PROBABILISTIC MULTIMODAL REPRESENTATION LEARNING VIA PRODUCT-OF-EXPERTS AND QUANTUM TENSOR NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal representation learning must fuse heterogeneous signals while accounting for differences in modality reliability. Deterministic fusion typically does not represent this uncertainty, whereas generative fusion can be constrained by compact latent spaces and posterior collapse. We introduce QUTA (Quantum Tensor Autoencoder), which combines a Product-of-Experts posterior with a quantum Tree Tensor Network (TTN). The PoE forms a shared probabilistic latent from modality-specific Gaussian experts, and a 16-qubit TTN transforms it into nonlinear quantum features using only 75 circuit parameters. A classical skip and an objective combining task, reconstruction, and KL losses make the hybrid model trainable end-to-end. Because fusion is a Product of Experts, the same trained model can run with missing or noisy test modalities, which we verify on CH-SIMS. The architecture defines a deployment path from local simulator-based training to remote quantum-processor inference. We further evaluate the frozen circuit with noisy density-matrix simulation and observe stable predictive performance. Across three benchmarks, QUTA achieves the best result in the matched-protocol CH-SIMS comparison and remains competitive on MOSI and MELD, while using substantially fewer trainable parameters than fine-tuned-BERT systems.

1 INTRODUCTION

Multimodal learning jointly derives representations and predictions from heterogeneous sources such as language, vision, and audio, on the premise that complementary modalities encode information unavailable from any single channel (Ngiam et al., 2011; Baltrušaitis et al., 2019). Representative discriminative methods, such as the Tensor Fusion Network (TFN) (Zadeh et al., 2017), Low-rank Multimodal Fusion (LMF) (Liu et al., 2018), and the Multimodal Transformer (MulT) (Tsai et al., 2019), combine modality-specific features and model cross-modal interactions to learn joint representations for downstream prediction. Their fusion mechanisms primarily emphasize predictive interactions, however, without explicitly representing modality-specific uncertainty. When modalities are missing, corrupted by noise, or uneven in quality, this lack of explicit uncertainty modeling can limit the reliability of the fused representation and poses a challenge to robust prediction (Ma et al., 2022).

Generative approaches offer an alternative by modeling multimodal observations through shared probabilistic latent variables. Variational autoencoders (Kingma & Welling, 2014) with product- or mixture-of-experts posteriors (Wu & Goodman, 2018; Shi et al., 2019) provide a framework for representing latent uncertainty and inferring a shared code from available modalities. In particular, Gaussian product-of-experts fusion weights modality contributions by their learned precision. Yet probabilistic fusion alone does not ensure an expressive downstream representation: compact latent spaces can constrain the modeling of complex cross-modal relationships, while variational training remains susceptible to posterior collapse (Lucas et al., 2019). Increasing latent dimensionality or network capacity to enrich these representations can introduce additional parameter and computational costs. This motivates enhancing the nonlinear expressiveness of probabilistic multimodal representations under a limited parameter budget.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Quantum machine learning (QML) offers a promising direction for enhancing representations under a limited parameter budget. By exploiting superposition and entanglement, quantum models encode data into a 2^n-dimensional Hilbert space using n qubits; data encoding, parameterized circuits, and measurement together yield nonlinear feature maps (Schuld & Killoran, 2019; Havlíček et al., 2019). However, devices in the noisy intermediate-scale quantum (NISQ) era remain constrained by qubit counts, noise, and feasible circuit depth (Preskill, 2018). Hybrid quantum–classical architectures, particularly variational quantum algorithms (VQAs), combine classical feature processing and optimization with shallow parameterized quantum circuits, providing a practical route to exploring the representational potential of quantum systems under near-term hardware constraints (Cerezo et al., 2021; Peruzzo et al., 2014).

Motivated by these limitations, we propose QuTA (Quantum Tensor Autoencoder), a generative framework for probabilistic multimodal representation learning. Compared with conventional discriminative fusion, QuTA incorporates generative uncertainty modeling to accommodate missing modalities and noisy inputs. We modify the autoencoder architecture by integrating a quantum Tree Tensor Network (TTN) into its latent representation pathway. This module maps the compact multimodal latent code into a high-dimensional quantum feature space and extracts nonlinear features, enriching the representation of complex cross-modal relationships with only $\mathcal{O}(n)$ trainable circuit parameters for n qubits. The resulting hybrid architecture combines probabilistic handling of imperfect inputs with expressive feature transformation and a compact trainable model. Our contributions are threefold:

- We introduce QuTA, a framework that integrates PoE-based probabilistic multimodal fusion with quantum latent transformation. It models modality-dependent uncertainty and supports inference from available modalities, while enriching the shared latent representation through a nonlinear quantum feature map.
- We employ a quantum TTN as a compact latent transformation module, with $5(n-1)$ trainable circuit parameters and $\mathcal{O}(\log n)$ parallel depth for n qubits. Our 16-qubit instance uses only 75 circuit parameters, providing a structured approach to nonlinear representation learning under a limited circuit parameter budget.
- We evaluate predictive performance and parameter requirements on MOSI, MELD, and CH-SIMS using 0.59–1.08M trainable parameters, excluding frozen feature extractors. The model achieves the highest mean Acc-3 in the matched-protocol CH-SIMS comparison. Module ablations, test-time modality removal and input corruption, and controlled gate-noise simulations assess component effects and model behavior under distinct sources of perturbation.

2 Background

2.1 Probabilistic Multimodal Latents

The variational autoencoder (VAE) learns a latent code through an encoder $q_\phi(z \mid x)$ and decoder $p_\theta(x \mid z)$ by maximizing the evidence lower bound (ELBO) (Kingma & Welling, 2014; Rezende et al., 2014),

$$\log p(x) \;\geq\; \mathbb{E}_{q_\phi(z|x)}\big[\log p_\theta(x \mid z)\big] \;-\; D_{\mathrm{KL}}\big(q_\phi(z \mid x) \,\|\, p(z)\big), \tag{1}$$

with $p(z) = \mathcal{N}(0, I)$. The reparameterization $z = \mu + \varepsilon \odot \sigma$, $\varepsilon \sim \mathcal{N}(0, I)$, permits gradient-based learning. For multimodal inputs, a Product-of-Experts (PoE) multiplies unimodal posteriors over the same latent (Hinton, 2002; Wu & Goodman, 2018). Gaussian products remain Gaussian: their precisions add, so an expert's learned inverse variance controls its coordinate-wise contribution. Unavailable experts can simply be omitted. Section 4.1 gives the fusion equations used by QuTA.

