Ambiguity-aware Truncated Flow Matching for Ambiguous Medical Image Segmentation

Fanding Li[1], Xiangyu Li[1]*, Xianghe Su[1], Xingyu Qiu[1], Suyu Dong[2], Wei Wang[3], Kuanquan Wang[1], Gongning Luo[1], Shuo Li[4,5]

[1]Faculty of Computing, Harbin Institute of Technology, Harbin, China
[2]College of Computer and Control Engineering, Northeast Forestry University, Harbin, China
[3]Faculty of Computing, Harbin Institute of Technology, Shenzhen, China
[4]Department of Computer and Data Science, Case Western Reserve University, Cleveland, Ohio 44106, United States
[5]Department of Biomedical Engineering, Case Western Reserve University, Cleveland, Ohio 44106, United States
lixiangyu@hit.edu.cn

Abstract

A simultaneous enhancement of accuracy and diversity of predictions remains a challenge in ambiguous medical image segmentation (AMIS) due to the inherent trade-offs. While truncated diffusion probabilistic models (TDPMs) hold strong potential with a paradigm optimization, existing TDPMs suffer from entangled accuracy and diversity of predictions with insufficient fidelity and plausibility. To address the aforementioned challenges, we propose Ambiguity-aware Truncated Flow Matching (ATFM), which introduces a novel inference paradigm and dedicated model components. Firstly, we propose Data-Hierarchical Inference, a redefinition of AMIS-specific inference paradigm, which enhances accuracy and diversity at data-distribution and data-sample level, respectively, for an effective disentanglement. Secondly, Gaussian Truncation Representation (GTR) is introduced to enhance both fidelity of predictions and reliability of truncation distribution, by explicitly modeling it as a Gaussian distribution at T_{trunc} instead of using sampling-based approximations. Thirdly, Segmentation Flow Matching (SFM) is proposed to enhance the plausibility of diverse predictions by extending semantic-aware flow transformation in Flow Matching (FM). Comprehensive evaluations on LIDC and ISIC3 datasets demonstrate that ATFM outperforms SOTA methods and simultaneously achieves a more efficient inference. ATFM improves GED and HM-IoU by up to 12% and 7.3% compared to advanced methods.

Code — https://github.com/PerceptionComputingLab/ATFM
Extended version — https://arxiv.org/abs/2511.06857

Introduction

Generating a series of predictions with high accuracy and diversity to estimate the distribution of annotation space is of significant importance in ambiguous medical image segmentation (AMIS) (Chavhan et al. 2008; Hong et al. 2021). High diversity reflects the inherent ambiguity present in medical images and high accuracy is critical for supporting dependable clinical decision-making, making both essential components of a reliable AMIS framework (Baumgartner et al. 2019; Zhang et al. 2022).

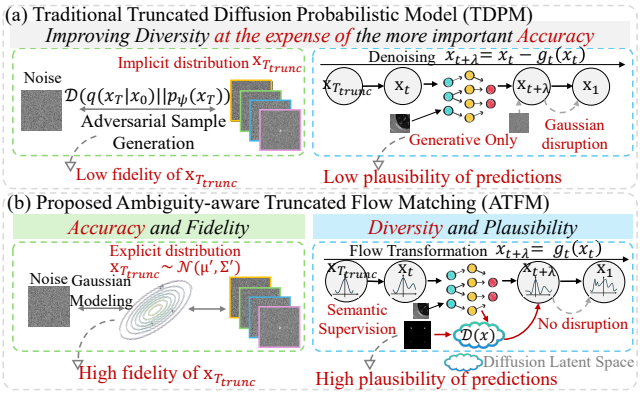


Figure 1: (a) Traditional TDPMs face the challenges of low fidelity and plausibility of predictions by improving diversity at the expense of accuracy. (b) The proposed ATFM enhances fidelity and plausibility of predictions by assigning distinct inference goals into two stages.

However, simultaneously improving prediction accuracy and diversity remains challenging in AMIS due to the inherent trade-off between these objectives in existing methods. Stochastic approaches (Baumgartner et al. 2019; Rahman et al. 2023) enhance diversity at the expense of the more important accuracy, yielding low-confidence diagnoses. Zhang et al. (Zhang et al. 2022) are able to regulate this trade-off but cannot enhance both properties simultaneously. Multi-rater-aware techniques (Zepf et al. 2023; Ji et al. 2021) improve accuracy and diversity by modeling annotators' labeling styles, yet their annotator-centric design inherently suppresses low-frequency modes, degrading segmentation quality. Broader application of these methods is still constrained by the inherent trade-off between entangled accuracy and diversity among predictions.

Truncated Diffusion Probabilistic Models (TDPMs) (Zheng et al. 2022) have shown great potential in simul-

*Corresponding Author

taneously improving accuracy and diversity thanks to the inference paradigm shift. TDPMs are proposed to inference within fewer steps by leveraging an auxiliary network to model the distribution at a predefined truncation point, shifting the original inference for acceleration. TDPMs are widely used in multiple areas such as high-resolution MRSI generating (Dong et al. 2025), autonomous driving (Liao et al. 2025), remote sensing (He et al. 2023), demonstrating superiority in both effectiveness and efficiency of inference.

However, as Fig. 1(a), directly applying TDPMs to achieve synergistic optimization of prediction accuracy and diversity in AMIS remains challenging due to: **(1) The uniform inference objective across all stages in conventional TDPMs inherently constrains their capacity for simultaneous accuracy and diversity improvement.** Vanilla TDPMs adopt a two-stage inference process with one objective mainly for acceleration. However, the deterministic and ambiguous components remain entangled throughout the inference process, making it difficult to achieve a simultaneous enhancement of both accuracy and diversity. **(2) Sub-optimal approximation of the underlying distribution at the truncation point fundamentally compromises prediction fidelity.** Distribution at T_{trunc} is estimated by drawing samples from adversarial networks rather than explicitly modeling in vanilla TDPMs, leading to degraded fidelity caused by inconsistent predictions and omission of low-frequency modes (clinically plausible yet rare). **(3) The absence of semantic guidance following truncation in conventional TDPMs adversely affects the plausibility of generated predictions.** Vanilla TDPMs rely on a vanilla diffusion process guided solely by generation quality after T_{trunc}. In AMIS tasks, this final stage lacks explicit semantic constraints for segmentation, which, although enhancing diversity, significantly compromises the more important accuracy and plausibility.

