# ECHO: Efficient Chest X-ray Report Generation with One-step Block Diffusion

Lifeng Chen[1,2*], Tianqi You[1,3*], Hao Liu[1†‡], Zhimin Bao[1], Jile Jiao[1], Xiao Han[1], Zhicai Ou[1], Tao Sun[1], Xiaofeng Mou[1], Xiaojie Jin[2], and Yi Xu[1†]

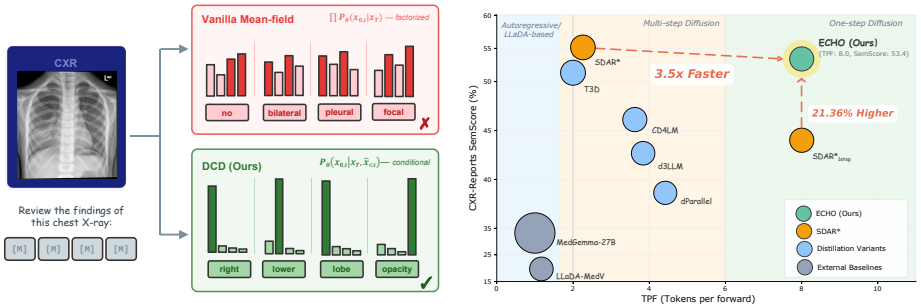
[1] AIRC, Midea Group
[2] Beijing Jiaotong University
[3] Dalian University of Technology
[*]Equal Contribution, [†]Corresponding Author, [‡]Project Leader.

**Abstract.** Chest X-ray report generation (CXR-RG) has the potential to substantially alleviate radiologists' workload. However, conventional autoregressive vision–language models (VLMs) suffer from high inference latency due to sequential token decoding. Diffusion-based models offer a promising alternative through parallel generation, but they still require multiple denoising iterations. Compressing multi-step denoising to a single step could further reduce latency, but often degrades textual coherence due to the mean-field bias introduced by token-factorized denoisers. To address this challenge, we propose **ECHO**, an efficient diffusion-based VLM (dVLM) for chest X-ray report generation. ECHO enables stable one-step-per-block inference via a novel Direct Conditional Distillation (DCD) framework, which mitigates the mean-field limitation by constructing unfactorized supervision from on-policy diffusion trajectories to encode joint token dependencies. In addition, we introduce a Response-Asymmetric Diffusion (RAD) training strategy that further improves training efficiency while maintaining model effectiveness. Extensive experiments demonstrate that ECHO surpasses state-of-the-art autoregressive methods, improving RaTE and SemScore by **64.33%** and **60.58%**, respectively, while achieving up to **an** 8× inference speedup with negligible degradation in clinical accuracy. Code, data, and models are publicly available at `https://echo-midea-airc.github.io/`.

**Keywords:** Chest X-ray Report Generation · One-step Block Diffusion · Direct Conditional Distillation

## 1 Introduction

In recent years, Vision-Language Models (VLMs) [3, 21, 26, 28, 33, 44, 48], where visual features are aligned with language instructions to enable complex cross-modal understanding, have demonstrated significant progress in many fields, notably medical image analysis [6, 23, 27, 36, 39, 41, 46, 54, 60]. Within this field, automated chest X-ray report generation (CXR-RG) [7, 25, 37, 45, 50, 55] has emerged as a critical application. As one of the most common clinical imaging




(a) Illustration of the mean-field bias and our DCD

(b) TPF vs. SemScore across methods

**Fig. 1: Motivation and performance of ECHO. (a)** Decoding all tokens simultaneously in one step produces incoherent outputs, as standard diffusion models predict each position independently. Our Direct Conditional Distillation (DCD) distills from a non-factorized target, yielding coherent one-step-per-block outputs. **(b)** Compared to both autoregressive and diffusion-based baselines, `ECHO` achieves a favorable trade-off between generation quality (SemScore) and decoding throughput (tokens per forward pass). Here, SDAR* denotes SDAR [9] implemented under the multimodal large language model setting.

exams, CXR's high volume places a heavy diagnostic burden on radiologists, creating strong demand for high-throughput automated reporting systems to ease workloads. Despite the promising performance achieved, canonical autoregressive (AR) VLMs often suffer from high inference latency due to their sequential decoding mechanism, which becomes a primary bottleneck for producing reports rapidly and at scale. Fortunately, thanks to parallel decoding capabilities, emerging diffusion-based Vision Language Models (dVLMs) [9,30,58,59] offer a highly promising route toward achieving fast report generation.

Despite the promise of parallel decoding, dVLMs in practice necessitate multiple denoising steps to ensure output coherence. This requirement stems from the mean-field approximation underlying their token-factorized denoisers, which introduces a structural bias that scales monotonically with the noise ratio. Although compressing decoding into a single step theoretically maximizes throughput, it compels the model to predict all tokens simultaneously from a fully masked input, precisely the scenario where mean-field bias is most acute. As illustrated in Fig. 1, the resulting inter-token incoherence leads to substantial quality degradation, prompting the critical question: ***Can we achieve the upper bound of decoding speed without compromising output fidelity?***

To address this challenge, distillation emerges as a natural solution, seeking to transfer the quality of multi-step decoding into a student model that operates in fewer passes. Specifically, several dLLMs [8, 12, 31, 38, 61] implement this via self-distillation, which aligns token-wise predictions with teacher's ones from less corrupted states. However, these teacher targets remain factorized across positions, ignoring dependencies between concurrently predicted tokens. For few-step inference, this factorization remains tolerable, as progressive un-

masking provides the inter-token context that token-independent targets fail to capture. In contrast, one-step decoding lacks such a corrective mechanism, allowing the full mean-field bias to resurface unchecked. Therefore, enabling reliable one-step decoding requires a fundamentally different training objective, one that is unfactorized and directly encodes the joint dependencies among tokens predicted in parallel. To achieve this goal, we propose a new Direct Conditional Distillation (DCD), which constructs supervision from the teacher's on-policy remasking trajectory by conditioning each target on committed high-confidence context. The constructed supervision can encode inter-token dependencies into the training signal and enable stable one-step-per-block parallel inference.

Building upon DCD, we present **ECHO** (**E**fficient **Ch**est X-ray Report Generation with **O**ne-step Block Diffusion), a new foundational dVLM for CXR report generation that achieves both clinical accuracy and one-step inference efficiency. More concretely, we first train an enhanced AR-based CXR-RG VLM using curated data enriched with finding-level content and symptom descriptions. Based on this, we propose Response-Asymmetric Diffusion (RAD) adaptation to efficiently convert the AR model into a block diffusion decoding paradigm. Subsequently, we apply DCD to distill this multi-step model into a one-step counterpart.

Notably, compared with state-of-the-art autoregressive methods, `ECHO` improves RaTE and SemScore by 64.33% and 60.58%, respectively, while achieving up to an 8× theoretical speedup (5.1× in practice). Our main contributions are summarized below:

- We present **ECHO**, a novel dVLM for CXR report generation that delivers strong clinical accuracy through one-step-per-block parallel decoding, outperforming both autoregressive and diffusion-based models by a large margin across standard benchmarks.
- We propose a novel **Direct Conditional Distillation (DCD)**, to the best of our knowledge the first one-step distillation framework for discrete diffusion language models, achieving up to a 390% inference speedup with only marginal quality degradation.
- We design **Response-Asymmetric Diffusion (RAD)** adaptation, which further reduces theoretical training FLOPs by 72.3%, translating to a 3.61× speedup in training efficiency.

## 2 Related Works

### 2.1 Multimodal Medical Foundation Models

Large-scale vision-language models (VLMs) [3, 21, 26, 28, 33, 44, 48] have established strong multimodal understanding by aligning visual representations with language instructions, achieving remarkable performance across diverse visual reasoning tasks. Motivated by this progress, researchers have adapted VLMs to the medical domain [6, 14, 15, 23, 27, 36, 39, 41, 46, 54, 60, 64–66], where precise alignment between clinical images and textual descriptions is essential. Within

medical imaging, automated CXR report generation has emerged as a central task [7,25,37,45,50,55], where existing methods predominantly follow an autoregressive paradigm and achieve strong clinical accuracy but at the cost of sequential decoding throughput, motivating the need for faster generation paradigms.

