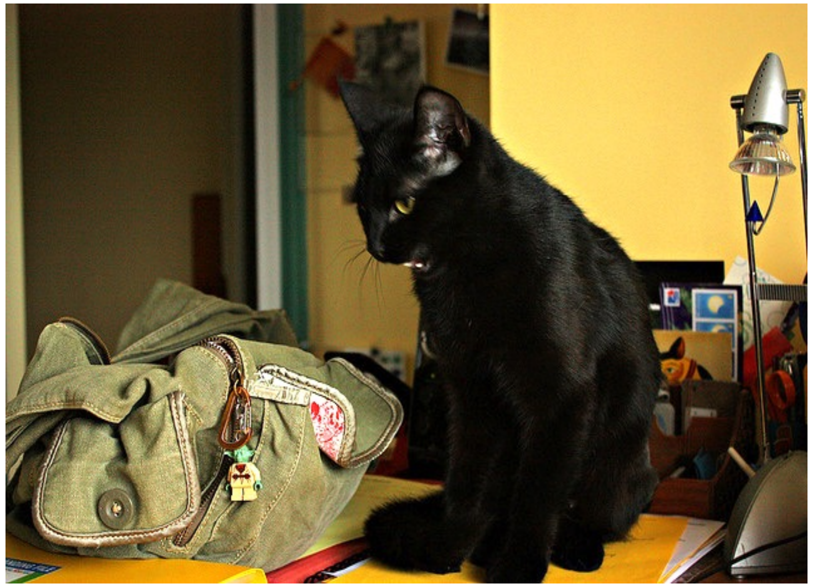
Bidirectional Contrastive Split Learning for Visual Question Answering

Yuwei Sun and Hideya Ochiai

東京大学
THE UNIVERSITY OF TOKYO

Center for
Advanced Intelligence Project

Multi-modal machine learning



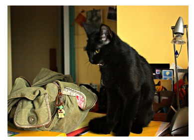
Q: Where is the bag?

A: On the table

- **V**isual **Q**uestion **A**nswering: Answering natural language questions based on the contents of a presented image

Multi-modal machine learning

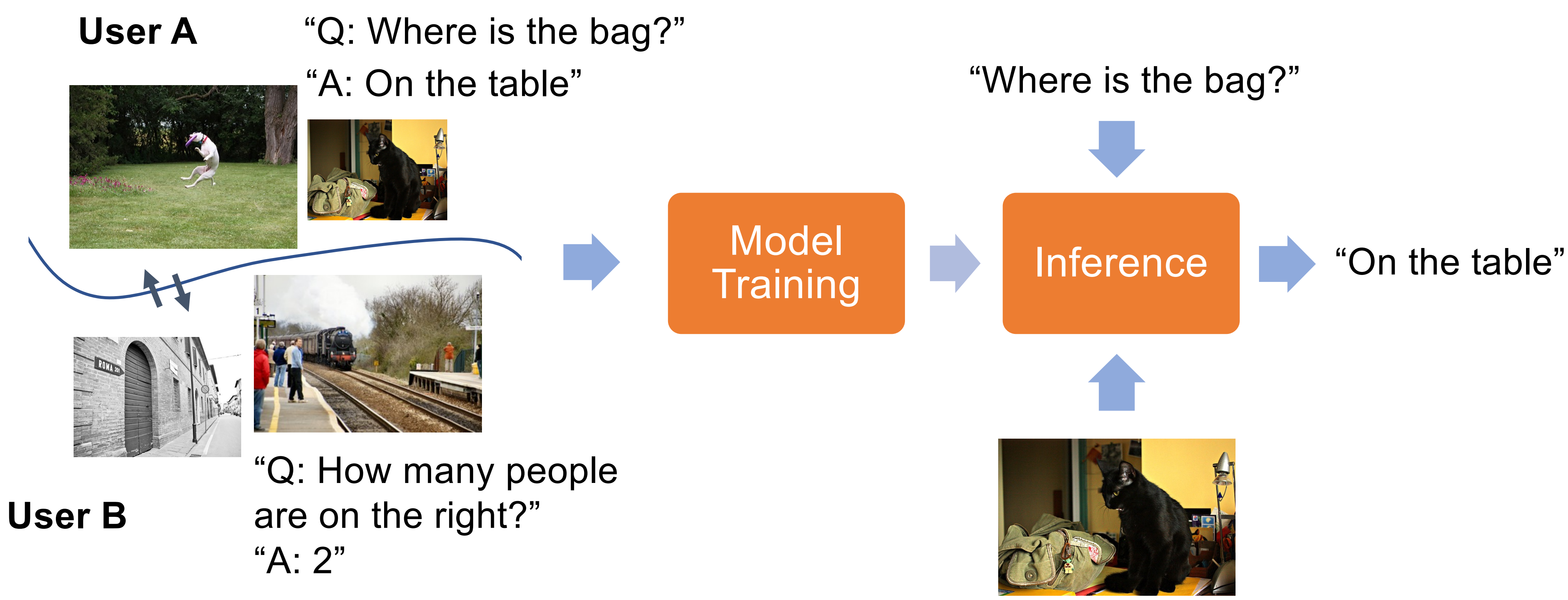
User A “Q: Where is the bag?”
“A: On the table”