In this work, we propose the Ambiguity-aware Truncated Flow Matching (ATFM) to achieve a synergistic enhancement of accuracy and diversity of predictions in AMIS tasks, supported by three designs (as Fig. 1(b)): (1) We propose Data-Hierarchical Inference, an innovative AMIS-specific inference paradigm redefinition inspired by TDPMs, where stochasticity during diffusion is marginalized for an effective disentanglement of accuracy and diversity. Specifically, ATFM performs truncated steps at the data distribution level to prioritize accuracy with diversity marginalized, whereas the final diffusion stage operates at the data sample level to enhance diversity without sacrificing accuracy. (2) We propose Gaussian Truncation Representation (GTR), which explicitly models the Gaussian latent distribution at truncation point to enhance prediction fidelity. While traditional TDPMs approximate the implicit distribution via adversarial sampling, GTR encodes image-level semantic features to logit distribution, thereby directly modeling the reliable distribution at the truncation point. This design preserves low-frequency modes and improves consistency, leading to predictions with higher fidelity in AMIS tasks. (3) We propose Segmentation Flow Matching (SFM), which introduces semantic-aware flow transformation to increase diversity and enhance plausibility simultaneously. While traditional TDPMs focus solely on generative quality after truncated steps, the proposed SFM leverages Flow Matching (FM) to overcome Gaussian limitations that disrupt fine-grained predictions and incorporates explicit semantic consistency modeling to ensure the plausibility of segmentation predictions in AMIS tasks.

In summary, our main contributions are as follows:

- We propose ATFM with Data-Hierarchical Inference redefining a more suitable inference paradigm for AMIS for the first time, where Data-Hierarchical Inference effectively decouples accuracy and diversity by marginalizing the stochasticity during diffusion process, thereby enabling a simultaneous improvement of both.
- The proposed GTR in ATFM pioneers explicitly modeling a Gaussian distribution at truncation point, which effectively preserves low-frequency modes and ensures sample consistency, thereby significantly improving prediction fidelity to ground truth distribution and reliability of the distribution at truncation point.
- The proposed SFM in ATFM pioneers semantic-aware flow transformation by modeling semantic consistency at each timestep, thereby enhancing plausibility while emphasizing sample-wise diversity, and is built upon the Flow Matching (FM) process that inherently avoids disturbance from Gaussian constraints.
- A comprehensive evaluation on the LIDC and ISIC3 subset datasets demonstrates that proposed ATFM significantly improves the SOTA methods and simultaneously achieves a more efficient inference.

Related Work

Ambiguous Medical Image Segmentation

Existing AMIS approaches fall into four main paradigms: model ensemble, multi-head frameworks, conditional variational autoencoder (cVAE)-based models, and diffusion-based models. All face a fundamental trade-off between prediction accuracy and sample diversity (Zhang et al. 2022).

Model ensemble (Monteiro et al. 2020; Lipman et al. 2023) and multi-head models (Ho, Jain, and Abbeel 2020; Kohl et al. 2018) generate multiple predictions via diverse architectures or output heads but do not change the original inference process. Consequently, prediction quality heavily depends on model selection, limiting simultaneous improvement of accuracy and diversity.

CVAE-based (Baumgartner et al. 2019; Kohl et al. 2019) and diffusion-based methods (Rahman et al. 2023; Zbinden et al. 2023) inject stochasticity to enhance diversity, yet both follow a one-stage inference paradigm that entangles accuracy and diversity optimization. This leads to inherent conflicts where gains in diversity often reduce accuracy.

To address the challenges, in the proposed ATFM, we introduce Data-Hierarchical Inference to redefine the inference paradigm for AMIS inspired by TDPMs. Specifically, a principled decoupling of the two objectives is achieved across different data hierarchies, where accuracy is enhanced at distributional level and diversity is promoted at sample level.

Truncated Diffusion Probabilistic Models

Truncated Diffusion Probabilistic Models (TDPMs) accelerate inference by truncating the diffusion process at $T_{\text{trunc}} \ll$