2.2 Variational Quantum Models

A parameterized quantum model encodes classical input x into an n-qubit state $|\psi(x)\rangle$, applies a unitary circuit $U(\theta)$, and returns observable expectations

$$f_\theta(x) = \langle \psi(x) | U^\dagger(\theta) O U(\theta) | \psi(x) \rangle. \tag{2}$$

The circuit and a classical optimizer form a variational quantum algorithm (VQA) (Cerezo et al., 2021; Benedetti et al., 2019). For the rotation generators used here, the parameter-shift rule evaluates derivatives through two circuit calls,

$$\frac{\partial f}{\partial \theta_k} = \frac{1}{2}\left[f(\theta_k + \tfrac{\pi}{2}) - f(\theta_k - \tfrac{\pi}{2})\right], \tag{3}$$

with the loss gradient obtained by the chain rule (Mitarai et al., 2018; Schuld et al., 2019). Simulator-based optimization is standard in model-level QML because it isolates architecture and optimization before scarce hardware access (Cerezo et al., 2021). Real NISQ devices additionally introduce sampling, gate, readout, and decoherence effects (Preskill, 2018; Wang et al., 2021). We therefore train locally and separately test the frozen circuit with density-matrix gate noise before considering QPU inference.

3 Related Works

Discriminative fusion captures cross-modal interactions through tensor products or low-rank factors (TFN, LMF) and cross-modal attention (MulT, CLIP) (Zadeh et al., 2017; Liu et al., 2018; Tsai et al., 2019; Radford et al., 2021). Probabilistic alternatives instead infer a shared distribution: MVAE multiplies unimodal Gaussian experts (Wu & Goodman, 2018), MMVAE uses modality-specific mixture inference (Shi et al., 2019), and MoPoE mixes products to address subset sampling bias (Sutter et al., 2021). These models expose latent uncertainty but remain vulnerable to the capacity and optimization limits of variational latents (Lucas et al., 2019). QuTA retains closed-form PoE fusion and adds a nonlinear quantum transformation.

Quantum-related multimodal work includes density-matrix fusion (Li et al., 2021), classical relational tensor networks (Sahay et al., 2018), QTN-VQC (Qi et al., 2021), and quantum-jump sentiment models (Chen et al., 2025). Much of it is quantum-inspired or uses tensor networks as classical fusion modules. In contrast, our hardware-executable variational TTN acts after a probabilistic PoE latent, separating uncertainty-aware fusion from the quantum nonlinear transformation.

4 Method

Figure 1 summarizes QuTA: modality-wise Gaussian experts form a shared PoE latent, which is transformed by a quantum Tree Tensor Network (TTN) and a classical readout; auxiliary decoders reconstruct pooled features.

4.1 Probabilistic Multimodal Fusion

For the available modalities $X_{\mathcal{S}} = \{x_m : m \in \mathcal{S}\}$, each encoder g_m parameterizes a d-dimensional diagonal Gaussian,

$$q_m(z \mid x_m) \;=\; \mathcal{N}\big(z;\, \mu_m,\, \Sigma_m\big), \qquad \Sigma_m \;=\; \mathrm{diag}(\sigma_m^2), \qquad (\mu_m,\, \log \sigma_m^2) \;=\; g_m(x_m), \tag{4}$$

where $d = 64$. Their normalized product gives the joint posterior

$$q_\phi(z \mid X_{\mathcal{S}}) \;\propto\; \prod_{m \in \mathcal{S}} q_m(z \mid x_m). \tag{5}$$

with closed-form precision, covariance, and mean

$$\Lambda_{\mathcal{S}} = \sum_{m \in \mathcal{S}} \Lambda_m, \qquad \Sigma_z = \Lambda_{\mathcal{S}}^{-1}, \qquad \mu_z = \Sigma_z \sum_{m \in \mathcal{S}} \Lambda_m \mu_m. \tag{6}$$

Thus, predicted precision weights each latent coordinate; missing modalities are handled by excluding their factors from $\mathcal{S}$. Training samples $z = \mu_z + \varepsilon \odot \sigma_z$, $\varepsilon \sim \mathcal{N}(0, I)$, by reparameterization (Kingma & Welling, 2014; Rezende et al., 2014); inference uses $z = \mu_z$.

