AdKDD 2026 | KDD Workshop, Jeju, South Korea

RoLGaL

Robust Learning for CVR Prediction via Guardrail-Guided Labeling

Xiaoli Gao, Qiuping Xu (presenter), Amir Roshankar, Rong Jin

Meta

TL;DR

The Challenge

- Production CVR models are retrained daily on fresh data.
- Noisy conversion labels can overwrite previously learned knowledge, leading to **catastrophic forgetting**.

Our Solution

- Learn **robust representations** using **guardrail-guided corrected labels**.
- Keep the **serving prediction head unchanged** to preserve calibration.
- Remove the auxiliary branch at inference, resulting in **zero latency overhead**.

Impact

Offline	Production
+0.25% AUC on Criteo	**+0.23%** lift in a primary revenue-related business metric
+0.31% AUC under noisy labels	Deployed to **10+ Meta production CVR/CTR models,** more to come

Takeaway: RoLGaL makes continual CVR training more robust without changing the serving model or adding any inference cost.

Problem

Daily model refresh

Adapt to new users, advertisers, and market shifts

↓

Noisy conversion labels

Delayed attribution · accidental clicks · logging artifacts

↓

Naïve retraining

Recent noisy gradients dominate the update

↓

Catastrophic forgetting

Stable historical preferences can be overwritten

Production constraints

✓ Minimize impact on the model's existing calibration

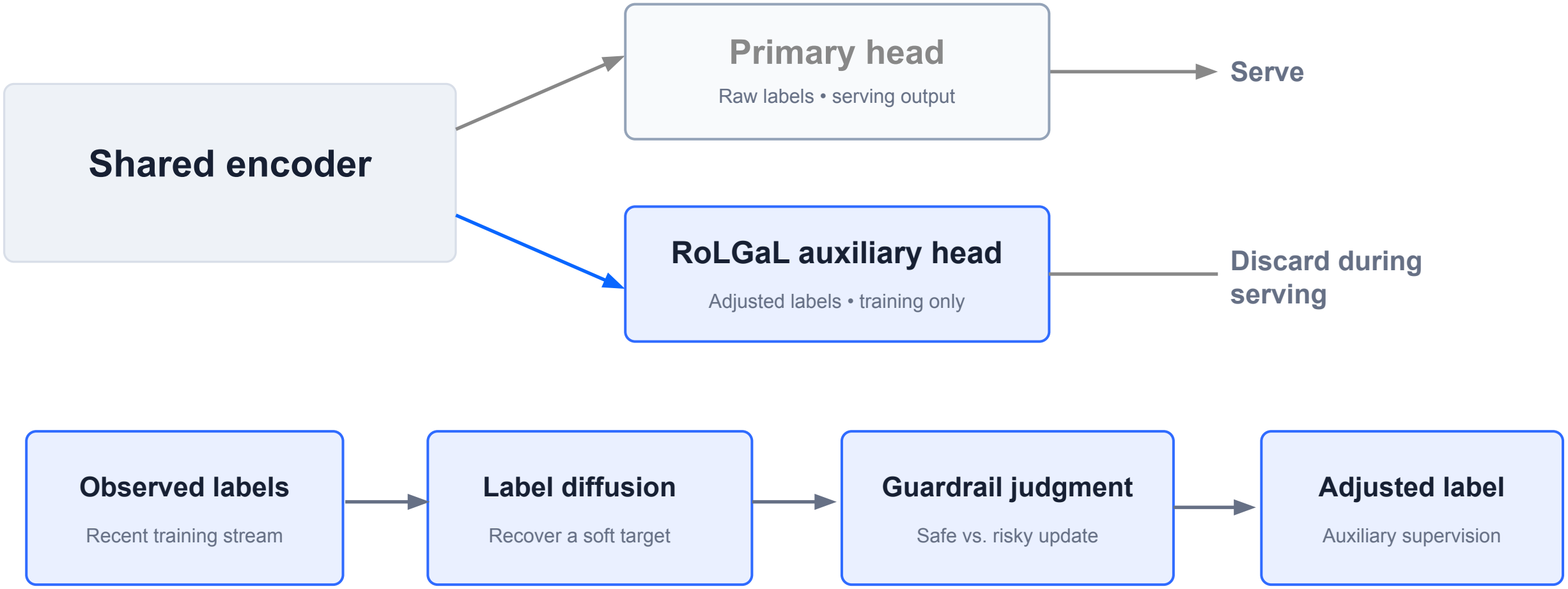
CVR estimates drive auction bidding; label replacement can shift the serving distribution.