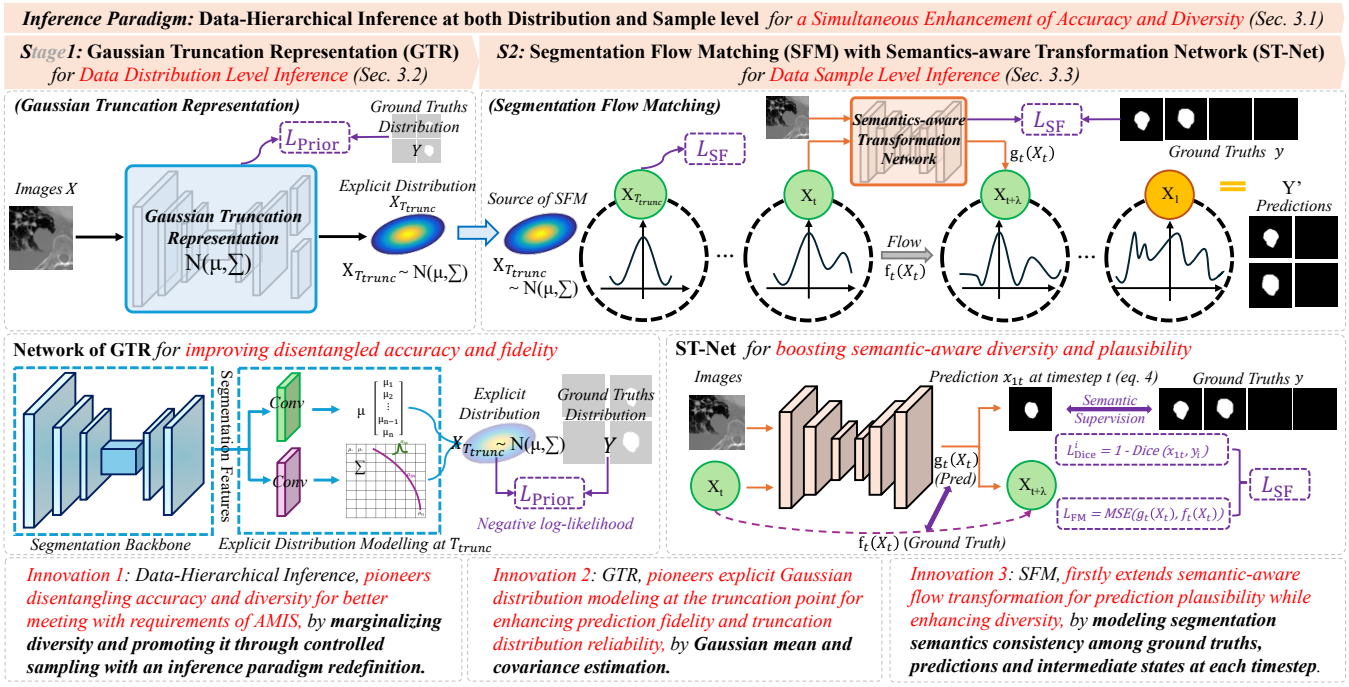


Figure 2: The proposed ATFM addresses the challenge of a synergistic optimization by boosting accuracy at distribution level and diversity at sample level within Data-Hierarchical Inference (Sec. 3.1) while enhancing fidelity and plausibility with GTR (Sec. 3.2) and SFM (Sec. 3.3), respectively.

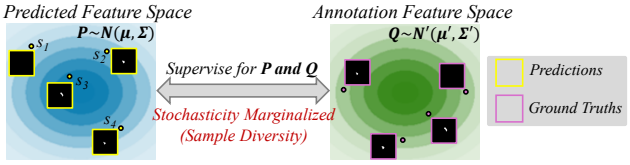


Figure 3: Data-Hierarchical Inference disentangles accuracy and diversity by marginalizing stochasticity during diffusion with a data-distribution level supervision.

T, splitting inference into estimating the distribution at T_{trunc} and reverse diffusion thereafter. Existing TDPMs approximate the distribution at T_{trunc} via adversarial sampling or perturbations (Zheng et al. 2022; Dong et al. 2025), improving speed but without redefining the inference paradigm. Consequently, they lack explicit modeling and semantic supervision at T_{trunc}, limiting performance on AMIS tasks.

The proposed ATFM addresses these gaps by introducing Data-Hierarchical Inference, which marginalizes stochasticity before truncation to disentangle and enhance both accuracy and diversity. Consequently, ATFM explicitly models a Gaussian distribution at T_{trunc} for high-fidelity sampling and applies semantic supervision after truncation to improve plausibility, making ATFM a tailored solution for AMIS.

Methods

The proposed ATFM (Fig. 2) redefines the inference paradigm for disentanglement and simultaneous enhancement of accuracy and diversity by marginalizing stochasticity during truncation, forming an AMIS-specific defining solution. (*Data-Hierarchical Inference, Sec. 3.1*). Specifically, ATFM firstly improves prediction fidelity and truncation distribution reliability by explicitly modeling the Gaussian distribution at truncation point, thereby preserving low-frequency modes often missed by sampling-based approximation and ensuring consistency across predictions (*GTR, Sec. 3.2*). Secondly, ATFM extends semantic-aware flow transformation by modeling semantic consistency among labels, predictions, and intermediate states during flow matching (FM), thereby enhancing prediction plausibility while promoting diversity. Additionally, Gaussian constraints are avoided for fine-grained predictions by FM (*SFM, Sec. 3.3*).

Data-Hierarchical Inference Enables Principled Disentanglement and Joint Enhancement of Accuracy and Diversity

The proposed Data-Hierarchical Inference forms disentanglement and simultaneous enhancement of prediction accuracy and diversity by marginalizing the stochasticity during diffusion, which introduces a redefinition of AMIS-specific inference paradigm. While preserving the efficiency gains from truncation, Data-Hierarchical Inference introduces a principled separation between distribution-level and sample-level inference, dedicated to optimizing overall accuracy and prediction diversity, respectively.

Specifically, the overall inference process firstly focuses on improving accuracy by supervising an accurate explicit

distribution at T_{trunc} as eq. 1, and then enhances diversity by generating varied samples from this distribution after T_{trunc} as eq. 2, ensuring each prediction remains both distinct and consistent with the underlying semantics.

$$\underbrace{\{s_1, s_2, \ldots, s_n\}}_{\text{diversity marginalized}} \in \underbrace{P \sim \mathcal{N}(\mu, \Sigma) \overset{\text{supervise}}{\Longleftrightarrow} Q \sim \mathcal{N}'(\mu', \Sigma')}_{\text{distribution level for accuracy}} \tag{1}$$

$$\underbrace{s_i \xrightarrow{\text{diffusion}} \text{pred}_i \quad \{\text{pred}_i\}_{i=1}^{n} \overset{\text{supervise}}{\Longleftrightarrow} \{\text{gt}_i\}_{i=1}^{n}}_{\text{sample level for diversity}} \tag{2}$$

where P denotes the explicit intermediate distribution estimated by our Data-Hierarchical Inference at T_{trunc}, and Q represents the corresponding distribution derived from ground truths. s_i denote latent samples, pred_i are the corresponding predictions, and gt_i represent the ground truths.

As illustrated in Fig. 3 and eq. 1, Data-Hierarchical Inference fundamentally redefines the inference paradigm by marginalizing stochasticity during truncation to achieve a principled disentanglement of accuracy and diversity. This paradigm optimization enables an improvement in accuracy without compromising diversity. Sample-level diffusion builds upon the globally aligned explicit distribution at T_{trunc}, leading to a unified and robust enhancement of both prediction fidelity and diversity.