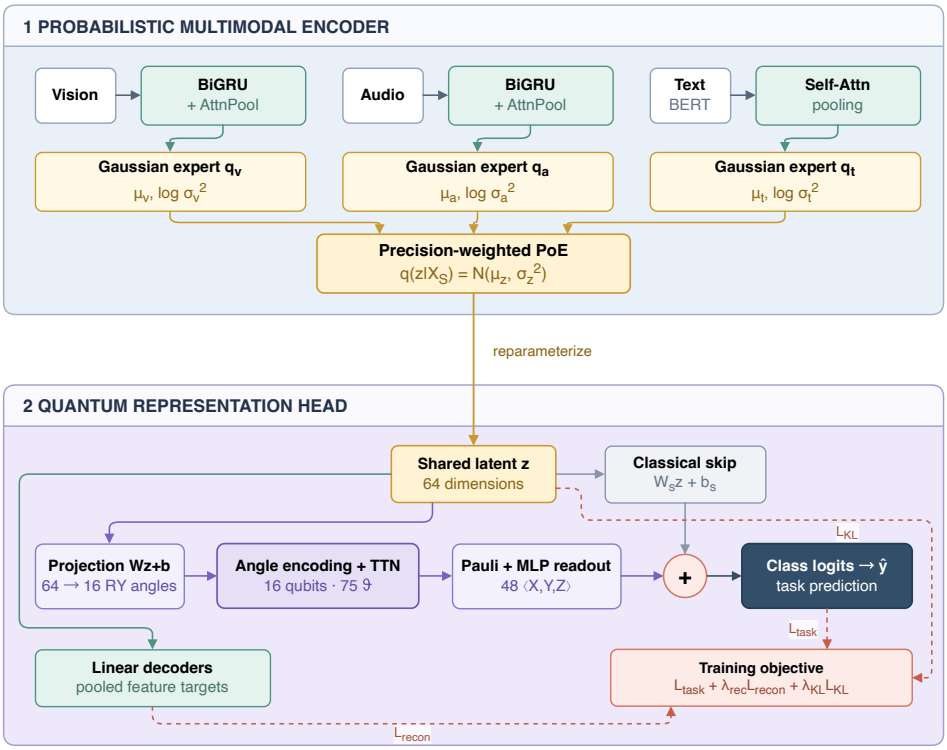


Figure 1: QuTA architecture. Available Gaussian experts form a 64-dimensional PoE posterior. Text uses self-attention pooling over frozen BERT tokens; vision and audio use BiGRUs. A $64 \to 16$ affine map encodes the latent into a 16-qubit TTN; 48 Pauli expectations feed a classical readout and linear skip. Decoders reconstruct pooled modality features.

4.2 Quantum Latent Transformation

A learned affine map produces one encoding angle per qubit, $\tilde{a}(z) = \tilde{W}z + \tilde{b} \in \mathbb{R}^n$ with $n = 16$. The feature map is

$$|\psi(z)\rangle = U_{\text{enc}}(z)|0\rangle^{\otimes n}, \qquad U_{\text{enc}}(z) = \left[\bigotimes_{i=1}^{n} R_Y(\tilde{a}_i(z))\right] H^{\otimes n}. \tag{7}$$

The implementation parameterizes $\tilde{a}$ as four consecutive same-axis R_Y layers, $\tilde{a}_i = \sum_{j=1}^{4} a_{j,i}$ with $a(z) = Wz+b \in \mathbb{R}^{4n}$. Because $R_Y(\alpha)R_Y(\beta) = R_Y(\alpha+\beta)$, this is an overparameterized $64 \to 16$ affine encoding and does not enlarge the noiseless feature map, though the extra parameters can ease training of the encoding.

We then apply a binary tree of two-qubit blocks (Figure 2),

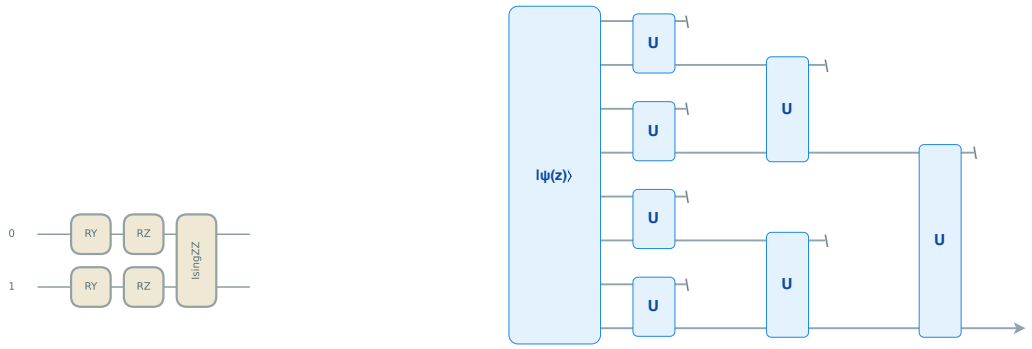
$$U_{\text{TTN}}(\vartheta) \;=\; \prod_{k=1}^{n-1} U_k(\vartheta_k), \tag{8}$$

where each U_k contains two R_Y, two R_Z, and one IsingZZ gate. The $n-1$ blocks therefore use $5(n-1) = 75$ circuit parameters and have $\mathcal{O}(\log n)$ parallel depth.

We measure single-qubit Pauli observables

$$o(z) \;=\; \big(\langle P_i\rangle\big)_{i\in[n],\, P\in\{X,Y,Z\}} \in \mathbb{R}^{3n}, \qquad \langle P_i\rangle \;=\; \langle\psi(z)|\, U_{\text{TTN}}^{\dagger} P_i\, U_{\text{TTN}}\, |\psi(z)\rangle, \tag{9}$$

to obtain $3n = 48$ features. The 75 parameters count only the variational circuit; the projection and readout are included in total model size. The same circuit definition can run on a compatible quantum processor, while exact classical simulation may scale exponentially with n.



(a) Five-parameter block U_k.

(b) Eight-qubit TTN example.

Figure 2: TTN circuit after the PoE latent. The 16-qubit model stacks $n - 1 = 15$ copies of the five-parameter block; panel (b) shows an eight-qubit instance of the same tree.

4.3 Prediction and Reconstruction Heads

The two-layer readout and linear skip produce

$$\begin{aligned} \psi_q(z) &= W_2 \,\mathrm{Dropout}(\phi(W_1 o(z) + b_1)) + b_2, \\ \psi_s(z) &= W_s z + b_s, \qquad f(z) = \psi_q(z) + \psi_s(z), \end{aligned} \tag{10}$$

where ϕ is LeakyReLU. The two branches emit class logits of the same dimension and are combined by addition, not concatenation. $f(z)$ is the discrete task head. Sentiment benchmarks also require a scalar $s(z)$: the softmax expectation of $f(z)$ over class centers and, when MAE/Corr are reported, a linear readout on μ_z. If both scores exist, a mixing weight is chosen on validation and frozen at test. Official accuracy follows the dataset protocol (binning of $s(z)$ or $\arg\max f(z)$); MAE and Corr use $s(z)$. The skip supplies a direct gradient path and lets the quantum branch complement a linear predictor, mitigating—without eliminating—optimization sensitivity in variational circuits (McClean et al., 2018; Cerezo et al., 2021).

Each linear decoder $d_m(z)$ reconstructs the standardized temporal mean $\bar{x}_m \in \mathbb{R}^{d_m}$ of valid frames. We normalize by feature dimension:

$$\mathcal{L}_{\text{recon}} = \frac{1}{|\mathcal{S}|} \sum_{m \in \mathcal{S}} \frac{1}{d_m} \left\| d_m(z) - \bar{x}_m \right\|_2^2 . \tag{11}$$

4.4 Learning Objective and Hybrid Inference

We optimize task loss, reconstruction, and latent regularization jointly:

$$\mathcal{L} \;=\; \mathcal{L}_{\text{task}} + \lambda_{\text{rec}} \mathcal{L}_{\text{recon}} + \lambda_{\text{KL}} \mathcal{L}_{\text{KL}}. \tag{12}$$

For classification, $\mathcal{L}_{\text{task}} = -\alpha_y (1 - p_y)^{\gamma_f} \log p_y$ is class-weighted focal cross-entropy (Lin et al., 2017) on $f(z)$; dual-task sentiment additionally supervises $s(z)$ with Smooth-L1.