### 2.2 Diffusion Language Model

Discrete diffusion language models (dLLMs) represent one such paradigm, enabling parallel token prediction in place of sequential decoding. Specifically, dLLMs [2, 4, 35, 42, 56] define a forward process that progressively masks tokens and a learned reverse process that recovers them over a discrete token vocabulary. Building on this foundation, dVLMs [9, 30, 58, 59] extend the framework to vision-language tasks through a two-stage process of visual encoder alignment followed by instruction fine-tuning, enabling image-conditioned parallel generation. Block diffusion [1, 16, 19] further refines these models by adopting a semi-autoregressive decoding scheme that generates outputs block by block in causal order, naturally supporting variable-length sequence generation. A further line of work reduces this cost by adapting pretrained autoregressive models into block diffusion models [4, 5, 9, 17, 51], offering a practical and scalable alternative. Collectively, these advances have improved both the training scalability and inference throughput of dLLMs and dVLMs.

### 2.3 Acceleration of dLLMs

As dLLMs and dVLMs continue to mature, further accelerating their inference has attracted growing research attention. This has been pursued from two complementary directions: inference-time optimization techniques [10, 20, 29, 34, 49, 52] that reduce per-step or per-token computation overhead, and training-time distillation [8, 12, 31, 38, 61] that compresses the multi-step denoising process into fewer steps. Along the distillation direction, methods such as SDTT [12] and dParallel [8] progressively align predictions at higher noise levels with those at lower ones, using either cross-entropy loss on self-generated trajectories or KL divergence between predicted distributions. Most recently, T3D [61] takes a distinct angle, employing a DPO-style optimization objective to directly penalize mean-field bias during training rather than matching noise-level predictions. Despite these advances, all existing methods retain token-factorized prediction targets and therefore still require multiple denoising steps to produce coherent outputs, leaving dLLM and dVLM throughput far below its theoretical upper bound. `ECHO` addresses this gap with Direct Conditional Distillation (DCD), a distillation framework for dLLMs and dVLMs that supports one-step denoising, applied here to block diffusion.

## 3 Preliminaries

To motivate our approach, we formalize the dLLM framework and analyze how its token-factorized parameterization gives rise to the parallelism bottleneck

identified in Section 1. Let $\mathcal{V}$ denote the vocabulary and $L$ the sequence length. A token sequence is $\mathbf{x}_0 = (x_{0,1}, \ldots, x_{0,L}) \in \mathcal{V}^L$. Discrete diffusion language models (dLLMs) [40, 42, 56] define a forward process $q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)$ that progressively masks tokens, and a reverse process $p_\theta(\mathbf{x}_{0:T-1} \mid \mathbf{x}_T)$ that recovers them.

**Mean-field parameterization.** Since the joint posterior $p(\mathbf{x}_0 \mid \mathbf{x}_t, t)$ over $\mathcal{V}^L$ requires exponentially many parameters, existing dLLMs [1, 35, 40] universally adopt a token-factorized (mean-field) approximation [2, 53, 57, 63]:

$$p_\theta(\mathbf{x}_0 \mid \mathbf{x}_t, t) \;=\; \prod_{i=1}^{L} p_\theta(x_{0,i} \mid \mathbf{x}_t, t), \tag{1}$$

trained by minimizing the per-token denoising objective:

$$\mathcal{L}_{\mathrm{MF}}(\theta) \;=\; \mathbb{E}_{\mathbf{x}_0,\, t,\, \mathbf{x}_t}\left[ -\sum_{i=1}^{L} \log\, p_\theta(x_{0,i} \mid \mathbf{x}_t, t) \right]. \tag{2}$$

As Eq. (2) decomposes into $L$ independent per-position cross-entropies, the global optimum is attained when each factor independently matches the true conditional marginal, i.e. $p^\star_\theta(x_{0,i} \mid \mathbf{x}_t, t) = q(x_{0,i} \mid \mathbf{x}_t, t)$. The resulting optimal factorized joint $p^\star_{\mathrm{MF}} = \prod_i q(x_{0,i} \mid \mathbf{x}_t, t)$ aligns each position's marginal with the true posterior, but cannot capture the cross-positional correlations present in $q(\mathbf{x}_0 \mid \mathbf{x}_t, t)$. We measure this gap by the mean-field bias:

$$\epsilon_{\mathrm{MF}}(\mathbf{x}_t, t) \;\triangleq\; \mathrm{KL}\big(q(\mathbf{x}_0 \mid \mathbf{x}_t, t) \,\big\|\, p^\star_{\mathrm{MF}}(\mathbf{x}_0 \mid \mathbf{x}_t, t)\big), \tag{3}$$

which quantifies the irreducible joint dependence structure across token positions that no factorized distribution can represent.

**Why multi-step sampling mitigates the bias.** The expected mean-field bias $\bar{\epsilon}_{\mathrm{MF}}(t) \triangleq \mathbb{E}_{\mathbf{x}_t}[\epsilon_{\mathrm{MF}}(\mathbf{x}_t, t)]$ is governed by the corruption level: when $\mathbf{x}_t$ has few masked tokens, the remaining unknowns are sparsely distributed and largely conditionally independent, so $\bar{\epsilon}_{\mathrm{MF}}(t)$ is small; when all tokens are masked ($t = T$), the posterior exhibits strong cross-positional dependence and $\bar{\epsilon}_{\mathrm{MF}}(T)$ reaches its maximum. In general, $\bar{\epsilon}_{\mathrm{MF}}(t)$ grows monotonically with the number of masked positions. [57] Multi-step reverse sampling systematically exploits this monotonicity. At each reverse step $s$, tokens decoded with high confidence from $p_\theta(\cdot \mid \mathbf{x}_s, s)$ are committed as observed context, reducing the number of masked positions from $m_s$ to $m_{s-1} < m_s$. Consequently, the model at step $s-1$ operates under a strictly reduced corruption level and, accordingly, a lower expected bias $\bar{\epsilon}_{\mathrm{MF}}(s-1) \le \bar{\epsilon}_{\mathrm{MF}}(s)$. By iterating this process, the multi-step chain progressively shifts decoding from the high-bias fully-masked regime toward a low-bias, nearly-observed state, thereby recovering the inter-token dependence structure that Eq. (1) otherwise discards.

## 4 Methodology

### 4.1 Overview

As illustrated in Fig. 2, the training pipeline of `ECHO` comprises three successive stages. In Stage 1, we perform continued pre-training (CPT) on Lingshu-7B [54]

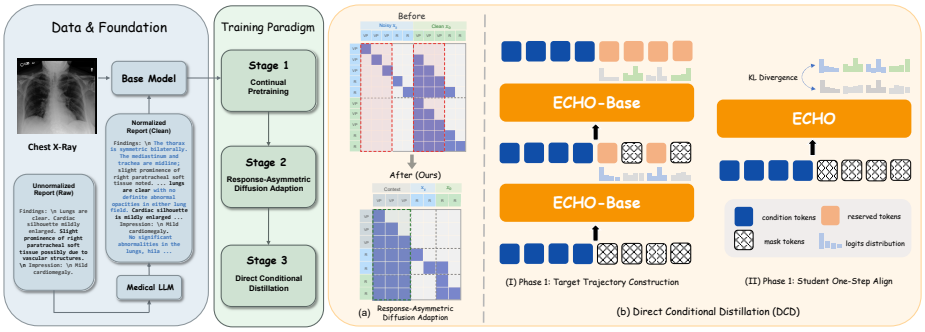



**Fig. 2: Overview of the ECHO training pipeline.** ECHO is built in three successive stages: continued pre-training (Stage 1) produces $\texttt{ECHO}_{\texttt{AR}}$; Response-Asymmetric Diffusion adaptation (Stage 2) converts it into the block diffusion model $\texttt{ECHO}_{\texttt{Base}}$; and Direct Conditional Distillation (Stage 3) distills $\texttt{ECHO}_{\texttt{Base}}$ into the final one-step-per-block model ECHO. **(a)** RAD duplicates only the response portion of the training sequence, avoiding the redundant duplication of long vision token contexts required by prior two-stage conversion methods. **(b)** DCD proceeds in two phases: the teacher's confidence-heuristic remasking trajectory is collected to form a joint, non-factorized supervision target, which is then used to align the student's single-step prediction via KL divergence.

using a curated CXR corpus. This produces $\texttt{ECHO}_{\texttt{AR}}$, an autoregressive vision-language model specialized in radiology report generation. Subsequently, we in Stage 2 propose Response-Asymmetric Diffusion (RAD) adaptation, which converts $\texttt{ECHO}_{\texttt{AR}}$ into $\texttt{ECHO}_{\texttt{Base}}$, a block diffusion decoding model that retains the domain knowledge of $\texttt{ECHO}_{\texttt{AR}}$, achieving an initial inference speedup. Lastly, in Stage 3, to further push decoding speed toward its theoretical upper bound, we apply Direct Conditional Distillation (DCD) to train $\texttt{ECHO}_{\texttt{Base}}$ toward one-step-per-block decoding, yielding the final ECHO model.