Data-Hierarchical Inference inherently addresses the core challenges of AMIS by reconciling high prediction accuracy with plausible diversity. Through principled disentanglement of accuracy and diversity, Data-Hierarchical Inference establishes a robust and efficient solution that redefines the inference paradigm and application of TDPMs in AMIS.

Summarized Advantage: Data-Hierarchical Inference, pioneers disentangling accuracy and diversity for better meeting with requirements of AMIS, by marginalizing diversity and promoting it through controlled sampling within an inference paradigm redefinition in two consecutive stages.

Gaussian Truncation Representation Improves Fidelity via Explicit Gaussian Modeling

The proposed GTR models the explicit distribution at T_{trunc} for prediction fidelity by parameterizing it as a Gaussian distribution. This explicit modeling, supervised to ensure the overall accuracy and serving as the truncation step, enhances fidelity of predictions and reliability of distribution at T_{trunc} which preserves low-frequency modes and improves sample consistency compared to sampling-based approximations.

Theorem 1. *The marginal distribution of the latent variable at any diffusion timestep τ can be parameterized as*

$$\mathcal{N}(\mu, \Sigma), \quad \textit{with } \Sigma = DD^{\top} + L. \tag{3}$$

Theorem 2. *For any Gaussian distribution $\mathcal{N}(\mu_0, \Sigma_0)$, there exists a specific timestep τ^* at which the diffusion process produces an identical distribution.*

The proof of Theorems 1 and 2 is provided in the appendix of extended version.

According to Theorems 1 and 2, together with the controllability of diffusion trajectories between adjacent timesteps (Qiu et al. 2025), it can be concluded that arbitrary Gaussian distributions are admissible as distribution within the diffusion framework. Hence, the Gaussian distribution on the logit map modeled as eq. 4 following the formulation in Theorem 1 is selected as the truncation distribution, as it most closely approximates the predictions and enables supervision to achieve optimal accuracy and reliability.

$$Z = f_\theta(X), \mu = g_\phi(Z), \Sigma = h_\psi(Z), X_{T_{\text{trunc}}} \sim \mathcal{N}(\mu, \Sigma) \tag{4}$$

where f_θ, g_ϕ, and h_ψ denote the segmentation backbone and separate convolutional layers for estimating the mean and covariance, respectively, and $X_{T_{\text{trunc}}}$ denotes the Gaussian distribution at the truncation point.

The L_{Prior} of GTR explicitly supervises accuracy between the truncation distribution and ground truths, defined as:

$$\begin{aligned} L_{\text{Prior}} &= -\log \int p(Y|X_{T_{\text{trunc}}}) p(X_{T_{\text{trunc}}}|X) dX_{T_{\text{trunc}}} \\ &\approx \frac{1}{M} \sum_{i=1}^{M} -\log p(Y|X^i_{T_{\text{trunc}}}) \end{aligned} \tag{5}$$

where X^i_{trunc} is the i^th sample from $X_{T_{\text{trunc}}}$, Y is the ground truth, and M is the number of Monte Carlo samples. Minimizing the negative log-likelihood in L_{Prior} optimizes the explicit Gaussian X_0 for accuracy and fidelity. The network is then frozen for subsequent inference.

Summarized Advantage: GTR, pioneers explicit Gaussian distribution modeling at the truncation point for enhancing prediction fidelity and truncation distribution reliability via mean and covariance parameterization and estimation.

Segmentation Flow Matching Enhances Plausibility via Semantic Consistency Modeling

The proposed SFM extends semantic-aware flow transformation for plausibility by modeling semantic consistency among labels, predictions, and intermediate states at each timestep after T_{trunc} during FM training. It incorporates a Semantic-aware Transformation Network (ST-Net) at each timestep to ensure that flow transformation proceeds under semantic constraints. SFM aligns well with the ambiguity-resolving requirements of AMIS tasks by enhancing plausibility while promoting diversity. Moreover, by employing Flow Matching instead of DDPM, SFM avoids the Gaussian constraints that introduces disturbances in fine-grained predictions.

Algorithm 1 is the summarized training procedure of SFM.

Computing the intermediate prediction corresponding to timestep t (line 2 and 3 in Algorithm 1): The flow transformation follows an Optimal Transformation (OT) schedule (Lu and Song 2024), representing the shortest path between source and target distributions. Under the OT framework, the diffusion trajectory in the latent space forms a line segment. Therefore, we perform analytic geometry in the latent space: the intermediate state X_t at timestep t is computed by linear interpolation between the source endpoint $X_{T_{\text{trunc}}}$ and the

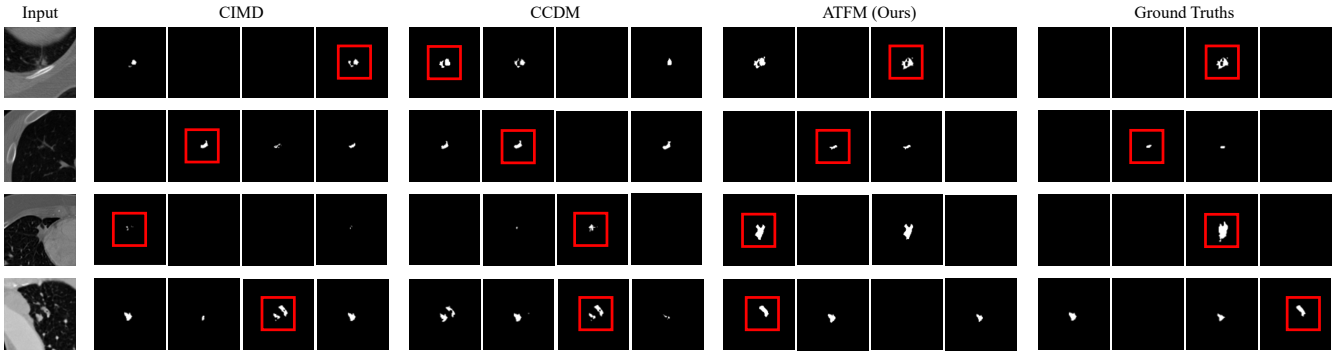


Figure 4: Comparative qualitative results on LIDC dataset among ground truths, two advanced methods and the proposed ATFM demonstrate both better alignment with ground truths and higher per-sample accuracy.