For the diagonal posterior, the KL averaged over latent coordinates is

$$\mathcal{L}_{\text{KL}} = \frac{1}{2d} \sum_{j=1}^{d} \left(\mu_{z,j}^2 + \sigma_{z,j}^2 - 1 - \log \sigma_{z,j}^2 \right). \tag{13}$$

Equation 12 is ELBO-motivated but, because its weights are selected independently, is a multi-task objective rather than the ELBO itself. The KL term regularizes toward $p(z) = \mathcal{N}(0, I)$ and provides a rate proxy (Alemi et al., 2017); it does not by itself guarantee posterior calibration.

As Figure 3 shows, training jointly optimizes all classical and TTN parameters with a local simulator in a hybrid VQA loop (Cerezo et al., 2021). At deployment, the classical service computes $z = \mu_z$ and 16 encoding angles; a remote QPU executes the frozen TTN and returns observables for the classical readout and skip.

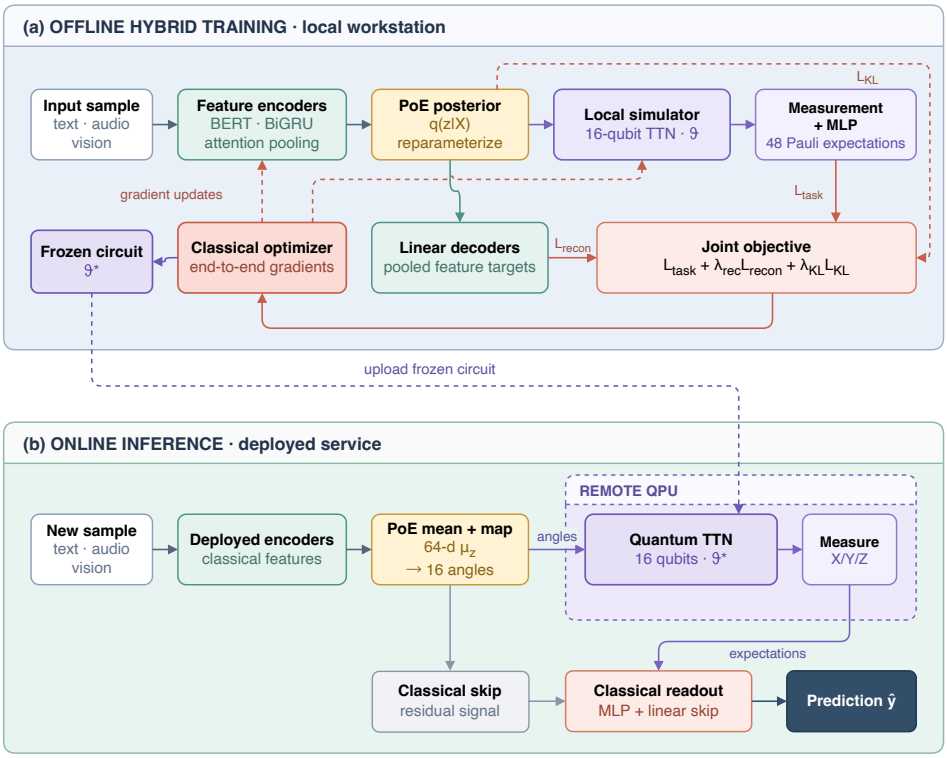


Figure 3: Hybrid workflow. Training uses a local simulator and all three losses; inference sends projected angles to a remote QPU and classifies its returned observables with the frozen classical heads.

5 Experiments

5.1 Datasets and Setup

We evaluate on MOSI (English video sentiment), MELD (dialogue emotion), and CH-SIMS v2 (Chinese video sentiment), using their standard splits, preprocessing, and metrics. All variants use frozen language features, a 64-dimensional PoE latent, and a 16-qubit TTN. Text is pooled with a linear self-attention encoder; vision and audio use BiGRUs, except MELD audio, which uses the same attention encoder as text. MOSI and MELD report seed 42, following their single-run comparisons, while CH-SIMS reports mean±std over three prespecified seeds. MOSI and CH-SIMS report discrete accuracy together with MAE/Corr from the validation-selected sentiment score $s(z)$. Hyperparameters and full parameter-counting rules are in Appendix A.

5.2 Comparative Study and Analysis

We count all trainable weights, including BERT only when it is fine-tuned; offline frozen embeddings are excluded for every method. This distinction is shown explicitly in Tables 1 and 2.

5.2.1 CH-SIMS

Table 1: CH-SIMS v2 under the official protocol. Ours and reproduced MVAE report mean±std over seeds 42, 3407, and 114514; other baselines follow the unified comparison of Mai & Han (2026). FT denotes BERT fine-tuning. Best mean in bold.

Method	BERT	Params	Acc-3	F1	MAE	Corr
No BERT fine-tuning						
AV-MC (Liu et al., 2022)	frozen	0.07M[§]	73.20[¶]	80.70[¶]	0.301[¶]	0.721[¶]
CmIR (Mai & Han, 2026)	frozen	≈0.37M[§]	75.00	**82.00**	0.286	0.742
MVAE (Wu & Goodman, 2018)	frozen	0.72M[§]	$73.95_{\pm 0.87}$	$81.47_{\pm 0.35}$	$0.300_{\pm 0.006}$	$0.740_{\pm 0.004}$
Fine-tuned BERT						
MISA (Hazarika et al., 2020)	FT	104.8M[§]	68.9[¶]	78.3[¶]	0.342[¶]	0.671[¶]
DLF (Wang et al., 2025)	FT	105.2M[§]	70.0[¶]	77.9[¶]	0.346[¶]	0.683[¶]
MMIM (Han et al., 2021)	FT	102.6M[§]	70.4[¶]	77.8[¶]	0.339[¶]	0.641[¶]
MAG-BERT (Rahman et al., 2020)	FT	103.8M[§]	70.6[¶]	77.1[¶]	0.346[¶]	0.641[¶]
Self-MM (Yu et al., 2021)	FT	102.5M[§]	72.7[¶]	78.6[¶]	0.315[¶]	0.691[¶]
KuDA (Feng et al., 2024)	FT	122.5M[§]	74.30[¶]	80.10[¶]	0.289[¶]	0.741[¶]
Ours (QUTA + TTN)	frozen	0.72M[§]	**75.47**$_{\pm 1.06}$	**82.00**$_{\pm 0.88}$	**0.281**$_{\pm 0.005}$	**0.750**$_{\pm 0.003}$