### 4.2 From ECHO-AR to ECHO

**Response-Asymmetric Diffusion Adaptation.** While $\texttt{ECHO}_{\texttt{AR}}$ achieves high-quality report generation, its token-by-token decoding mechanism limits inference throughput. To address this, our first objective is to convert $\texttt{ECHO}_{\texttt{AR}}$ into $\texttt{ECHO}_{\texttt{Base}}$, a block diffusion model that preserves the domain knowledge acquired during pretraining. Prior work [9] has implemented this through a two-stage adaptation, i.e., a pretraining phase followed by SFT. In these methods, the entire training sequence, covering vision, instruction, and response tokens, is duplicated across both stages to construct block-causal teacher-forcing targets. For CXR report generation, where vision token sequences are substantially long, this full-sequence duplication incurs prohibitive training costs. To overcome this inefficiency, we propose Response-Asymmetric Diffusion (RAD), which achieves this adaptation through a single SFT stage. As illustrated in Fig. 2, RAD duplicates only the response portion of the training sequence, and constructs a

block attention mask such that each noisy response block attends to all vision and instruction tokens as well as all previously decoded blocks. As illustrated in Fig. 3(b), this asymmetric design eliminates the redundant duplication of long vision token contexts, substantially reducing training FLOPs while consolidating the originally two-stage conversion into one. Furthermore, we conduct a series of experiments (Sec. 5.4) to investigate the relationship between training data volume, model quality, and inference speed during the RAD conversion. We find that $\texttt{ECHO}_{\texttt{Base}}$ attains performance comparable to $\texttt{ECHO}_{\texttt{AR}}$ using only a small fraction of the original training data, demonstrating the high knowledge-transfer efficiency of RAD.

**Direct Conditional Distillation.** With $\texttt{ECHO}_{\texttt{Base}}$ established, our next objective is to convert it into a model capable of one-step-per-block decoding, fully maximizing inference throughput. However, as discussed in Sec. 3, discrete diffusion language models inherently rely on multi-step decoding to progressively reduce the mean-field bias, and our dVLM setting is no exception. This calls for a training target that is itself non-factorized: by aligning the student to such a target, the inter-token dependencies accumulated across the multi-step denoising trajectory can be captured in a single forward pass. To this end, we propose Direct Conditional Distillation (DCD), which constructs this non-factorized target by collecting and stitching the high-confidence token distributions at each step of the teacher's multi-step denoising process, forming a joint supervision signal that encodes cross-token dependencies.

The full self-distillation procedure is detailed in Algorithm 1 and consists of the following two phases, applied iteratively during training.

**Phase 1: On-policy Teacher Trajectory Collection.** The teacher trajectory is the trajectory sampled by its multi-step denoising process. To obtain this, we first run $\texttt{ECHO}_{\texttt{Base}}$ through a confidence-heuristic denoising process, where tokens are progressively unmasked in order of their prediction confidence $c_i \triangleq \max_v p_\theta(x_i{=}v \mid \mathbf{x}_{\text{curr}})$. Specifically, at each step of a block's denoising, we record the predicted distribution and the committed token label for every position as it is unmasked. The collected token labels form the pseudo labels $\hat{\mathbf{b}}_n$, which are used to construct the training sequence for the subsequent student alignment phase. The collected distributions are then stitched together into the joint target $\mathcal{P}_{\text{tch}}^{(n)}$ for the $n$-th block, as the distillation target in Phase 2.

**Phase 2: Student One-Step Alignment.** With the pseudo-labels and stitched joint distribution collected in Phase 1, the next step is to align this distribution with the student's one-step prediction. Concretely, we first use the pseudo-labels to construct a block teacher-forcing training sequence following the RAD scheme (Sec. 4.2), ensuring that the student makes its prediction under exactly the same conditions as the teacher. Letting $b_{n,i}$ denote the token at position $i$ of block $n$ and $\mathcal{Q}_\phi^{(n)} \triangleq \left\{p_\phi(b_{n,i} \mid \mathbf{x}_{\text{train}})\right\}_{i=1}^{L}$ the student's one-step predicted distribution over that block, we minimize the forward Kullback–Leibler (KL) divergence $D_{\text{KL}}\left(\mathcal{P}_{\text{tch}}^{(n)} \,\|\, \mathcal{Q}_\phi^{(n)}\right)$ for each block and aggregate the loss over all blocks. Intuitively, a token that is unmasked at a later step within a block requires stronger conditioning to reach a high-confidence prediction, indicating that its position is

**Algorithm 1:** Direct Conditional Distillation (DCD)

**Input :** Context $\mathbf{x}_{\text{ctx}}$, teacher $\theta$, student $\phi$, $N$ blocks, block length $L$, sampling confidence threshold $\tau$

**Output:** Updated student $\phi$

`// Phase 1: On-policy Teacher Trajectory Collection`

**for** $n = 1, \ldots, N$ **do**
  Initialize $\mathcal{B} \leftarrow \{1, \ldots, L\}$,;
  $\mathbf{x}_{\text{curr}} \leftarrow (\mathbf{x}_{\text{ctx}}, \hat{\mathbf{b}}_{1:n-1})$;
  **while** $\mathcal{B} \neq \emptyset$ **do**
    Query teacher $p_\theta(x_i \mid \mathbf{x}_{\text{curr}})$ for all $i \in \mathcal{B}$;
    $\mathcal{U} \leftarrow \{i \in \mathcal{B} \mid c_i \geq \tau\}$;;
    **if** $\mathcal{U} = \emptyset$ **then** $\mathcal{U} \leftarrow \{\arg\max_{i \in \mathcal{B}} c_i\}$;
    $\mathcal{P}^{(n)}_{\text{tch}}[i] \leftarrow p_\theta(x_i \mid \mathbf{x}_{\text{curr}})$,;
    $\hat{x}_i \leftarrow \arg\max\, p_\theta(x_i \mid \mathbf{x}_{\text{curr}})$ for all $i \in \mathcal{U}$;
    $\mathbf{x}_{\text{curr}} \leftarrow \mathbf{x}_{\text{curr}} \cup \{\hat{x}_i\}_{i \in \mathcal{U}}$,;
    $\mathcal{B} \leftarrow \mathcal{B} \setminus \mathcal{U}$;
  $\hat{\mathbf{b}}_n \leftarrow (\hat{x}_1, \ldots, \hat{x}_L)$;

`// Phase 2: Student One-Step Alignment (cf. Sec. 4.2 and Fig. 2)`

Construct RAD training sequence $\mathbf{x}_{\text{train}}$ using $(\mathbf{x}_{\text{ctx}}, \hat{\mathbf{b}}_{1:N})$;

$\mathcal{L}_{\text{DCD}} \leftarrow \sum_{n,i} D_{\text{KL}}\big(\mathcal{P}^{(n)}_{\text{tch}}[i] \,\big\|\, p_\phi(b_{n,i} \mid \mathbf{x}_{\text{train}})\big)$;

Update $\phi \leftarrow \text{Optimizer}(\phi, \nabla_\phi \mathcal{L}_{\text{DCD}})$;

subject to more severe mean-field bias. Therefore, we introduce a token reweighting scheme that assigns each position a weight proportional to the step at which it is unmasked. Under this scheme, positions with larger mean-field bias receive stronger supervision, while tokens committed at early steps serve as a regularization signal for the alignment process.