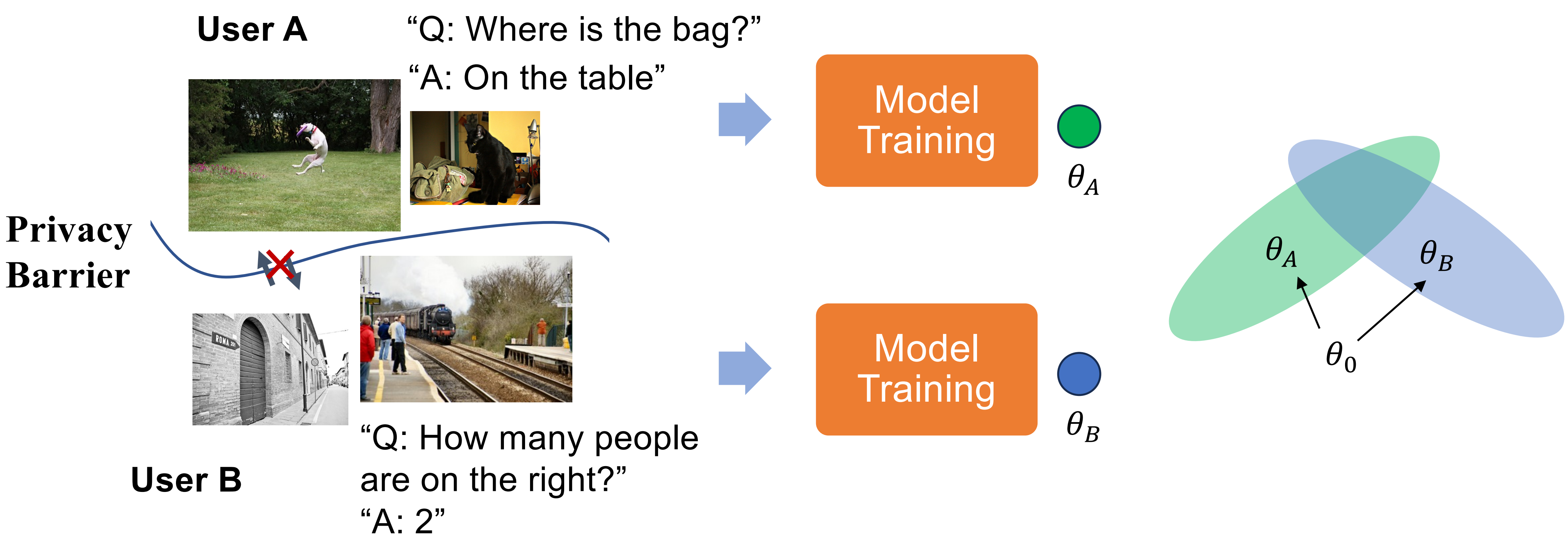
Model Training → Inference

User B “Q: How many people are on the right?”
“A: 2”