[§]Trainable parameters; frozen embeddings are excluded. [¶]Scores are unified-protocol re-evaluations from CmIR Table 2, while BERT and parameter columns describe each method's native design. QUTA obtains the best Acc-3, MAE, and Corr with 0.72M trainable parameters. MVAE is reproduced with frozen features (718,609 trainable parameters). Compared with MVAE, QUTA improves mean Acc-3 by 1.52 percentage points and F1 by 0.53 points, while reducing MAE from 0.300 to 0.281 and increasing Corr from 0.740 to 0.750.

5.2.2 MOSI

Table 2: MOSI comparison. Acc-2 and F1 exclude neutral samples. FT denotes BERT fine-tuning and Params counts all trainable weights. Best published results are bold; ours and reproduced MVAE report seed-42 tests.

Method	Text	Params	Acc-7	Acc-2	F1	MAE	Corr
No BERT fine-tuning							
TFN[‡] (Zadeh et al., 2017)	frozen	9.50M[§]	34.9	80.8	80.7	0.901	0.698
LMF[‡] (Liu et al., 2018)	frozen	0.51M[§]	33.2	82.5	82.4	0.917	0.695
MulT (Tsai et al., 2019)	GloVe	1.54M[§]	40.0	83.0	82.8	0.871	0.698
MVAE (Wu & Goodman, 2018)	frozen	1.08M[§]	33.53	75.76	75.86	1.005	0.632
Fine-tuned BERT							
MMIM (Han et al., 2021)	FT	109.8M[§]	**46.65**	86.06	85.98	**0.700**	**0.800**
DEVA (Wu et al., 2025)	FT	113.0M[§]	46.32	**86.29**	**86.30**	0.730	0.787
Ours (QUTA + TTN)[†]	frozen	1.08M[§]	41.84	82.77	82.50	0.854	0.716

[†]Single seed-42 test with validation-only model selection. [‡]TFN/LMF use the BERT-feature re-evaluations of Hazarika et al. (2020); Han et al. (2021); MulT uses word-aligned GloVe. [§]Frozen embeddings are excluded from trainable parameters. QUTA improves over the frozen-feature baselines on Acc-7, MAE, and Corr, while the fine-tuned-BERT systems are substantially larger. Against the reproduced MVAE baseline, QUTA improves all five reported metrics, including 8.31 points in Acc-7 and 7.01 points in Acc-2.

5.2.3 MELD

Table 3: MELD 7-class emotion recognition. wF1 is the official metric. MVAE is a reproduced seed-42 baseline; other baseline values follow the original papers.

Method	Params	wF1
text-CNN (Poria et al., 2019)	1.86M[§]	55.02
cMKL (Poria et al., 2019)	n/a (kernel)	55.51
bcLSTM (T+A) (Poria et al., 2019)	5.13M[§]	59.25
DialogueRNN (T+A) (Majumder et al., 2019; Poria et al., 2019)	2.67M[§]	60.25
COSMIC (Ghosal et al., 2020)	9.86M[§]	65.21
MMGCN (Hu et al., 2021)	2.55M[§]	58.65
MVAE (Wu & Goodman, 2018)	—	59.18
Ours (QuTA + TTN)	0.59M[§]	59.30

[§]Trainable task-module parameters; offline embeddings are excluded. With 0.59M parameters, QuTA is comparable to bcLSTM and MMGCN, below DialogueRNN and COSMIC on wF1. MVAE obtains Acc=60.11, wF1=59.18, and macro-F1=40.35 on seed 42; its parameter count is not reported for this run. QuTA and MVAE are close on wF1 (59.30 versus 59.18), with a difference of only 0.12 percentage points in this single-seed comparison.

5.3 Ablation Study

We run three CH-SIMS studies. Module ablations replace the TTN, skip, or PoE one factor at a time (Table 4a); every module row reports mean$\pm$std over seeds 42, 3407, and 114514. Modality ablations retrain each subset with unused experts omitted from the PoE (Table 4b, seed 42). To test the architecture under missing or noisy inputs, we evaluate the frozen full checkpoint while dropping experts or corrupting one or two test modalities (Table 4c, Figure 4; Table 8).

(a) Module ablation.

Variant	Acc-3	Acc-5	MAE	Corr
Full model (TTN + skip + PoE)	**75.47**$_{\pm 1.06}$	**57.35**$_{\pm 0.89}$	**0.281**$_{\pm 0.005}$	0.750$_{\pm 0.003}$
- Quantum TTN $\rightarrow$ MLP	74.76$_{\pm 0.42}$	56.54$_{\pm 1.41}$	0.284$_{\pm 0.010}$	0.751$_{\pm 0.009}$
- Remove skip	75.47$_{\pm 1.26}$	55.77$_{\pm 0.54}$	0.286$_{\pm 0.003}$	**0.751**$_{\pm 0.006}$
- PoE $\rightarrow$ concat	74.05$_{\pm 0.30}$	50.84$_{\pm 1.57}$	0.309$_{\pm 0.009}$	0.730$_{\pm 0.013}$

(b) Retrained modality subsets.

Modalities	Acc-3	Acc-5	MAE	Corr
T	70.02	47.58	0.3434	0.6576
V	61.90	40.72	0.3951	0.5344
A	46.32	33.66	0.4820	0.2355
T+V	74.56	55.32	0.2936	0.7319
T+A	69.44	48.65	0.3351	0.6585
T+V+A	**76.69**	**58.12**	**0.2788**	**0.7509**

(c) Frozen full-model inference with missing test modalities.