After DCD, each block is decoded in a single step, enabling fast inference while incurring no more than 2–5% quality degradation, as empirically demonstrated in Sec. 5.

### 4.3 Hallucination Mitigation

Hallucination is a critical concern in CXR report generation, where inaccurate descriptions of clinical findings can directly undermine diagnostic reliability. During the development of our vision-language model, we identify two predominant hallucination patterns in medical dLLMs: inaccurate identification of symptoms, and degenerate repetition loops that fail to terminate.

For inaccurate symptom identification, we attribute this pattern to an implicit negative bias inherent in the training data. In routine clinical practice, radiologists follow a "reporting by exception" convention, describing only abnormal findings while omitting normal structures or dismissing them with a single catch-all phrase such as "no other significant abnormality." Consequently, for

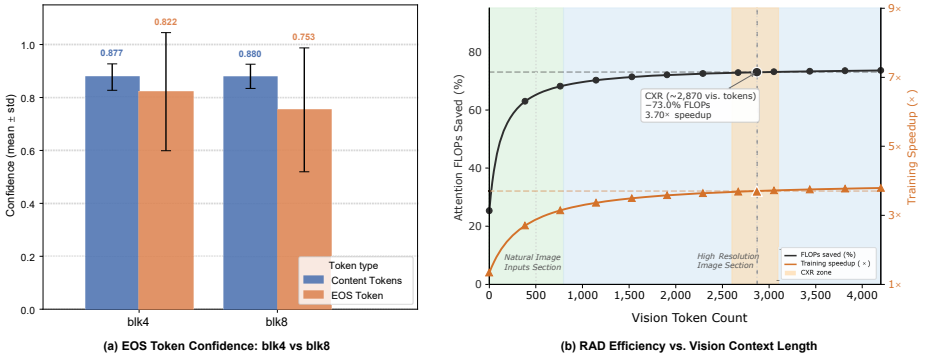



**Fig. 3: (a)** Confidence analysis of <eos> tokens vs. content tokens under block sizes blk4 and blk8 in standard multi-step inference.**(b)** RAD attention FLOPs saved (%) and training speedup (×) as a function of vision token count.

the majority of anatomical regions that are clinically unremarkable, their normal status is absent from the report rather than explicitly stated as negative. This systematic omission leaves the model without explicit negative evidence during training, leaving its conditional distribution under-constrained. At inference time, this under-constraint manifests as two complementary failure modes: false-positive hallucinations, where the model fabricates findings for normal regions, and false-negative omissions, where genuine abnormalities are overlooked. To address this, we propose a data normalization paradigm tailored for CXR report generation. Specifically, we reformulate every training report so that each predefined anatomical region receives an explicit annotation, either a positive finding or a negative assertion. This ensures unambiguous supervision at every position, directly eliminating the implicit omission bias described above. Our experiments in Sec. 5.5 show that the hallucination reduction brought by data normalization consistently benefits both $\texttt{ECHO}_{\texttt{AR}}$ and the final ECHO model.

Beyond inaccurate symptom identification, ECHO also suffers from degenerate repetition loops that fail to terminate. To understand the root cause, we examine $\texttt{ECHO}_{\texttt{Base}}$'s prediction confidence of content tokens and <eos> tokens under standard confidence-heuristic inference. As shown in Fig. 3(a), <eos> tokens exhibit systematically lower mean confidence and substantially greater variance compared to content tokens. Using such a flat and unstable distribution as the distillation target for the <eos> token makes it difficult for the one-step model to terminate generation with high confidence. This problem is further exacerbated as the block size grows: as shown in Fig. 3(a), increasing the block size leaves content token confidence nearly unaffected while causing a notable decline in <eos> confidence. To address this, we apply an additional cross-entropy loss on the <eos> token during distillation, explicitly pulling its predicted distribution toward a sharp, high-confidence one-hot target. Together, the two hallucination mitigation strategies consistently reduce generation errors in ECHO, contributing to improved clinical accuracy.

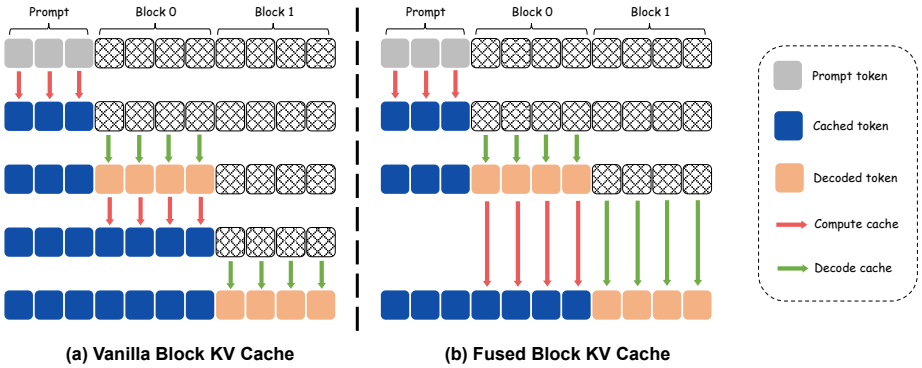



**Fig. 4:** **(a)** Vanilla block KV cache: after all tokens of a block are committed, a dedicated forward pass is performed to update the KV cache. **(b)** Fused block KV cache: the KV cache update for the preceding block is fused into the current block's denoising forward, eliminating the dedicated KV update pass.

### 4.4 Inference Paradigm

To further maximize the inference throughput of `ECHO`, we optimize the inference paradigm used in dLLMs. A common technique in semi-autoregressive decoding is block KV cache [52], which caches the key-value states of previously decoded blocks to avoid redundant recomputation. As illustrated in Fig. 4(a), after all tokens of a block are committed, a dedicated forward pass is performed solely to update the KV cache with the newly decoded block. This additional forward pass incurs acceptable inference time overhead in multi-step decoding, but for our one-step model it doubles the total number of forward passes for each block. To eliminate this overhead, we propose **Fused Block KV Cache.** As illustrated in Fig. 4(b), after a block is decoded, we defer its KV cache update and fuse it into the next block's denoising forward. In this fused forward, the model simultaneously computes and caches the key-value states for the preceding decoded block while denoising the current block's fully-masked tokens, removing the dedicated KV update pass entirely. As proven in Supplementary Sec. **??**, Fused Block KV Cache introduces no additional FLOPs while halving the number of forward passes, directly reducing inference latency.

## 5 Experiments

In this section, we present experimental results to assess the effectiveness of `ECHO` for fast and reliable CXR report generation. We compare `ECHO` against state-of-the-art autoregressive baselines on report quality, and against diffusion-based distillation baselines on distillation efficiency. We then ablate the individual components of DCD, and analyze the effect of training data volume during RAD adaptation and the impact of report normalization across all training stages.

### 5.1 Experimental Setups

**Dataset** We conduct experiments on a unified CXR report corpus built from multiple public datasets, including MIMIC-CXR [24], CheXpert-Plus [22], ReX-Gradient [62], and IU-Xray [11].
For training, we apply a standardized cleaning and normalization pipeline across all source datasets to construct a bilingual training set, with a small subset of LLaVA-ReCap-558K [28] included to mitigate catastrophic forgetting. For evaluation, we sample 2,000 English and 2,000 Chinese reports from each of normalized MIMIC-CXR, CheXpert-Plus, and ReXGradient, ensuring balanced coverage across datasets and languages. For MIMIC-CXR, evaluation samples are drawn exclusively from the official patient-level test split. Full preprocessing details and data statistics are provided in Supplementary Sec. **??**.

**Models** All `ECHO` models are initialized from Lingshu-7B [54], an autoregressive VLM pretrained on large-scale medical corpora, which provides strong domain-specific knowledge and vision alignment for CXR report generation.