Methods	$GED_{16} \downarrow$	$GED_{32} \downarrow$	$GED_{100} \downarrow$	$HM\text{-}IoU_{32} \uparrow$	$MDM_{32} \uparrow$
Prob. Unet(Kohl et al. 2018)	$0.310_{\pm 0.010}$	$0.303_{\pm 0.010}$	$0.252_{\pm 0.004}$	$0.548_{\pm 0.000}$	$0.681_{\pm 0.020}$
PHiSeg(Baumgartner et al. 2019)	$0.262_{\pm 0.000}$	$0.247_{\pm 0.000}$	$0.224_{\pm 0.004}$	$0.595_{\pm 0.000}$	$0.704_{\pm 0.020}$
SSN(Monteiro et al. 2020)	$0.259_{\pm 0.000}$	$0.243_{\pm 0.010}$	$0.225_{\pm 0.002}$	$0.550_{\pm 0.010}$	-
MoSE(Gao et al. 2023)	$0.218_{\pm 0.003}$	$0.195_{\pm 0.002}$	$0.189_{\pm 0.002}$	$0.624_{\pm 0.004}$	$0.767_{\pm 0.004}$
P^2SAM(Li et al. 2024)	$0.208_{\pm 0.000}$	$0.206_{\pm 0.000}$	-	$0.627_{\pm 0.000}$	$0.939_{\pm 0.000}$
CIMD(Rahman et al. 2023)	-	-	$0.321_{\pm 0.000}$	$0.592_{\pm 0.002}$	$0.915_{\pm 0.004}$
AB(Chen, Zhang, and Hinton 2023)	$0.213_{\pm 0.001}$	$0.196_{\pm 0.002}$	$0.193_{\pm 0.002}$	$0.619_{\pm 0.001}$	$0.792_{\pm 0.002}$
CCDM(Zbinden et al. 2023)	$0.212_{\pm 0.001}$	$0.194_{\pm 0.001}$	$0.183_{\pm 0.002}$	$0.631_{\pm 0.002}$	$0.790_{\pm 0.003}$
ATFM(Ours)	$\mathbf{0.206_{\pm 0.002}}$	$\mathbf{0.188_{\pm 0.001}}$	$\mathbf{0.162_{\pm 0.002}}$	$\mathbf{0.667_{\pm 0.002}}$	$\mathbf{0.948_{\pm 0.001}}$

Table 1. Quantitative results on LIDC dataset show the superior performance of ATFM. Bold represents the best per column. Arrows indicate the increasing performance of the metrics.

Algorithm 1: Training Procedure for proposed SFM

Require: Source distribution $X_{T_{\text{trunc}}}$, Target distribution X_1, Output of ST-net $g_\theta(X_t)$ at timestep t, ground truths y_i

1: **repeat**
2: $X_t = t \times X_1 + (\mathbf{1} - t) \times X_{T_{\text{trunc}}}$
3: $x_{1t} = x_t + g_\theta(X_t) \times (\mathbf{1} - t)$ (*calculation of prediction corresponding to timestep t*)
4: $L^i_{\text{Dice}} = 1 - Dice(x_{1t}, y_i), i = 1, 2, \ldots, N$
5: $L_{\text{SF}} = L_{\text{FM}} + \frac{1}{N} \sum_{i=1}^{N} \alpha \times L^i_{\text{Dice}}$ (*semantic consistency modeling*)
6: $\theta \leftarrow \theta - \eta \nabla_\theta L_{\text{SF}}$ (*gradient update*)
7: **until** $||\nabla_\theta L_{\text{SF}}|| < \delta$ (*convergence*)

target endpoint X_1. Then, using the direction vector of the segment $g_t(X_t)$ and the position of X_t, a predicted result x_{1t} is derived by projecting along the diffusion trajectory starting from x_t.

Semantic consistency modeling (line 4 to 6 in Algorithm 1): By computing the Dice loss between the predicted result x_{1t} and all ground truth annotations, semantic consistency at timestep t can be explicitly modeled. This supervision acts as an auxiliary constraint to the Flow Matching loss, encouraging the transformation to preserve plausibility and consistency throughout the diffusion process for diversity.

The aforementioned SFM training process not only ensures accurate flow transformation, but also explicitly models the semantic consistency among the state, predicted result and ground truths at each timestep t. This dual-objective optimization enhances the semantic plausibility of predictions, and simultaneously capturing diverse sample-level variations via flow matching, positioning SFM as an indispensable module of the proposed ATFM framework for AMIS.

Summarized Advantage: SFM, firstly extends semantic-aware flow transformation for prediction plausibility while enhancing diversity, by modeling segmentation semantic consistency among ground truths, predictions and intermediate states at each timestep.

Experiments

Experimental Setup

Datasets. In our experiments, we applied two public datasets for ambiguous medical image segmentation: LIDC-IDRI (Kalpathy-Cramer et al. 2016) and ISIC3 subset (Codella et al. 2019; Zepf et al. 2023). The LIDC-IDRI dataset consists of lung CT scans with multiple expert-annotated lesion segmentations, highlighting diagnostic ambiguities. Following preprocessing as described in (Kohl et al. 2018, 2019), the dataset includes 15,096 slices, each with four corresponding segmentation labels. The ISIC3 dataset provides dermoscopic images of skin lesions with annotations for lesion boundaries. Using the preprocessed