Modalities	Acc-3	Acc-5	MAE	Corr
T	68.76	42.17	0.3838	0.5985
V	61.80	36.65	0.4563	0.4787
A	42.36	31.24	0.5285	0.1001
T+V	76.69	50.77	0.3079	0.7369
T+A	66.73	50.29	0.3589	0.5993
V+A	62.67	43.81	0.3788	0.5511
T+V+A	**76.69**	**58.12**	**0.2788**	**0.7509**

Table 4: CH-SIMS ablations under the official configuration. Module rows with $\pm$ are mean$\pm$std over seeds 42, 3407, and 114514. Table 4b uses seed 42. The MLP substitute uses two linear layers with widths 64-1-3 (71 trainable parameters in the head, versus 75 TTN parameters). Table 4b retrains each subset; Table 4c uses the frozen full checkpoint.

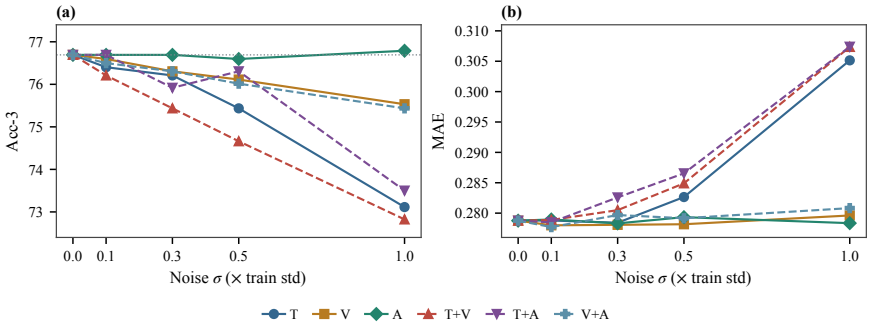


Figure 4: Frozen CH-SIMS inference under feature noise: (a) Acc-3 and (b) MAE. Solid lines corrupt one modality; dashed lines corrupt two. Full metrics appear in Appendix Table 8.

5.4 Noise Robustness under Density-Matrix Simulation

We assess the frozen seed-42 model on a stratified 200-example CH-SIMS subset with ideal Lightning and the 16-qubit Amazon Braket DM1 density-matrix simulator (Preskill, 2018; Amazon Web Services, 2026). All parameters and the validation-selected sentiment score remain fixed. DM1 applies single-qubit depolarizing noise after each H, R_Y, and R_Z gate and two-qubit noise after each IsingZZ gate, using the uncompiled four-layer angle encoding. We test $p = 0.01$ on all 200 examples and sweep $p \in \{0, 0.005, 0.01, 0.02\}$ on a fixed 64-example subset. Comparisons pair identical examples; $p = 0$ uses ideal Lightning.

At $p = 0.01$, only two of 200 Acc-3 decisions change; Acc-3 moves from 77.00 to 77.50, while Acc-2 and weighted binary F1 are unchanged (Appendix Table 7). Scores retain correlation 0.9999 with mean absolute displacement 0.0030. On the 64-example sweep, classification metrics remain unchanged through $p = 0.02$, while mean score displacement rises to 0.0081 (Figure 5). This supports stability under the tested noise settings, not a benefit from noise.

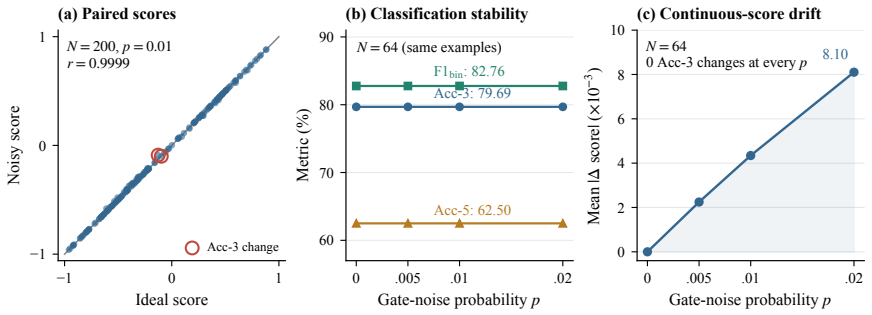


Figure 5: Gate-noise robustness: (a) paired scores on 200 examples at $p = 0.01$, with two Acc-3 changes circled; (b) classification stability and (c) score drift on the fixed 64-example sweep.

6 Conclusion

We presented QuTA, which combines precision-weighted probabilistic fusion, feature reconstruction, and a 75-parameter quantum TTN head. It leads the matched-protocol CH-SIMS comparison and remains competitive on MOSI and MELD with roughly 1M trainable parameters. Its hybrid workflow trains on a local simulator and can execute the frozen TTN on a remote QPU at inference. Paired density-matrix simulation on CH-SIMS shows stable predicted scores: only two of 200 Acc-3 decisions change at $p = 0.01$, and a fixed 64-example sweep retains its classification metrics through $p = 0.02$.

AI USE STATEMENT

We used Claude (Anthropic) and Codex (OpenAI) for four categories of assistance: (1) writing and polishing, including language editing, terminology consistency, and LaTeX formatting; (2) retrieval and discovery, including finding related work and relevant documentation; (3) research ideation and execution, including discussing experimental designs, implementing and modifying experimental code, and interpreting results; and (4) drafting manuscript sections, including the abstract, introduction, background, related work, and conclusion. AI assistance informed the research workflow; reported empirical measurements were obtained from experimental runs. The authors reviewed and edited AI-assisted material and take responsibility for the final manuscript, code, citations, and claims. No AI tool is listed as an author.

REPRODUCIBILITY STATEMENT

The full training and evaluation code will be released with the camera-ready version. The implementation is built on PyTorch 2 with PennyLane 0.45 (software-only, device-agnostic); the variational quantum circuit is evaluated primarily with a classical state-vector simulator, with the frozen CH-SIMS circuit additionally evaluated under gate noise using density-matrix simulation. Vision and audio are encoded by a bidirectional GRU with attention pooling; text uses a linear projection and self-attention pooling over offline frozen BERT token features that are never updated. The fusion head is a 16-qubit tree tensor network (75 trainable parameters) with a linear skip. The latent dimension is 64. We optimize with AdamW and a focal classification loss plus the reconstruction and KL terms in equation 12; hyperparameters, data preprocessing, and the dataset splits follow the compared methods.