Multi-modal machine learning



More robust decentralized multi-modal learning

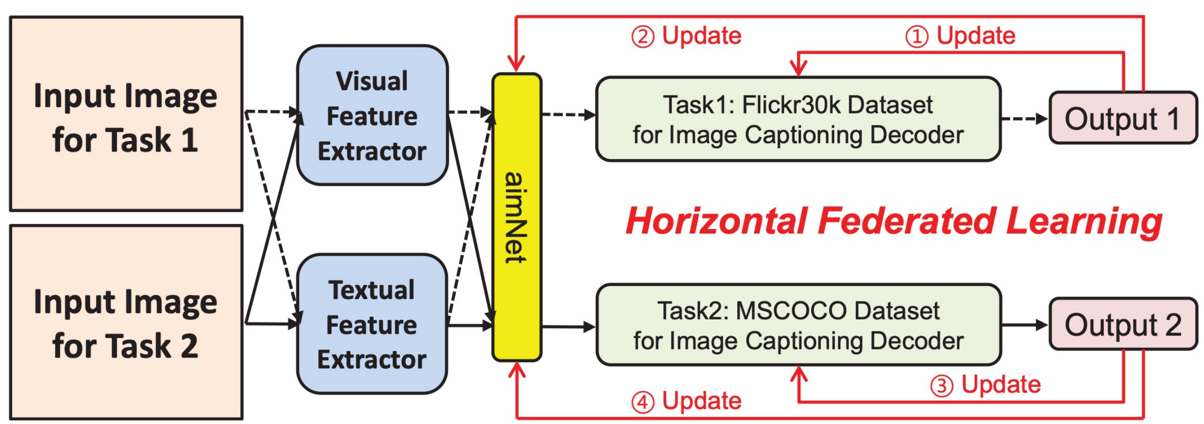


- ➢ The collected vast amount of user data for training raises critical privacy concerns.
- ➢ Transferring and aggregating the knowledge from these individually learned models is crucial for achieving the training goal across the entire data distribution.

Decentralized VQA

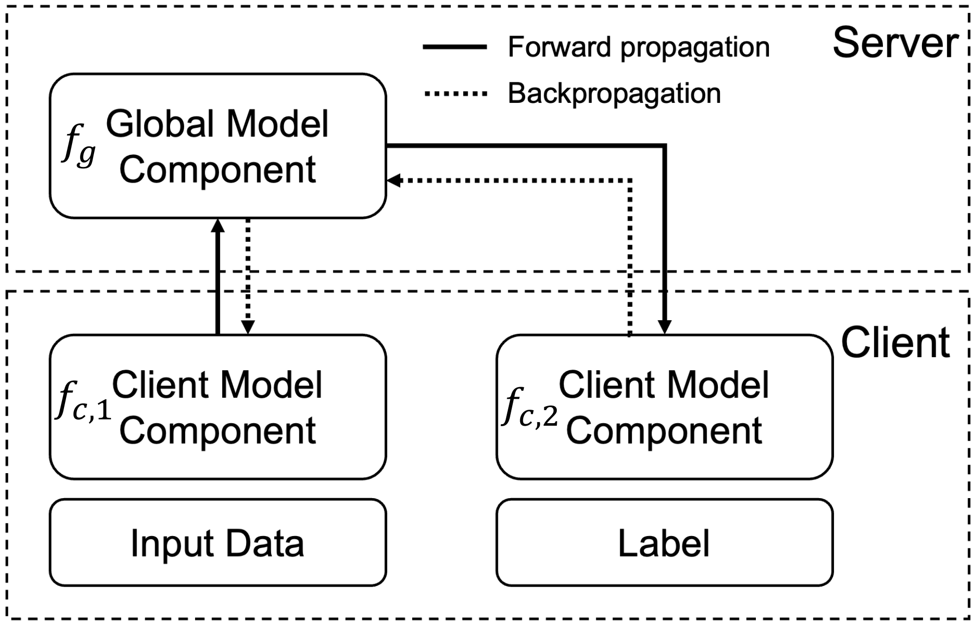
Methods	Shared Data	Shared Model	Learning Framework	Loss Function
MMNas	✓	✓	Single fusion	Cross entropy
QICE	✓	✓	Single fusion	Contrastive loss
aimNet	×	✓	Federated Learning	Cross entropy
BiCSL (Ours)	×	×	Split Leaning	Contrastive loss