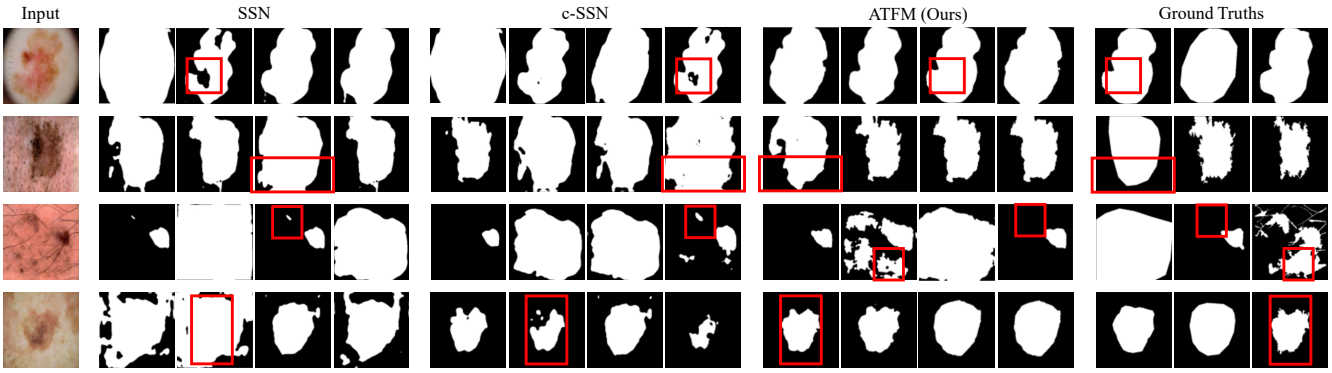


Figure 5: Comparative qualitative results on ISIC3 subset dataset among ground truths, two advanced methods and the proposed ATFM demonstrate both better alignment with ground truths and higher per-sample accuracy.

Methods	$GED_{16} \downarrow$	$GED_{32} \downarrow$	$GED_{100} \downarrow$	$HM\text{-}IoU_{32} \uparrow$	$MDM_{32} \uparrow$
Prob. Unet(Kohl et al. 2018)	$0.202_{\pm 0.003}$	$0.187_{\pm 0.003}$	$0.171_{\pm 0.002}$	$0.697_{\pm 0.005}$	$0.927_{\pm 0.003}$
SSN(Monteiro et al. 2020)	$0.197_{\pm 0.007}$	$0.181_{\pm 0.004}$	$0.167_{\pm 0.002}$	$0.700_{\pm 0.004}$	$0.939_{\pm 0.001}$
c-Prob. Unet(Zepf et al. 2023)	$0.208_{\pm 0.004}$	$0.202_{\pm 0.005}$	$0.179_{\pm 0.002}$	$0.719_{\pm 0.004}$	$0.925_{\pm 0.004}$
c-SSN(Zepf et al. 2023)	$0.200_{\pm 0.007}$	$0.195_{\pm 0.007}$	$0.177_{\pm 0.001}$	$0.725_{\pm 0.004}$	$0.931_{\pm 0.003}$
ATFM(Ours)	$\mathbf{0.183_{\pm 0.001}}$	$\mathbf{0.152_{\pm 0.002}}$	$\mathbf{0.147_{\pm 0.003}}$	$\mathbf{0.732_{\pm 0.003}}$	$\mathbf{0.942_{\pm 0.002}}$

Table 2. Quantitative results on ISIC3 subset dataset show the superiority of ATFM. Bold represents the best per column. Arrows indicate the increasing performance of the metrics.

Models	Inference Steps and $Time_{100}$
CIMD	$T = 100$ steps $\approx$ 420s
AB	$T = 250$ steps $\approx$ 1050s
CCDM	$T = 250$ steps $\approx$ 1100s
ATFM	GTR + ($T_{\text{Trunc}} = 25$ steps) $\approx$ **113s**

Table 3. Time comparison for generating 100 samples on the LIDC dataset for diffusion-based methods demonstrates the superior time efficiency of proposed ATFM.

Models	$GED_{100} \downarrow$	$HM\text{-}IoU_{32} \uparrow$
Act. GTR	$0.230_{\pm 0.001}$	$0.550_{\pm 0.010}$
SFM w/o L_{SF}	$0.185_{\pm 0.002}$	$0.624_{\pm 0.002}$
SFM	$0.176_{\pm 0.002}$	$0.631_{\pm 0.002}$
ATFM w/o L_{SF}	$0.249_{\pm 0.001}$	$0.597_{\pm 0.003}$
ATFM	$\mathbf{0.162_{\pm 0.002}}$	$\mathbf{0.667_{\pm 0.002}}$

Table 4. Ablation Study on LIDC dataset shows the validity of all components in proposed ATFM.

ISIC3 subset from (Zepf et al. 2023), we work with 300 images, each featuring exactly three annotations.

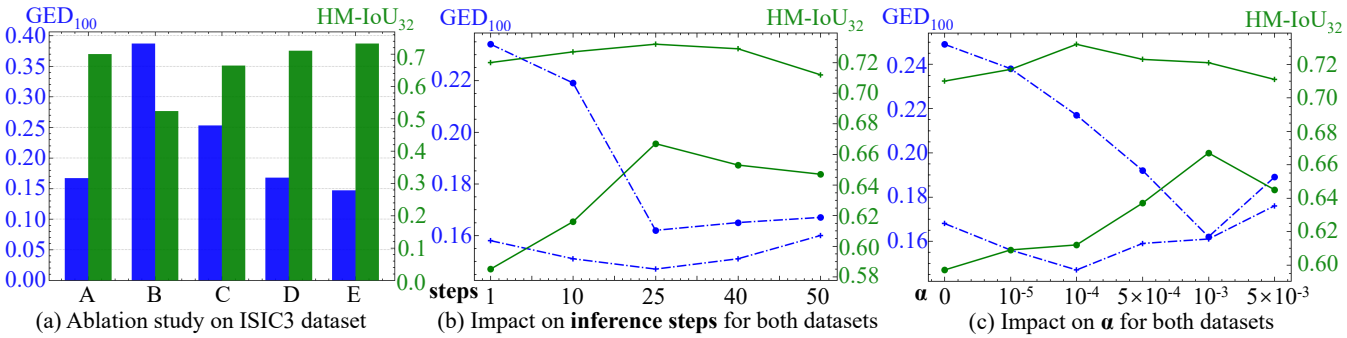
Implementation Details. All training and inference procedures are conducted on a single RTX 3090 GPU with 24GB memory. SFM in ATFM is trained for 200 epochs on LIDC with GTR pretrained for 1000 epochs, and for 120 epochs on ISIC3 with a 400-epoch GTR. We set $\lambda = 10^{-3}$ (i.e. $T = 1000$) with a linear schedule for all experiments. Both GTR and ST-Net of the SFM are optimized using an Adam optimizer (Kingma and Ba 2014) with a learning rate of 10^{-4}. Hyper-parameter M is set to 20 following (Zepf et al. 2023) and α is set to 10^{-3} and 10^{-4} for LIDC and ISIC3 respectively according to hyper-parameter studies.