MOSI and MELD report a single seed (42), consistent with the compared baselines. For CH-SIMS we report mean $\pm$ std over three prespecified seeds (42, 3407, 114514), with model selection performed on validation only. Full configuration files, per-seed logs, and prediction hashes for every reported table will be included in the release, enabling the results in Sec. 5.2 to be reproduced end-to-end.

REFERENCES

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*, 2017.

Amazon Web Services. Amazon braket density matrix simulator (DM1). `https://docs.aws.amazon.com/braket/latest/developerguide/braket-submit-tasks-simulators.html`, 2026. Accessed 2026-09-06.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019. doi: 10.1109/tpami.2018.2798607.

Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, 4(4):043001, 2019. doi: 10.1088/2058-9565/ab4eb5.

Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. Variational quantum algorithms. *Nature Reviews Physics*, 3:625–644, 2021.

Yiwei Chen, Kehuan Yan, Yu Pan, and Daoyi Dong. Qinn-qj: A quantum-inspired neural network with quantum jump for multimodal sentiment analysis. arXiv preprint arXiv:2510.27091, 2025.

Xinyu Feng, Yuming Lin, Lihua He, You Li, Liang Chang, and Ya Zhou. Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14755–14766, 2024.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. Cosmic: Commonsense knowledge for emotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2470–2481, 2020.

Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9186–9206, 2021. doi: 10.18653/v1/2021.emnlp-main.723.

Vojtěch Havlíček, Antonio D. Córcoles, Kristan Temme, Aram W. Harrow, Abhinav Kandala, Jerry M. Chow, and Jay M. Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747):209–212, 2019.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1122–1131, 2020.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5666–5675, 2021.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Qiuchi Li, Dimitris Gkoumas, Christina Lioma, and Massimo Melucci. Quantum-inspired multimodal fusion for video sentiment analysis. *Information Fusion*, 2021.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, and Xiaoteng Li. Make acoustic and visual cues matter: CH-SIMS v2.0 dataset and AV-Mixup consistent module. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, 2022.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2247–2256, 2018.

James Lucas, George Tucker, Roger B. Grosse, and Mohammad Norouzi. Don't blame the ELBO! a linear VAE perspective on posterior collapse. *arXiv preprint arXiv:1911.02469*, 2019.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18177–18186, 2022. doi: 10.1109/cvpr52688.2022.01764.

Sijie Mai and Shiqin Han. Learning invariant modality representation for robust multimodal learning from a causal inference perspective. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 45698–45721, 2026.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6818–6825, 2019.

Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9:4812, 2018.

Kosuke Mitarai, Makoto Negoro, Masahiro Kitagawa, and Keisuke Fujii. Quantum circuit learning. *Physical Review A*, 98(3):032309, 2018. doi: 10.1103/physreva.98.032309.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 689–696, 2011.

Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J. Love, Alán Aspuru-Guzik, and Jeremy L. O'Brien. A variational eigenvalue solver on a photonic quantum processor. *Nature Communications*, 5:4213, 2014.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, 2019. doi: 10.18653/v1/P19-1050.

John Preskill. Quantum computing in the NISQ era and beyond. *Quantum*, 2:79, 2018. doi: 10.22331/q-2018-08-06-79.

Jun Qi, Chao-Han Huck Yang, and Pin-Yu Chen. Qtn-vqc: An end-to-end learning framework for quantum neural networks. arXiv preprint arXiv:2110.03861, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Wasifur Rahman, Md. Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2359–2369, 2020.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning (ICML)*, 2014.

Saurav Sahay, Shachi H. Kumar, Rui Xia, Jonathan Huang, and Lama Nachman. Multimodal relational tensor network for sentiment and emotion classification. arXiv preprint arXiv:1806.02923, 2018.

Maria Schuld and Nathan Killoran. Quantum machine learning in feature hilbert spaces. *Physical Review Letters*, 122(4):040504, 2019.

Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. Evaluating analytic gradients on quantum hardware. *Physical Review A*, 99(3):032331, 2019. doi: 10.1103/physreva.99.032331.

Yuge Shi, Siddharth Narayanaswamy, Brooks Paige, and Philip H. S. Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Thomas M. Sutter, Imant Daunhawer, and Julia E. Vogt. Multimodal generative learning utilizing Jensen-Shannon divergence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6558–6569, 2019.

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. Dlf: Disentangled-language-focused multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21180–21188, 2025.

Samson Wang, Enrico Fontana, M. Cerezo, Kunal Sharma, Andrew Sornborger, Lukasz Cincio, and Patrick J. Coles. Noise-induced barren plateaus in variational quantum algorithms. *Nature Communications*, 12:6961, 2021.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Sheng Wu, Xiaobao Wang, Longbiao Wang, Dongxiao He, and Jianwu Dang. Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6575–6583, 2021.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1103–1114, 2017.

A Implementation Details

A.1 Shared Model Configuration

All models share the same architecture: per-modality encoders map each modality to a Gaussian factor, a precision-weighted Product-of-Experts layer fuses them into a 64-dimensional latent, and a 16-qubit TTN (75 trainable circuit parameters) followed by a classical MLP readout and a linear skip produces the task output (Sec. 4). Angle encoding uses four consecutive R_Y layers that compose to the single-layer map in Eq. 7. Encoders are ModalityEncoderV2 with dropout 0.3. Text is an attention-only pool: a linear map into width $2h$ followed by a learned self-attention over frozen BERT tokens (Figure 1). Vision and audio use a one-layer bidirectional GRU of per-direction size h, except on MELD, where audio is also attention-only. Here $h = 128$ on CH-SIMS and $h = 192$ on MOSI and MELD. The quantum branch is evaluated exactly with a state-vector simulator (`lightning.qubit`, shots disabled); the noise-robustness evaluation in Sec. 5.4 instead uses Amazon Braket DM1 with the frozen model. Optimization uses AdamW with learning rate 10^{-3}, weight decay 10^{-2}, a linear warmup followed by cosine decay, and a focal classification loss with $\gamma_f = 2$ combined with the reconstruction and KL terms of Sec. 4.4. Latent dimension is 64 throughout; each per-modality linear decoder reconstructs the standardized temporal mean of the corresponding input features.