✓ Minimize inference overhead

Any robustness layer must not slow down real-time serving.

Key question: Can we improve robustness without changing the serving prediction?

RoLGaL: Guardrail-Guided Optimization and Off-Path Serving



RoLGaL: Adaptive Label Adjustment and Multiple-Head

- **Label diffusion**

Inspired by diffusion-style denoising, it uses a lightweight one-step probabilistic relaxation: an observation model that accounts for label noise, followed by Bayesian recovery of a soft target

$$y_i^* = P(\tilde{y}_i = 1 \mid y_i^{obs}) = \frac{1}{1 + \left(\frac{1-p_i}{p_i}\right) \exp\left(\frac{1-2y_i^{obs}}{2\sigma^2}\right)},$$

- **Guardrail-guided judgment**

Compare diffused label y^* with guardrail label y^g to access the risk score r of the label

$$r_i = I\left(\mathrm{Dist}(y_i^*, y_i^g) > \tau\right).$$

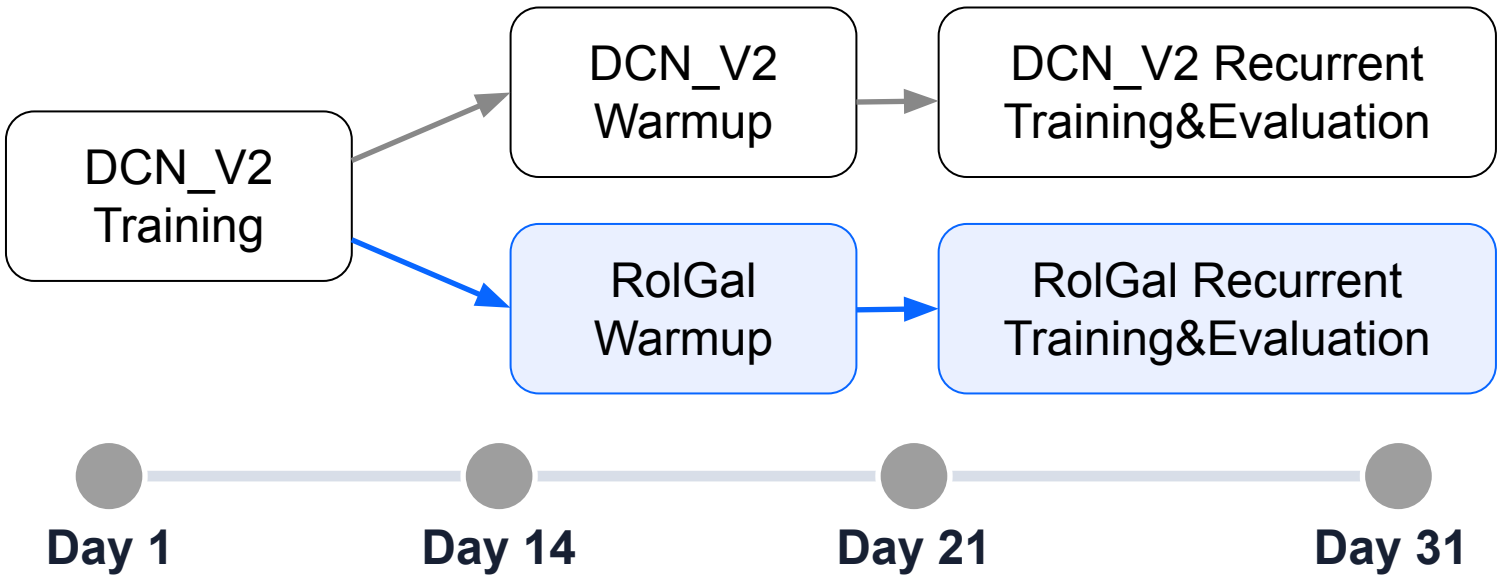
- **Adaptive label adjustment**

Leverage risk score r to generate final adjust label y^a. c is parameter to control the reliability of the guardrail label y^g

$$y_i^a = (1 - r_i) \cdot y_i^* + r_i \cdot \left(c_i y_i^g + (1 - c_i) y_i^*\right),$$

Experiment: Setup

- Criteo: 16.5M+ chronologically ordered interactions spanning 31 days
- DCN-v2 used as the backbone for all methods
- 14-day warm-up, 7-day training, and 10-day streaming evaluation
- The previous checkpoint generates guardrail predictions before each daily update



Experiment: Parity with KD

- Compared against Label Smoothing and Knowledge Distillation

Table 1: Main Results on Criteo. We report Mean ± Std and relative lift from the baseline across 3 trials.