**Implementation Details** For RAD, we use the same data as the SFT stage in continued pre-training. The vision encoder and projector weights of $\texttt{ECHO}_{\texttt{AR}}$ are frozen throughout, and only the LLM backbone is trained. For DCD, we randomly sample 30,000 examples from the RAD training set, drawn proportionally across datasets. During teacher trajectory collection, samples that produce degenerate repetition loops are automatically discarded.

**Evaluation Metrics** To provide a comprehensive assessment of the generated chest X-ray reports, we evaluate our model across three complementary dimensions: linguistic quality, clinical fidelity, and structural stability. These metrics enable a rigorous comparison between `ECHO` and existing state-of-the-art (SOTA) models while specifically addressing the unique challenges of diffusion-based generation.

- **Linguistic Quality Metrics (LQM).** We use standard natural language generation NLG metrics, including ROUGE-L [32] and CIDEr [47], to measure lexical overlap and fluency between generated and reference reports, enabling direct comparison with prior methods.
- **Clinical Fidelity Metrics (CFM).** Linguistic fluency does not guarantee accurate symptom identification, so we use RaTEScore [67] and SemScore [43] to assess clinical content fidelity. Since our evaluation set consists of normalized reports, a fair comparison across models requires focusing solely on symptom identification accuracy rather than reporting style. To this end, we only adopt the positive-finding score of RaTEScore, and SemScore is insensitive to negative findings by design.
- **Structural Stability Metrics (SSM).** To assess generation stability specific to discrete diffusion models, we report Perplexity (PPL) computed with

**Table 1: Performance comparison with state-of-the-art models.** We report average metrics across the MIMIC-CXR, CheXpert-Plus, and ReXGradient test sets. **Bold** values indicate the best performance. $\text{ECHO}_{\text{blk}k}$ denotes block size $k$ ($k$ tokens decoded per forward pass). Please refer to Supplementary Sec. **??** for detailed results on individual datasets.

| Methods | CheXpert-Plus | | | | ReXGradient | | | | MIMIC-CXR | | | | Speed | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ROUGE-L | CIDEr | RaTEScore | SemScore | ROUGE-L | CIDEr | RaTEScore | SemScore | ROUGE-L | CIDEr | RaTEScore | SemScore | TPF | TPS |
| **Proprietary general models** | | | | | | | | | | | | | | |
| Gemini3-Pro [18] | 26.95 | 1.53 | 40.86 | 29.72 | 32.10 | 1.81 | 33.52 | 39.02 | 27.08 | 1.55 | 41.42 | 30.45 | ×1.0 | – |
| Qwen3-Max [3] | 27.19 | 1.53 | 41.63 | 29.24 | 30.50 | 1.72 | 39.12 | 38.13 | 26.86 | 1.51 | 40.78 | 27.32 | ×1.0 | – |
| **Autoregressive Medical Models** | | | | | | | | | | | | | | |
| LLaVA-Med [27] | 6.92 | 0.65 | 10.11 | 11.39 | 7.36 | 0.72 | 10.01 | 17.04 | 6.69 | 0.65 | 10.07 | 9.22 | ×1.0 | 36.33 |
| Lingshu-7B [54] | 21.95 | 1.12 | 31.34 | 27.54 | 23.63 | 1.25 | 20.26 | 34.82 | 22.24 | 1.16 | 32.79 | 26.94 | ×1.0 | 53.70 |
| Hulu-Med-7B [23] | 22.38 | 1.15 | 26.25 | 23.43 | 21.66 | 1.14 | 22.77 | 23.56 | 22.26 | 1.18 | 22.89 | 22.30 | ×1.0 | 38.78 |
| MedGemma-27B [41] | 20.30 | 0.79 | 38.83 | 31.79 | 23.41 | 0.91 | 26.00 | 37.74 | 20.80 | 0.82 | 38.48 | 30.24 | ×1.0 | 16.78 |
| Lingshu-32B [54] | 20.40 | 1.06 | 30.92 | 24.32 | 22.10 | 1.20 | 20.33 | 31.49 | 20.73 | 1.10 | 32.40 | 24.14 | ×1.0 | 23.60 |
| Hulu-Med-32B [23] | 19.46 | 0.96 | 27.17 | 22.07 | 22.31 | 1.11 | 23.01 | 26.67 | 17.96 | 0.99 | 24.54 | 18.23 | ×1.0 | 15.67 |
| **Diffusion Medical Methods** | | | | | | | | | | | | | | |
| LLaDA-MedV [13] | 7.69 | 0.22 | 26.72 | 17.76 | 9.72 | 0.23 | 18.19 | 22.09 | 9.60 | 0.23 | 27.47 | 17.26 | ×1.30 | 12.86 |
| $\text{ECHO}_{\text{Base}}$ | 56.90 | 4.25 | 59.12 | 52.64 | 66.46 | 5.64 | 63.02 | 66.68 | 54.45 | 4.16 | 55.85 | 46.10 | ×1.89 | 48.86 |
| CD4LM [31] | 49.21$^{13\%\downarrow}$ | 3.38$^{29\%\downarrow}$ | 55.90$^{5\%\downarrow}$ | 43.86$^{16\%\downarrow}$ | 58.73$^{11\%\downarrow}$ | 4.44$^{21\%\downarrow}$ | 58.36$^{7\%\downarrow}$ | 55.62$^{16\%\downarrow}$ | 46.21$^{15\%\downarrow}$ | 3.13$^{24\%\downarrow}$ | 52.11$^{6\%\downarrow}$ | 38.76$^{15\%\downarrow}$ | ×3.61 | 98.49 |
| d3LLM [38] | 44.04$^{22\%\downarrow}$ | 2.92$^{31\%\downarrow}$ | 55.39$^{6\%\downarrow}$ | 40.35$^{23\%\downarrow}$ | 54.18$^{18\%\downarrow}$ | 4.17$^{26\%\downarrow}$ | 57.99$^{7\%\downarrow}$ | 53.50$^{19\%\downarrow}$ | 41.13$^{24\%\downarrow}$ | 2.65$^{36\%\downarrow}$ | 50.99$^{8\%\downarrow}$ | 34.20$^{25\%\downarrow}$ | ×3.84 | 101.64 |
| dParallel [8] | 42.22$^{25\%\downarrow}$ | 3.17$^{25\%\downarrow}$ | 44.65$^{24\%\downarrow}$ | 36.72$^{30\%\downarrow}$ | 48.73$^{26\%\downarrow}$ | 4.02$^{28\%\downarrow}$ | 45.50$^{27\%\downarrow}$ | 54.53$^{18\%\downarrow}$ | 39.98$^{26\%\downarrow}$ | 2.96$^{28\%\downarrow}$ | 40.47$^{27\%\downarrow}$ | 31.15$^{32\%\downarrow}$ | ×4.42 | 110.12 |
| T3D [61] | 54.99$^{3\%\downarrow}$ | 4.06$^{4\%\downarrow}$ | 56.90$^{3\%\downarrow}$ | 49.86$^{5\%\downarrow}$ | 64.36$^{3\%\downarrow}$ | 5.23$^{7\%\downarrow}$ | 58.63$^{6\%\downarrow}$ | 61.65$^{7\%\downarrow}$ | 52.38$^{3\%\downarrow}$ | 3.85$^{7\%\downarrow}$ | 52.38$^{6\%\downarrow}$ | 42.51$^{7\%\downarrow}$ | ×2.00 | 60.25 |
| $\text{ECHO}_{\text{blk8}}$ | 55.21$^{2\%\downarrow}$ | 4.14$^{2\%\downarrow}$ | 56.85$^{3\%\downarrow}$ | 51.40$^{2\%\downarrow}$ | 65.13$^{2\%\downarrow}$ | 5.48$^{2\%\downarrow}$ | 59.92$^{4\%\downarrow}$ | 64.00$^{4\%\downarrow}$ | 53.07$^{2\%\downarrow}$ | 4.00$^{3\%\downarrow}$ | 52.97$^{5\%\downarrow}$ | 44.83$^{2\%\downarrow}$ | **×8.00** | **274.21** |
| $\text{ECHO}_{\text{blk4}}$ | **56.14**$^{1\%\downarrow}$ | **4.20**$^{1\%\downarrow}$ | **57.40**$^{2\%\downarrow}$ | **49.57**$^{5\%\downarrow}$ | **66.39**$^{0\%\downarrow}$ | **5.54**$^{1\%\downarrow}$ | **62.18**$^{1\%\downarrow}$ | **66.28**$^{0\%\downarrow}$ | **54.12**$^{0\%\downarrow}$ | **4.05**$^{2\%\downarrow}$ | **53.95**$^{3\%\downarrow}$ | **45.57**$^{1\%\downarrow}$ | ×4.00 | 129.19 |

Qwen3-1.7B as the judge model. Lower PPL indicates more stable and fluent generation.

