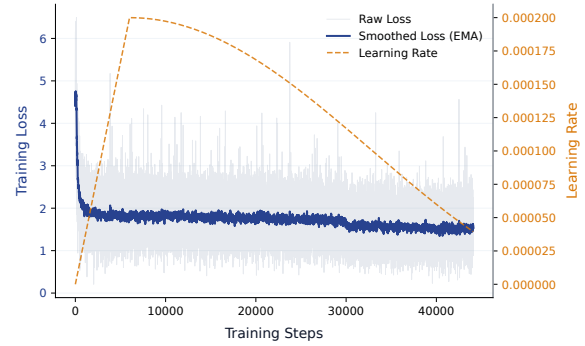


Figure 6: Training dynamics of FLUID-V during Stage I. Raw loss is noisy at the step level, while the exponential moving average decreases rapidly after warmup and remains stable as the learning rate decays.

evidence, while Top-1 feedback can over-amplify a transient wrong token.

A.4 Soft Semantic States Sharpen before Commitment

Figures 5(b–c) illustrate the desired behavior. In a histology question, an unresolved slot first mixes candidates such as “gli”, “GB”, “tissue”, “soft”, and “sar”. These candidates are semantically related but not yet reliable at the token level. Over refinement rounds, entropy decreases and the probability mass concentrates on the final token. FLUID-V therefore represents unresolved positions as entropy-gated Top- k embedding mixtures: confident slots move toward token semantics, while uncertain slots retain a softer state until commitment becomes reliable.

B Training Dynamics and Convergence

Figure 6 reports the Stage I training loss together with the learning-rate schedule. The curve shows a

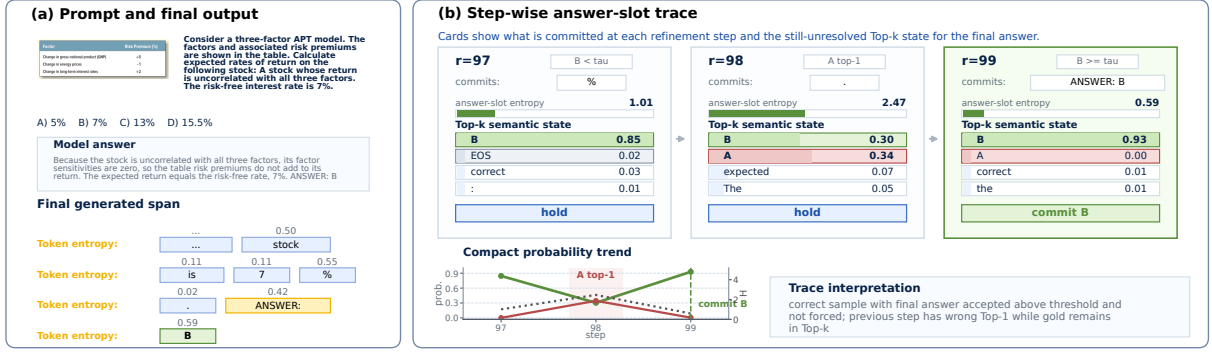


Figure 7: Trace-level case study on a correct MMMU validation sample. The final answer slot first favors the gold option B below the confidence threshold, then temporarily assigns the highest probability to the wrong option A while B remains in the Top- k state, and finally commits B with confidence above the threshold.

short adaptation transient followed by a stable downward trend, indicating that the pretrained AR-VLM can absorb the causal local-refinement objective without a prolonged destabilization phase. This behavior is consistent with FLUID-V’s design: the model is not trained as a diffusion-native VLM from scratch, but is adapted from a causal backbone while preserving clean multimodal conditioning and prefix-aligned supervision.

The raw loss remains visibly noisy because Stage I samples local refinement contexts with different horizons, restoration patterns, and uncertainty levels. The smoothed loss, however, decreases steadily after the initial warmup and continues to improve as the learning rate decays. This supports the use of a compact adaptation budget and suggests that the direct and recurrent training views provide compatible supervision rather than competing objectives.

C Additional Experimental Details

C.1 Loss Weight Ablation

Table 6 studies the recurrent embedding loss weight λ_{emb} . A small value under-supervises the intermediate-state space, making unresolved slots rely mainly on token-level cross entropy. A very large value over-constrains the soft state and weakens the discrete prediction objective. The best result is obtained at $\lambda_{\text{emb}} = 0.15$, which provides a stable balance between token prediction and embedding-space refinement.

C.2 Training Cost Analysis

Table 7 summarizes the additional adaptation cost. Stage I dominates the cost because it updates the LoRA adapter and trains the evolving intermediate states. Stage II is lightweight: the decoder is

λ_{emb}	MMMU	RWQA	OCRBench
0.05	54.2	70.5	76.8
0.10	54.8	71.0	77.5
0.15	55.8	71.4	78.8
0.20	55.0	71.0	77.5
0.25	55.4	70.9	77.7

Table 6: Ablation on the recurrent embedding weight λ_{emb} .

Stage	Trained Module	Hardware	Est. Time
Stage I	LoRA Adapter	4 × 48GB GPUs	~30 hours
Stage II	Horizon Head	4 × 48GB GPUs	negligible
Total	—	4 × 48GB GPUs	~120 GPU-hours

Table 7: Computational overhead of FLUID-V adaptation. The total cost is approximately 120 GPU-hours, with Stage I accounting for nearly all of the training time.

frozen and only the horizon head is trained, so its cost is negligible relative to Stage I. Overall, FLUID-V requires about 120 GPU-hours, making the adaptation substantially cheaper than training a diffusion VLM from scratch.

D Case Study

To complement the aggregate ablations, we inspect MMMU validation traces and record token-level entropy, Top- k candidates, accepted tokens, confidence-gate decisions, and refinement steps. We select a correct example whose final answer slot undergoes clear option-level competition before commitment. This trace is not intended as a standalone performance claim; rather, it illustrates the mechanism behind the aggregate improvements of evolving intermediate states.

As shown in Figure 7, the model eventually produces the correct answer B , but the answer

slot is not resolved monotonically. At refinement step 97, B is already the Top-1 candidate with probability 0.85. However, it remains below the case-study acceptance threshold $\eta_{\text{case}} = 0.9$, and the slot is therefore kept unresolved. At step 98, the wrong option A temporarily becomes Top-1 with probability 0.34, while B remains in the Top- k state with probability 0.30. At step 99, the distribution sharpens again: B returns to Top-1 with probability 0.93 and is accepted without forced low-confidence fallback.

This trajectory illustrates why unresolved positions should not be reduced too early to a binary mask-or-token state. A hard Top-1 feedback state would pass the transient A preference from step 98 into the next refinement context, potentially amplifying an unstable option-level prediction. In contrast, the soft Top- k state keeps B available during the uncertain transition, while confidence-gated commitment prevents the slot from entering the persistent causal prefix until the distribution becomes sufficiently sharp. The case therefore provides a trace-level example of FLUID-V’s main design claim: local refinement benefits from preserving partial semantic evidence in unresolved states and separating temporary refinement from causal commitment.