A.2 Parameter Accounting

We count every trainable weight from each method's official architecture and exclude offline frozen language features (BERT or GloVe), which are preprocessors never updated. On MOSI, TFN and LMF are counted from the official fusion architecture with MMSA BERT-as-features input sizes $(768, 5, 20)$: TFN uses hidden dimensions $[128, 32, 128]$, text-out 32 and post-fusion width 64; LMF uses $[128, 16, 128]$ and rank 3. MulT is counted from Table 5 of Tsai et al. (2019) ($d = 40$, $D = 4$, 10 heads) and the official encoder, with GloVe / COVAREP / Facet inputs $(300, 74, 35)$. MMIM and DEVA fine-tune BERT-base ($\approx$109.5M); their task heads alone are 0.28M / 3.48M. On CH-SIMS, CmIR's modules are recomputed as $\approx$0.37M. Table 5 lists our trainable counts from `numel` on the reported models; frozen BERT is excluded and the quantum TTN contributes 75 parameters.

A.3 Per-Dataset Configuration

Table 6 summarizes the dataset-specific settings. For MOSI, the reported test is a single seed-42 run with validation-only model selection. MELD uses text and audio only (the official protocol's visual features are not used), with balanced sampling. CH-SIMS uses the official v2 protocol with three prespecified seeds; mixing weights and checkpoints are selected on validation only.

For CH-SIMS, the reconstructed modalities are text, vision and audio; for MELD only text and audio (the vision encoder is omitted). Dataset-specific loss weights are fixed before test evaluation and listed in the released configuration files.

Table 5: Trainable parameters of the reported QuTA models ($\times 10^3$). Text is an attention encoder; vision/audio are BiGRUs except MELD audio (attention). Head includes the $64 \to 64$ angle projection, TTN, readout, skip, decoders, and the scalar sentiment readout when the benchmark reports MAE/Corr.

Module	CH-SIMS	MOSI	MELD
Text encoder (attn)	230	345	345
Vision encoder (BiGRU)	269	313	–
Audio encoder	152	358	165
Head (TTN + readout + skip + decoders)	71	65	78
Total	722	1,082	588

Table 6: Per-dataset configuration. T, V, A are the text, vision and audio feature dimensions after the frozen-feature preprocessing of Sec. 5.1; BERT features are frozen in all cases. Hidden dim h is the per-direction GRU size; attention encoders project to width $2h$.

	CH-SIMS	MOSI	MELD
Text dim T	768	768	768
Vision dim V	177	35	–
Audio dim A	25	74	300
Text encoder	attn	attn	attn
Vision encoder	BiGRU	BiGRU	–
Audio encoder	BiGRU	BiGRU	attn
Batch size	32	32	64
Epochs	25	8	20
Hidden dim	128	192	192
Warmup epochs	5	5	5
Focal γ_f	2	2	2
λ_{rec}	0.03	0.03	0.1
λ_{KL}	0.005	0.005	0.005
Seeds	42,3407,114514	42	42

A.4 Gate-noise Evaluation Details

Table 7 gives the complete paired metrics for the gate-noise experiments summarized in Sec. 5.4.

Table 7: Paired CH-SIMS noise evaluation with the frozen seed-42 model. All rows use the same reported sentiment score as Table 1. The 200-example evaluation and 64-example sweep are separate paired comparisons. $p = 0$ uses ideal Lightning; $p > 0$ uses DM1. F1_{bin} is weighted binary F1.

N	p	Acc-5	Acc-3	Acc-2	F1_{bin}	MAE	Corr
200	0	58.50	77.00	81.50	81.42	0.2675	0.7598
200	0.01	59.50	77.50	81.50	81.42	0.2658	0.7622
64	0	62.50	79.69	82.81	82.76	0.2366	0.8131
64	0.005	62.50	79.69	82.81	82.76	0.2360	0.8140
64	0.01	62.50	79.69	82.81	82.76	0.2354	0.8148
64	0.02	62.50	79.69	82.81	82.76	0.2344	0.8160

A.5 TEST-TIME FEATURE NOISE

Table 8 reports the full official-test metrics for the frozen seed-42 model under additive Gaussian corruption of one or two modalities, corresponding to Figure 4. Missing-expert inference is Table 4c; here every expert remains in the PoE.

Table 8: Frozen full-model CH-SIMS inference with additive Gaussian noise on one or two test modalities. σ multiplies the train-set per-dimension timestep standard deviation. All experts remain in the PoE. Official test split, seed-42 fused output.

Noisy	σ	Acc-5	Acc-3	MAE	Corr
–	0	58.12	76.69	0.2788	0.7509
T	0.1	58.51	76.40	0.2788	0.7490
T	0.3	58.61	76.21	0.2784	0.7484
T	0.5	57.54	75.44	0.2826	0.7395
T	1	55.32	73.11	0.3051	0.7071
V	0.1	58.12	76.60	0.2780	0.7523
V	0.3	58.03	76.31	0.2781	0.7529
V	0.5	58.12	76.11	0.2782	0.7527
V	1	57.45	75.53	0.2796	0.7522
A	0.1	58.03	76.69	0.2790	0.7506
A	0.3	58.61	76.69	0.2783	0.7527
A	0.5	58.70	76.60	0.2793	0.7519
A	1	58.41	76.79	0.2783	0.7558
T+V	0.1	57.83	76.21	0.2789	0.7518
T+V	0.3	56.67	75.44	0.2805	0.7515
T+V	0.5	56.19	74.66	0.2849	0.7458
T+V	1	53.19	72.82	0.3074	0.7040
T+A	0.1	58.61	76.69	0.2785	0.7515
T+A	0.3	58.51	75.92	0.2826	0.7473
T+A	0.5	58.51	76.31	0.2866	0.7403
T+A	1	55.51	73.50	0.3074	0.7026
V+A	0.1	58.03	76.50	0.2777	0.7535
V+A	0.3	57.54	76.31	0.2797	0.7510
V+A	0.5	57.25	76.02	0.2792	0.7518
V+A	1	56.87	75.44	0.2808	0.7508