- Existing decentralized methods depend on learned model weight sharing.
- However, sharing a complete model results in **adversarial attacks** and **inefficient training** due to constrained client resources.



aimNet [Liu et al. AAAI 2020]

Split Learning for model privacy



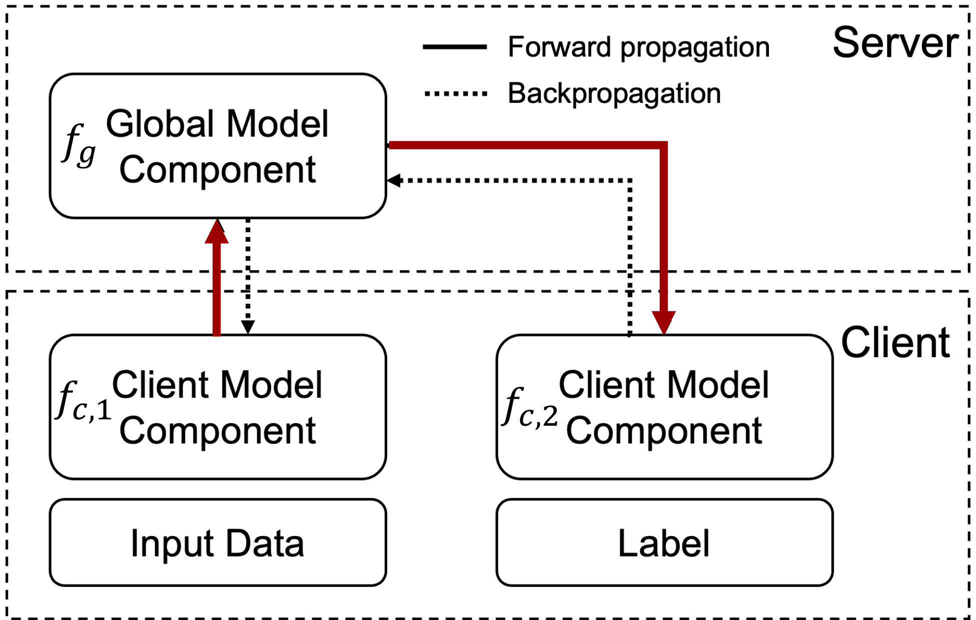
(a) Split Learning

Unidirectional, sequential

Activations: $f_{c,1} \rightarrow f_g \rightarrow f_{c,2}$

Gradients: $f_{c,1} \leftarrow f_g \leftarrow f_{c,2}$

Split Learning for model privacy



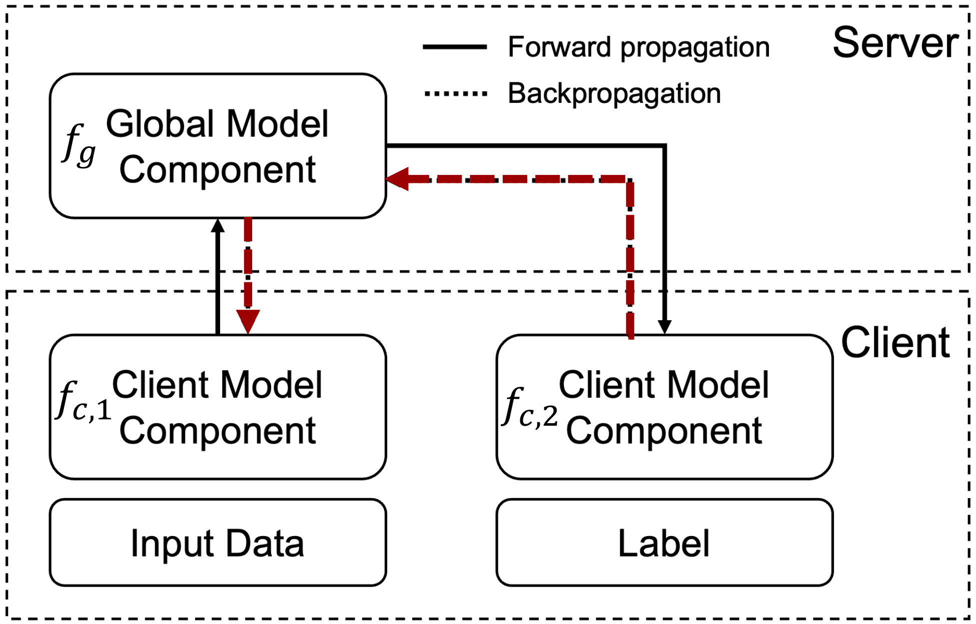