Evaluation Metrics. Three metrics are utilized for comprehensive evaluation from three aspects: For segmentation distribution, we utilize the Generalised Energy Distance (GED) (Bellemare et al. 2017) to evaluate the alignment among distribution of predictions and ground-truths. For sample fidelity, we utilize the Hungarian-Matching Intersection-over-Union (HM-IoU) (Gao et al. 2023) to provide an accurate representation of the performance on segmentation across all predictions. For individual segmentation accuracy, we utilize the Maximum Dice Matching (MDM) (Rahman et al. 2023) to evaluate the best Dice scores between each prediction result and each ground truth. We denote the metrics with subscript n to represent the metrics calculated using n samples. Results are reported as mean ± standard deviation over five independent runs.

Experimental Results

Comparison with SOTAs for Performance Superiority.

Quantitative Evaluation. Table 1 and Table 2 report the quantitative results on the LIDC and ISIC3 datasets, respectively. Across all key metrics—GED (for diversity), HM-IoU (for sample fidelity), and MDM (for individual



(a) Ablation study on ISIC3 dataset (b) Impact on **inference steps** for both datasets (c) Impact on **α** for both datasets

Figure 6: Comprehensive analysis through ablation and hyper-parameter studies demonstrates the effectiveness of innovations and training configurations in proposed ATFM.

accuracy)—ATFM consistently outperforms state-of-the-art methods. Notably, ATFM achieves an 11.5% improvement on GED_{100} in LIDC, along with consistent gains in GED_{16} and GED_{32} compared with the runner-up method, highlighting its ability to better capture the underlying segmentation label distribution. A minimum of 7.3% improvement in HM-IoU_{32} further demonstrates the superior fidelity of the generated samples. ATFM also leads in the MDM_{32} metric, validating its accuracy at the individual prediction level. Similar trends are observed on ISIC3, where ATFM outperforms the runner-up method by 12% in GED while also achieving top-tier results in HM-IoU and MDM, demonstrating comprehensive improvements in diversity, fidelity, and accuracy. These results showcase the effectiveness of the Data-Hierarchical Inference framework of ATFM, which explicitly models intermediate distributions and enforces semantic-aware flow transformations to enhance accuracy and diversity in AMIS.

Qualitative Results. Fig. 4 and Fig. 5 further compare visual results among advanced methods and proposed ATFM for LIDC dataset and ISIC3 dataset, respectively. Predictions from ATFM more faithfully reflect the range of plausible annotations and better preserve fine-grained structures, indicating its superior alignment with ambiguous and detailed ground truths, which is an essential and necessary requirement in AMIS tasks.

Inference Efficiency. Table 3 reports the total inference time for generating 100 plausible predictions. ATFM retains the inference efficiency advantage of truncated diffusion models, requiring only 25 diffusion steps and an estimation of truncation point (approx. 113s). Compared to other diffusion-based methods, it achieves both superior segmentation performance and significantly faster inference, highlighting its practicality and scalability in AMIS applications.

Ablation Studies Demonstrate Effectiveness of Innovations. Table 4 and Fig. 6(a) show the conducted ablation study on both datasets evaluating five model variants: GTR with activation layers (A), SFM with and without L_{SF} (B and C), and ATFM with and without L_{SF} (D and E). ATFM outperformed both Act. GTR and SFM by a minimum of 10% and 6% on both metrics, highlighting the benefit of Data-Hierarchical Inference and the high effectiveness and fidelity of proposed ATFM provided by GTR. The performance gap of an average of 11% between models with and without L_{SF} emphasizes the importance of semantics consistency modeling, underscoring the role of L_{FM} and SFM in preserving plausibility when enhancing diversity.

Hyper-parameter Studies Prove Effectiveness of Training Configurations. Fig. 6(b) and 6(c) illustrate the effects of inference step count in SFM and α in L_{SF} in SFM, respectively. Five values between 1 and 50 are set for inference steps based on the property of Euler Sampler (Song and Ermon 2020; Song et al. 2020). Optimal performance was achieved within 25 steps, offering a good balance between performance and efficiency. For α, values between 0 and 5×10^{-3} were tested. Small α values limits L_{SF}'s impact, while α with too large values diminishes the effect of L_{FM}. Values of α striking a balance were chosen as the final setting.

Conclusion

In this work, we proposed Ambiguity-aware Truncated Flow Matching (ATFM), addressing the challenge of jointly improving prediction accuracy and diversity via a novel inference paradigm and dedicated model components, tailored to the demands of AMIS tasks. Specifically, we firstly proposed Data-Hierarchical Inference, redefining a novel inference paradigm that disentangles prediction accuracy and diversity, which supervises a distribution for accuracy at T_{trunc} by marginalizing stochasticity and promoting diversity through controlled sampling in the following timesteps. On this foundation, we designed two key modules for ATFM: GTR, which explicitly models the Gaussian distribution at truncation point to ensure prediction fidelity and truncation distribution reliability for overall accuracy; and SFM, which extends semantic-aware flow transformation to model semantic consistency across predictions, annotations and intermediate states for enhancing prediction plausibility while promoting diversity. Experimental results on two public datasets showed that the proposed ATFM outperforms SOTA methods across all metrics and offers a more efficient inference process simultaneously. ATFM offers a versatile and reliable solution for AMIS across a broader range of scenarios through multifaceted analysis and outstanding performance.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 62501195, 62272135, 62372135 and 82527807, and the Key Research & Development Program of Heilongjiang Province under Grants 2023ZX01A08 and 2024ZX12C23, and the Natural Science Foundation of Heilongjiang Province under Grant LH2024F019.