– **Efficiency Metrics (EM).** To evaluate decoding efficiency, we report two complementary speed metrics. TPF (Tokens Per Forward Pass) measures theoretical throughput as the number of decoded tokens per forward pass, normalized relative to the AR baseline (×1.0). TPS (Tokens Per Second) measures practical decoding throughput as the number of tokens generated per second during inference.

### 5.2 Comparison with State-of-the-art Methods

In Table 1, to evaluate the effectiveness of ECHO, we compare it against three categories of state-of-the-art models: (1) General-purpose proprietary models, including Gemini3-Pro [18] and Qwen3-Max [3]. (2) Autoregressive (AR) medical VLMs, such as LLaVA-Med v1.5-7B [27], Lingshu-7B/32B [54], MedGemma-27B [41], and Hulu-Med-7B/32B [23]. (3) Diffusion-based medical models, including LLaDA-MedV-8B [13] and several distilled variants [8,31,38,61]. These variants share the same base model $\text{ECHO}_{\text{Base}}$ and distillation data, but with different distillation targets.

**Overall Performance.** ECHO consistently achieves superior performance compared to both general and specialized models. Notably, it significantly outperforms larger medical VLMs like MedGemma-27B, ranging from 17% to 40% in CFM. Even when compared to the most powerful models like Gemini3-Pro and Qwen3-Max, ECHO maintains a clear advantage in all medical clinical metrics.

**Efficiency of Distillation.** To further differentiate distillation strategies, we compare all methods under the challenging block size $L = 8$ configuration. As

**Table 2: Ablation on DCD components.** SW: step-wise token reweighting; CE: cross-entropy loss on the <eos> token; RKL: reverse KL divergence. Results are reported on CheXpert-Plus, ReXGradient, and MIMIC-CXR.

| DCD settings | | | CheXpert-Plus | | | | ReXGradient | | | | MIMIC-CXR | | | | PPL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SW | CE | RKL | ROUGE-L | CIDEr | RaTEScore | SemScore | ROUGE-L | CIDEr | RaTEScore | SemScore | ROUGE-L | CIDEr | RaTEScore | SemScore | |
| | | | 52.57 | 3.61 | 54.87 | 46.93 | 63.28 | 5.17 | 61.32 | 58.36 | 51.73 | 3.63 | 51.19 | 43.50 | 23.72 |
| ✓ | | | 52.44 | 3.53 | 56.30 | 46.66 | 62.90 | 5.07 | 61.33 | 59.90 | 51.76 | 3.65 | 53.14 | 43.70 | 21.07 |
| ✓ | ✓ | | **56.14** | **4.20** | **57.40** | **49.57** | **66.39** | **5.54** | **62.18** | **66.28** | **54.12** | **4.05** | **53.95** | **45.57** | **18.83** |
| ✓ | | ✓ | 52.51 | 3.41 | 55.20 | 45.32 | 61.71 | 4.94 | 61.26 | 59.55 | 50.91 | 3.48 | 53.31 | 42.43 | 20.23 |
| ✓ | ✓ | ✓ | 53.25 | 3.72 | 57.25 | 47.07 | 63.82 | 5.21 | 61.47 | 63.46 | 51.48 | 3.67 | 53.42 | 43.86 | 21.32 |

shown in Tab. 1, $\texttt{ECHO}_{\text{blk8}}$ achieves an 8× theoretical speedup over the AR baseline while incurring only 2–5% quality degradation relative to $\texttt{ECHO}_{\texttt{Base}}$ across all metrics. By contrast, T3D delivers only a 2× theoretical speedup yet still incurs a comparable quality drop, and dParallel achieves a 4.4× speedup at the cost of 18–32% degradation in clinical fidelity metrics. $\texttt{ECHO}_{\text{blk4}}$ further provides a more conservative working point, where a 4× theoretical speedup reduces quality degradation to within 1.25% on average. Taken together, these results confirm that DCD achieves a consistently superior quality–speed trade-off across both block size configurations and against all existing distillation methods.

### 5.3 Ablation Study of Distillation Components

To investigate the individual contributions of the components in our proposed Direct Conditional Distillation (DCD) strategy, we conduct extensive ablation studies. Specifically, we evaluate the impact of Step-wise Weighting (SW), Cross-Entropy loss for the <eos> token (CE), and Reverse KL divergence (RKL). The results across three benchmarks are summarized in Tab. 2.

**Effect of Step-wise Weighting (SW).** As shown in row 2 of Tab. 2, SW consistently improves generation stability across all datasets. Upweighting tokens that are unmasked later in the denoising trajectory, which tend to carry stronger inter-token dependencies, reduces PPL from 23.72 to 21.07. RaTEScore also improves steadily, e.g., from 54.87 to 56.30 on CheXpert-Plus, indicating that the reweighting helps the model better capture the conditioning provided by earlier committed tokens.

**Effect of <eos> Cross-Entropy Loss (CE).** Adding CE supervision on the <eos> token produces the largest single improvement across all settings. As discussed in Sec. 4.2, one-step diffusion models tend to assign low and unstable confidence to <eos> positions, leading to degenerate repetition loops. Explicitly supervising the <eos> token with a one-hot target directly addresses this failure mode, resulting in a substantial gain in ROUGE-L from 52.44 to 56.14 on CheXpert-Plus and CIDEr from 3.65 to 4.05 on MIMIC-CXR. PPL drops further to 18.83, its lowest value across all configurations, confirming that reliable termination is essential for overall report quality.

**Effect of Reverse KL (RKL).** We also replace the forward KL objective with reverse KL (RKL) to examine whether its mode-seeking behavior benefits distillation. As shown in rows 4 and 5 of Tab. 2, adding RKL consistently degrades

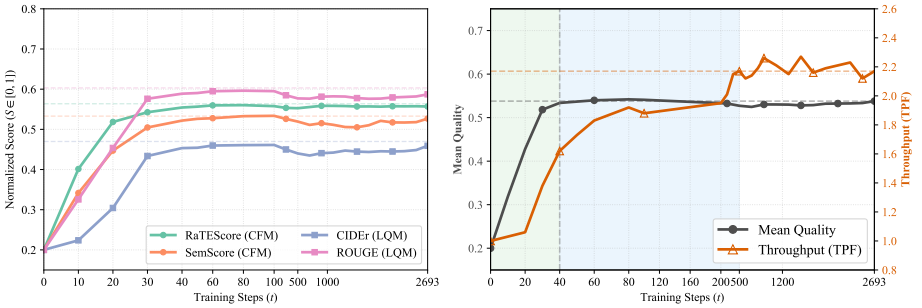



**Fig. 5: Effect of training data scale during RAD adaptation. (a) Quality Metric Convergence During RAD.** Each colored curve tracks one of four clinical and linguistic metrics across training steps, and dashed horizontal lines mark the corresponding $\texttt{ECHO}_{\texttt{AR}}$ baseline values. **(b) Decoupled Convergence of Quality and Throughput.** The black curve shows the mean quality score and the red curve shows inference throughput (TPF), and dashed horizontal lines mark respective final values.

performance. Compared to the SW-only baseline, CIDEr drops from 3.65 to 3.48 and SemScore from 43.70 to 42.43 on MIMIC-CXR. A likely cause is that clinical reports require the model to cover all plausible findings rather than concentrate probability mass on a single mode, which is the tendency encouraged by reverse KL. Forward KL, which preserves the full teacher distribution, is therefore better suited for this task.