(a) Split Learning

Unidirectional, sequential

Activations: $f_{c,1} \rightarrow f_g \rightarrow f_{c,2}$

Gradients: $f_{c,1} \leftarrow f_g \leftarrow f_{c,2}$

Split Learning for model privacy



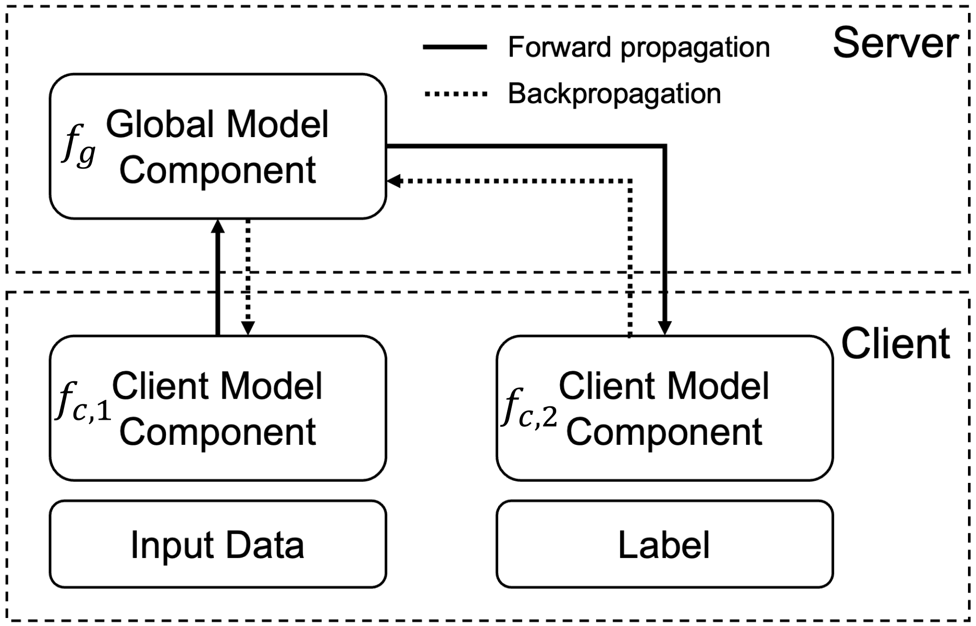
(a) Split Learning

Unidirectional, sequential

Activations: $f_{c,1} \rightarrow f_g \rightarrow f_{c,2}$

Gradients: $f_{c,1} \leftarrow f_g \leftarrow f_{c,2}$

BiCSL vs. Split Learning

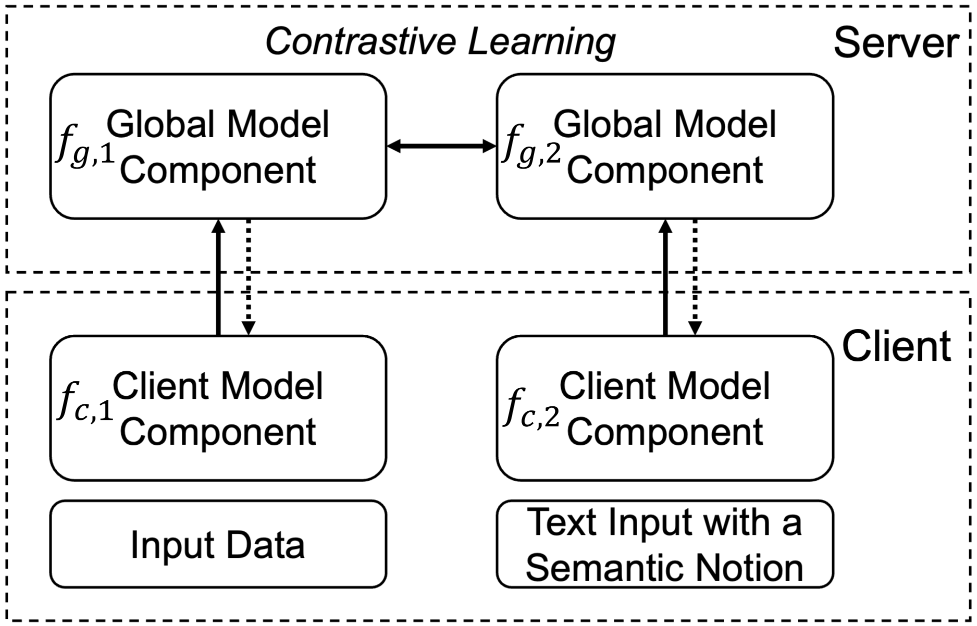


(a) Split Learning

Unidirectional, sequential