References

Baumgartner, C. F.; Tezcan, K. C.; Chaitanya, K.; Hötker, A. M.; Muehlematter, U. J.; Schawkat, K.; Becker, A. S.; Donati, O.; and Konukoglu, E. 2019. Phiseg: Capturing uncertainty in medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, 119–127. Springer.

Bellemare, M. G.; Danihelka, I.; Dabney, W.; Mohamed, S.; Lakshminarayanan, B.; Hoyer, S.; and Munos, R. 2017. The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*.

Chavhan, G. B.; Parra, D. A.; Oudjhane, K.; Miller, S. F.; Babyn, P. S.; and Salle, F. L. P. 2008. Imaging of ambiguous genitalia: Classification and diagnostic approach1. *Radiographics*.

Chen, T.; Zhang, R.; and Hinton, G. 2023. Analog Bits: Generating Discrete Data using Diffusion Models with Self-Conditioning. In *The Eleventh International Conference on Learning Representations*.

Codella, N.; Rotemberg, V.; Tschandl, P.; Celebi, M. E.; Dusza, S.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M.; et al. 2019. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*.

Dong, S.; Cai, Z.; Hangel, G.; Bogner, W.; Widhalm, G.; Huang, Y.; Liang, Q.; You, C.; Kumaragamage, C.; Fulbright, R. K.; et al. 2025. A Flow-based Truncated Denoising Diffusion Model for super-resolution Magnetic Resonance Spectroscopic Imaging. *Medical Image Analysis*, 99: 103358.

Gao, Z.; Chen, Y.; Zhang, C.; and He, X. 2023. Modeling Multimodal Aleatoric Uncertainty in Segmentation with Mixture of Stochastic Experts. In *The Eleventh International Conference on Learning Representations*.

He, J.; Li, Y.; Yuan, Q.; et al. 2023. Tdiffde: A truncated diffusion model for remote sensing hyperspectral image denoising. *arXiv preprint arXiv:2311.13622*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hong, S.; Bonkhoff, A. K.; Hoopes, A.; Bretzner, M.; Schirmer, M. D.; Giese, A.-K.; Dalca, A. V.; Golland, P.; and Rost, N. S. 2021. Hypernet-ensemble learning of segmentation probability for medical image segmentation with ambiguous labels. *arXiv preprint arXiv:2112.06693*.

Ji, W.; Yu, S.; Wu, J.; Ma, K.; Bian, C.; Bi, Q.; Li, J.; Liu, H.; Cheng, L.; and Zheng, Y. 2021. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12341–12351.

Kalpathy-Cramer, J.; Zhao, B.; Goldgof, D.; Gu, Y.; Wang, X.; Yang, H.; Tan, Y.; Gillies, R.; and Napel, S. 2016. A comparison of lung nodule segmentation algorithms: methods and results from a multi-institutional study. *Journal of digital imaging*, 29: 476–487.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kohl, S.; Romera-Paredes, B.; Meyer, C.; De Fauw, J.; Ledsam, J. R.; Maier-Hein, K.; Eslami, S.; Jimenez Rezende, D.; and Ronneberger, O. 2018. A probabilistic u-net for segmentation of ambiguous images. *Advances in neural information processing systems*, 31.

Kohl, S. A.; Romera-Paredes, B.; Maier-Hein, K. H.; Rezende, D. J.; Eslami, S.; Kohli, P.; Zisserman, A.; and Ronneberger, O. 2019. A hierarchical probabilistic u-net for modeling multi-scale ambiguities. *arXiv preprint arXiv:1905.13077*.

Li, C.; Lin, Z.; Liu, H.; Liu, Y.; Huang, Y.; Ding, X.; Tu, X.; Yuan, Y.; et al. 2024. Pˆ 2SAM: Probabilistically Prompted SAMs Are Efficient Segmentator for Ambiguous Medical Images. In *ACM Multimedia*.

Liao, B.; Chen, S.; Yin, H.; Jiang, B.; Wang, C.; Yan, S.; Zhang, X.; Li, X.; Zhang, Y.; Zhang, Q.; et al. 2025. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12037–12047.

Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2023. Flow Matching for Generative Modeling. In *The Eleventh International Conference on Learning Representations*.

Lu, C.; and Song, Y. 2024. Simplifying, Stabilizing and Scaling Continuous-Time Consistency Models. *arXiv preprint arXiv:2410.11081*.

Monteiro, M.; Le Folgoc, L.; Coelho de Castro, D.; Pawlowski, N.; Marques, B.; Kamnitsas, K.; van der Wilk, M.; and Glocker, B. 2020. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *Advances in neural information processing systems*, 33: 12756–12767.

Qiu, X.; Yang, M.; Ma, X.; Li, F.; Liang, D.; Luo, G.; Wang, W.; Wang, K.; and Li, S. 2025. Finding Local Diffusion Schrodinger Bridge using Kolmogorov-Arnold Network. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 23227–23236.

Rahman, A.; Valanarasu, J. M. J.; Hacihaliloglu, I.; and Patel, V. M. 2023. Ambiguous medical image segmentation using diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11536–11546.

Song, Y.; and Ermon, S. 2020. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33: 12438–12448.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Zbinden, L.; Doorenbos, L.; Pissas, T.; Huber, A. T.; Sznitman, R.; and Márquez-Neila, P. 2023. Stochastic segmentation with conditional categorical diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1119–1129.

Zepf, K. M.; Petersen, E. W.; Frellsen, J.; and Feragen, A. 2023. That Label's got Style: Handling Label Style Bias for Uncertain Image Segmentation. In *Eleventh International Conference on Learning Representations*.

Zhang, W.; Zhang, X.; Huang, S.; Lu, Y.; and Wang, K. 2022. A probabilistic model for controlling diversity and accuracy of ambiguous medical image segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4751–4759.

Zheng, H.; He, P.; Chen, W.; and Zhou, M. 2022. Truncated diffusion probabilistic models. *arXiv preprint arXiv:2202.09671*, 1(3.1): 2.