### 5.4 Analysis of Data Scale of RAD

We study how the amount of training data used in the RAD stage affects both model quality and inference throughput, by evaluating checkpoints at different training steps across all benchmarks. Fig. 5 reports five clinical and linguistic metrics alongside inference throughput (TPF) as functions of training steps.

As shown in Fig. 5(a), all four quality metrics follow a rapid climb-and-plateau pattern. Within approximately 60 training steps, performance on all metrics reaches or surpasses the $\texttt{ECHO}_{\texttt{AR}}$ baseline, indicating that RAD transfers the domain knowledge of the pretrained AR model efficiently and requires only a small fraction of the full training data to do so.

Fig. 5(b) reveals that quality and throughput converge at different rates. Quality saturates early (around step 60, corresponding to only 2.2% of the full RAD training data), whereas TPF continues to improve throughout training, increasing from 1.62 to 2.17 (+33.95%). A plausible interpretation is that while cross-modal semantic alignment is established quickly, additional training is needed for the model to stabilize the joint token distribution within each block, which in turn enables more tokens to be committed with high confidence per denoising step and raises throughput.

**Table 3: Ablation on report data types across all training stages.** Each stage contains two settings: normalized and unnormalized reports.

| Stage | Data Type | CheXpert-Plus | | | | ReXGradient | | | | MIMIC-CXR | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ROUGE-L | CIDEr | RaTEScore | SemScore | ROUGE-L | CIDEr | RaTEScore | SemScore | ROUGE-L | CIDEr | RaTEScore | SemScore |
| **Stage I** | Normalized | 56.89 | 4.47 | 59.18 | 52.94 | 67.07 | 5.68 | 64.39 | 66.60 | 54.55 | 4.32 | 56.58 | 46.70 |
| | Unnormalized | 23.82 | 1.53 | 45.59 | 32.44 | 25.67 | 1.49 | 26.05 | 40.77 | 25.06 | 1.56 | 42.72 | 33.54 |
| **Stage II** | Normalized | 56.90 | 4.25 | 59.12 | 52.64 | 66.46 | 5.64 | 63.02 | 66.68 | 54.45 | 4.16 | 55.85 | 46.10 |
| | Unnormalized | 23.47 | 1.51 | 39.60 | 30.46 | 24.97 | 1.45 | 23.90 | 36.97 | 24.54 | 1.53 | 37.76 | 32.64 |
| **Stage III** | Normalized | 56.14 | 4.20 | 57.40 | 49.57 | 66.39 | 5.54 | 62.18 | 66.28 | 54.12 | 4.05 | 53.95 | 45.57 |
| | Unnormalized | 18.79 | 1.19 | 42.08 | 27.53 | 19.98 | 1.20 | 24.98 | 31.70 | 19.61 | 1.24 | 37.55 | 28.02 |

### 5.5 Analysis of Report Normalization

To assess the impact of report normalization, we retrain the pipeline at each stage using unnormalized raw reports while keeping all other settings unchanged. Tab. 3 summarizes the results across all three training stages.

**Impact on Stage I (Continued Pre-training).** Normalized data yields substantial improvements in Stage I across all metrics. On CheXpert-Plus, ROUGE-L increases from 23.82 to 56.89 and RaTEScore from 45.59 to 59.18. This large gap indicates that raw reports, which frequently contain inconsistent phrasing and systematically omit negative findings, provide ambiguous supervision that prevents the model from learning reliable visual-textual mappings.

**Impact on Stage II and Stage III (Adaptation and Distillation).** The gap persists through Stage II and widens further in Stage III. Without normalization, ROUGE-L collapses to 18.79 on CheXpert-Plus and 19.98 on ReXGradient after distillation, substantially below the already-degraded Stage II unnormalized baseline. With normalization, the model retains strong clinical accuracy across all stages, with SemScore remaining above 45.0 on CheXpert-Plus even after one-step distillation. This progressive degradation under unnormalized training suggests that noisy supervision compounds across stages: the omission bias that distorts pretraining is further amplified by the mean-field bias of block diffusion, and becomes most severe when the student must reproduce the teacher's full conditional distribution in a single step.

## 6 Conclusion

We presented `ECHO`, a discrete diffusion VLM that demonstrates high clinical accuracy and extreme decoding efficiency are simultaneously achievable in automated chest X-ray reporting. By rethinking both the AR-to-diffusion conversion and the distillation target, `ECHO` reduces training cost, eliminates multi-step decoding, and consistently outperforms autoregressive and diffusion-based state-of-the-art methods across standard benchmarks. We hope this work provides a strong foundation for efficient and reliable generation in broader medical multimodal settings beyond CXR reporting.

## Acknowledgements

This work was supported by the Talent Fund of Beijing Jiaotong University under Grant No. 2025XKRC015.

## References


1. Arriola, M., Gokaslan, A., Chiu, J.T., Yang, Z., Qi, Z., Han, J., Sahoo, S.S., Kuleshov, V.: Block diffusion: Interpolating between autoregressive and diffusion language models. In: The Thirteenth International Conference on Learning Representations (2025)
2. Austin, J., Johnson, D.D., Ho, J., Tarlow, D., Van Den Berg, R.: Structured denoising diffusion models in discrete state-spaces. Advances in neural information processing systems (2021)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bie, T., Cao, M., Chen, K., Du, L., Gong, M., Gong, Z., Gu, Y., Hu, J., Huang, Z., Lan, Z., et al.: Llada2. 0: Scaling up diffusion language models to 100b. arXiv preprint arXiv:2512.15745 (2025)
5. Chandrasegaran, K., Thomas, A.W., Ku, J., Berto, F., Kim, J.M., Brixi, G., Nguyen, E., Massaroli, S., Poli, M.: Rnd1: Simple, scalable ar-to-diffusion conversion (2025)
6. Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Cai, Z., Ji, K., Wan, X., et al.: Towards injecting medical visual knowledge into multimodal llms at scale. In: Proceedings of the 2024 conference on empirical methods in natural language processing (2024)
7. Chen, Z., Varma, M., Xu, J., Paschali, M., Van Veen, D., Johnston, A., Youssef, A., Blankemeier, L., Bluethgen, C., Altmayer, S., et al.: A vision-language foundation model to enhance efficiency of chest x-ray interpretation. arXiv preprint arXiv:2401.12208 (2024)
8. Chen, Z., Fang, G., Ma, X., Yu, R., Wang, X.: dparallel: Learnable parallel decoding for dllms. arXiv preprint arXiv:2509.26488 (2025)
9. Cheng, S., Jiang, Y., Zhou, Z., Liu, D., Tao, W., Zhang, L., Qi, B., Zhou, B.: Sdar-vl: Stable and efficient block-wise diffusion for vision-language understanding. arXiv preprint arXiv:2512.14068 (2025)
10. Christopher, J.K., Bartoldson, B.R., Ben-Nun, T., Cardei, M., Kailkhura, B., Fioretto, F.: Speculative diffusion decoding: Accelerating language generation through diffusion. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (2025)
11. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association (2016)
12. Deschenaux, J., Gulcehre, C.: Beyond autoregression: Fast llms via self-distillation through time (2025)
13. Dong, X., Zhu, W., Chen, X., Wang, Z., Qiu, P., Tang, S., Li, X., Wang, Y.: Llada-medv: Exploring large language diffusion models for biomedical image understanding. arXiv preprint arXiv:2508.01617 (2025)
14. Du, S., Zou, Y., Li, J., Liu, M., Li, Y., Shang, C., Shen, Q.: Pansharpening for thin-cloud contaminated remote sensing images: a unified framework and benchmark dataset. In: Proceedings of the AAAI Conference on Artificial Intelligence (2026)
15. Du, S., Zou, Y., Wang, Z., Li, X., Li, Y., Shang, C., Shen, Q.: Unsupervised hyperspectral image super-resolution via self-supervised modality decoupling. International Journal of Computer Vision (2026)