Method	AUC (↑)	LogLoss (↓)
Baseline	0.9027 ± 0.0002	0.12563 ± 0.00013
Label Smoothing	0.9027 ± 0.0003 (+0.00%)	0.12556 ± 0.00008 (-0.08%)
Knowledge Distillation	0.9047 ± 0.0003 (+0.22%)	**0.12488 ± 0.00025 (-0.60%)**
Singlehead RoLGaL	0.9048 ± 0.0003 (+0.23%)	0.12506 ± 0.00017 (-0.46%)
Multihead RoLGaL	**0.9050 ± 0.0003 (+0.25%)**	0.12494 ± 0.00021 (-0.55%)

Experiment: Robust to Label Noise

- More robust than LS and KD under corrupted training labels

Table 2: Robustness under Injected Noise ($\eta = 0.01$). RoLGaL outperforms LS and KD under noisy supervision.

Method	AUC (↑)	Lift
Baseline	0.8989 ± 0.0002	–
Label Smoothing	0.8988 ± 0.0001	-0.01%
Knowledge Distillation	0.9006 ± 0.0002	+0.19%
Multihead RoLGaL	**0.9016 ± 0.0001**	**+0.31%**

- Relative AUC lift increases from +0.31% at η = 0.01 to +0.43% at η = 0.10. Maintains strong performance as injected label noise becomes more severe

Table 3: Robustness under Injected Label Noise.

Noise (η)	Baseline	Singlehead RoLGaL ($K = 1$)	Multihead RoLGaL ($K = 2$)
0.01	0.8989 ± 0.0002	0.9013 ± 0.0002 (+0.27%)	**0.9016 ± 0.0001 (+0.31%)**
0.05	0.8900 ± 0.0002	0.8930 ± 0.0002 (+0.34%)	**0.8931 ± 0.0001 (+0.34%)**
0.10	0.8814 ± 0.0003	0.8850 ± 0.0003 (+0.41%)	**0.8852 ± 0.0004 (+0.43%)**

Experiment: Ablation

- Selective guardrail-based label adjustment is the most important component in this setup
- Other components introduce no measurable degradation and provide complementary robustness
- In practice, performance depends mainly on guardrail selection and the choice of RoLGaL noise and risk thresholds

Table 4: Component Ablation ($\eta = 0.01$). We report AUC Mean ± Std and relative lift.

Configuration	AUC (Mean ± Std)	Lift (Δ%)
A. Multi-Head Scaling (Full RoLGaL)		
Baseline	0.8989 ± 0.0002	0.00%
RoLGaL ($K = 1$)	0.9013 ± 0.0002	+0.27%
RoLGaL ($K = 2$)	**0.9016 ± 0.00005**	**+0.31%**
RoLGaL ($K = 3$)	0.9016 ± 0.00008	+0.30%
B. Module Contribution (Fixed $K = 2$)		
No Guardrail	0.8999 ± 0.0001	+0.12%
No Label Diffusion	0.9015 ± 0.0001	+0.29%
Full RoLGaL	**0.9016 ± 0.00005**	**+0.31%**

Experiment: Production Impact

+0.23%

statistically significant lift in a primary revenue-related business metric

10+

launched to 10+ models

0

inference latency overhead

Prototype

Auxiliary-task integration

Offline validation

NE performance

Standalone A/B validation

Revenue-related lift

Bundle A/B tests&rollouts

10+ models

Takeaways

01. Robust Training

Robust continual CVR training under noisy labels

02. Calibration

Preserves existing calibration with minimal disruption

03. Efficiency

Delivers online gains with zero inference overhead

Thank you & Questions?

Contact: xqiuping@meta.com