16. Fathi, N., Scholak, T., Noel, P.A.: Unifying autoregressive and diffusion-based sequence generation. In: ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy (2025)
17. Gong, S., Agarwal, S., Zhang, Y., Ye, J., Zheng, L., Li, M., An, C., Zhao, P., Bi, W., Han, J., et al.: Scaling diffusion language models via adaptation from autoregressive models. In: The Thirteenth International Conference on Learning Representations (2025)
18. Google DeepMind: Gemini 3 pro model card (2025)
19. Han, X., Kumar, S., Tsvetkov, Y.: Ssd-lm: Semi-autoregressive simplex-based diffusion language model for text generation and modular control. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (2023)
20. Hu, Z., Meng, J., Akhauri, Y., Abdelfattah, M.S., Seo, J.s., Zhang, Z., Gupta, U.: Accelerating diffusion language model inference via efficient kv caching and guided diffusion. arXiv e-prints (2025)
21. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
22. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence (2019)
23. Jiang, S., Wang, Y., Song, S., Hu, T., Zhou, C., Pu, B., Zhang, Y., Yang, Z., Feng, Y., Zhou, J.T., et al.: Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. arXiv preprint arXiv:2510.08668 (2025)
24. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data (2019)
25. Lee, S., Youn, J., Kim, H., Kim, M., Yoon, S.H.: Cxr-llava: a multimodal large language model for interpreting chest x-ray images. European Radiology (2025)
26. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. Transactions on Machine Learning Research (2024)
27. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems (2023)
28. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895 (2024)
29. Li, G., Fu, Z., Fang, M., Zhao, Q., Tang, M., Yuan, C., Wang, J.: Diffuspec: Unlocking diffusion language models for speculative decoding. arXiv preprint arXiv:2510.02358 (2025)
30. Li, S., Kallidromitis, K., Bansal, H., Gokul, A., Kato, Y., Kozuka, K., Kuen, J., Lin, Z., Chang, K.W., Grover, A.: Lavida: A large diffusion language model for multimodal understanding. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
31. Liang, Y., Wang, Z., Chen, H., Sun, X., Wu, J., Yu, X., Liu, J., Barsoum, E., Liu, Z., Jha, N.K.: Cd4lm: Consistency distillation and adaptive decoding for diffusion language models. arXiv preprint arXiv:2601.02236 (2026)
32. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out (2004)

33. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems (2023)
34. Liu, Z., Yang, Y., Zhang, Y., Chen, J., Zou, C., Wei, Q., Wang, S., Zhang, L.: dllm-cache: Accelerating diffusion large language models with adaptive caching. arXiv preprint arXiv:2506.06295 (2025)
35. Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., ZHOU, J., Lin, Y., Wen, J.R., Li, C.: Large language diffusion models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
36. Pan, J., Liu, C., Wu, J., Liu, F., Zhu, J., Li, H.B., Chen, C., Ouyang, C., Rueckert, D.: Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2025)
37. Pellegrini, C., Özsoy, E., Busam, B., Navab, N., Keicher, M.: Radialog: A large vision-language model for radiology report generation and conversational assistance. arXiv preprint arXiv:2311.18681 (2023)
38. Qian, Y.Y., Su, J., Hu, L., Zhang, P., Deng, Z., Zhao, P., Zhang, H.: d3llm: Ultra-fast diffusion llm using pseudo-trajectory distillation. arXiv preprint arXiv:2601.07568 (2026)
39. Saab, K., Tu, T., Weng, W.H., Tanno, R., Stutz, D., Wulczyn, E., Zhang, F., Strother, T., Park, C., Vedadi, E., et al.: Capabilities of gemini models in medicine. arXiv preprint arXiv:2404.18416 (2024)
40. Sahoo, S., Arriola, M., Schiff, Y., Gokaslan, A., Marroquin, E., Chiu, J., Rush, A., Kuleshov, V.: Simple and effective masked diffusion language models. Advances in Neural Information Processing Systems (2024)
41. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
42. Shi, J., Han, K., Wang, Z., Doucet, A., Titsias, M.: Simplified and generalized masked diffusion for discrete data. Advances in neural information processing systems (2024)
43. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A.Y., Lungren, M.: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). pp. 1500–1519 (2020)
44. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
45. Thawakar, O.C., Shaker, A.M., Mullappilly, S.S., Cholakkal, H., Anwer, R.M., Khan, S., Laaksonen, J., Khan, F.: Xraygpt: Chest radiographs summarization using large medical vision-language models. In: Proceedings of the 23rd workshop on biomedical natural language processing (2024)
46. Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al.: Towards generalist biomedical ai. Nejm Ai (2024)
47. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2015)
48. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)

49. Wang, X., Xu, C., Jin, Y., Jin, J., Zhang, H., Deng, Z.: Diffusion llms can do faster-than-ar inference via discrete diffusion forcing. arXiv preprint arXiv:2508.09192 (2025)
50. Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. Nature Communications (2025)
51. Wu, C., Zhang, H., Xue, S., Diao, S., Fu, Y., Liu, Z., Molchanov, P., Luo, P., Han, S., Xie, E.: Fast-dllm v2: Efficient block-diffusion llm. arXiv preprint arXiv:2509.26328 (2025)
52. Wu, C., Zhang, H., Xue, S., Liu, Z., Diao, S., Zhu, L., Luo, P., Han, S., Xie, E.: Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding. arXiv preprint arXiv:2505.22618 (2025)
53. Xu, M., Geffner, T., Kreis, K., Nie, W., Xu, Y., Leskovec, J., Ermon, S., Vahdat, A.: Energy-based diffusion language models for text generation. In: The Thirteenth International Conference on Learning Representations (2025)
54. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
55. Yang, L., Wang, Z., Chen, Z., Liang, X., Zhou, L.: Medxchat: A unified multimodal large language model framework towards cxrs understanding and generation. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI) (2025)
56. Ye, J., Xie, Z., Zheng, L., Gao, J., Wu, Z., Jiang, X., Li, Z., Kong, L.: Dream 7b: Diffusion large language models. arXiv preprint arXiv:2508.15487 (2025)
57. Yoo, J., Kim, W., Hong, S.: Redi: Rectified discrete flow. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
58. You, Z., Nie, S., Zhang, X., Hu, J., Zhou, J., Lu, Z., Wen, J.R., Li, C.: Llada-v: Large language diffusion models with visual instruction tuning. arXiv preprint arXiv:2505.16933 (2025)
59. Yu, R., Ma, X., Wang, X.: Dimple: Discrete diffusion multimodal large language model with parallel decoding. arXiv preprint arXiv:2505.16990 (2025)
60. Zhang, K., Zhou, R., Adhikarla, E., Yan, Z., Liu, Y., Yu, J., Liu, Z., Chen, X., Davison, B.D., Ren, H., et al.: A generalist vision–language foundation model for diverse biomedical tasks. Nature medicine (2024)
61. Zhang, T., Zhang, X., Han, L., Shi, H., He, X., Li, Z., Wang, H., Xu, K., Srivastava, A., Pavlovic, V., et al.: T3d: Few-step diffusion language models via trajectory self-distillation with direct discriminative optimization. arXiv preprint arXiv:2602.12262 (2026)
62. Zhang, X., Acosta, J.N., Miller, J., Huang, O., Rajpurkar, P.: Rexgradient-160k: A large-scale publicly available dataset of chest radiographs with free-text reports. arXiv preprint arXiv:2505.00228 (2025)
63. Zhang, Y., Schwing, A., Zhao, Z.: Variational masked diffusion models. arXiv preprint arXiv:2510.23606 (2025)
64. Zhao, C., Cai, W., Dong, C., Hu, C.: Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2024)
65. Zhao, C., Chen, J., Li, H., Kang, Z., Lu, S., Wei, X., Zhang, K., Yang, J., Tai, Y.: LUVE : Latent-cascaded ultra-high-resolution video generation with dual frequency experts. In: Forty-third International Conference on Machine Learning (2026)

66. Zhao, C., Ci, E., Xu, Y., Fan, T., Guan, S., Ge, Y., Yang, J., Tai, Y.: Ultrahr-100k: Enhancing uhr image synthesis with a large-scale high-quality dataset. Advances in Neural Information Processing Systems (2025)
67. Zhao, W., Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Ratescore: A metric for radiology report generation. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (2024)