Activations: $f_{c,1} \rightarrow f_g \rightarrow f_{c,2}$

Gradients: $f_{c,1} \leftarrow f_g \leftarrow f_{c,2}$



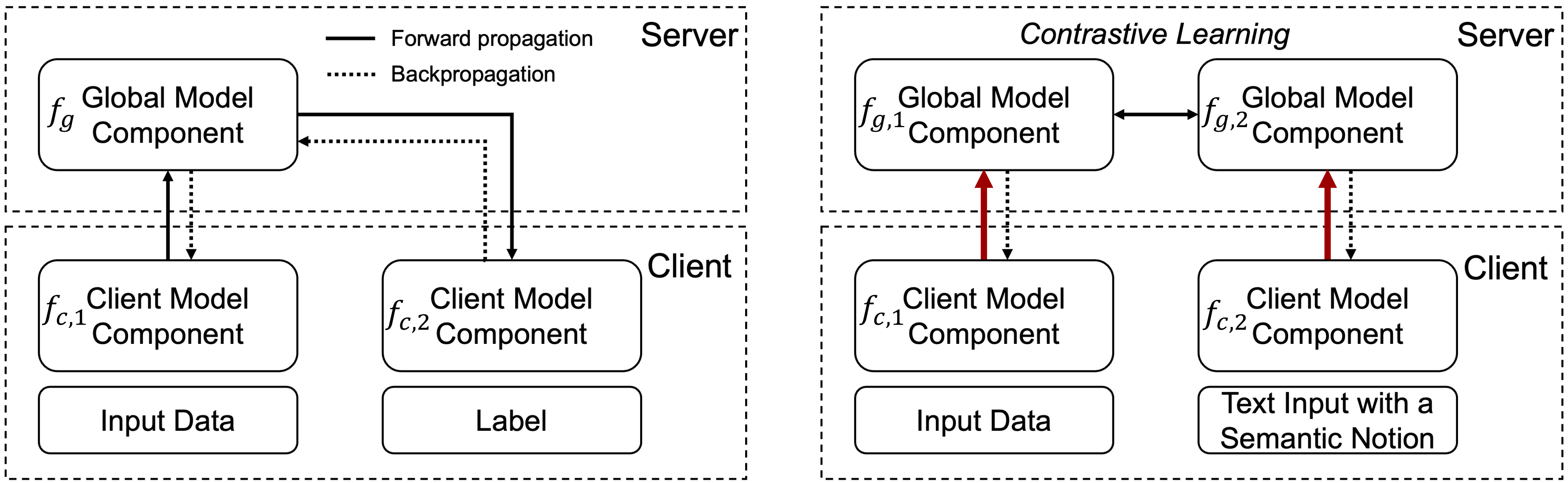
(b) Bidirectional Contrastive Split Learning (BiCSL, ours)

Bidirectional, concurrent

Activations: $f_{c,1} \rightarrow f_{g,1}$ $f_{c,2} \rightarrow f_{g,2}$

Gradients: $f_{c,1} \leftarrow f_{g,1}$ $f_{c,2} \leftarrow f_{g,2}$

BiCSL vs. Split Learning



(a) Split Learning

Unidirectional, sequential

Activations: $f_{c,1} \rightarrow f_g \rightarrow f_{c,2}$

Gradients: $f_{c,1} \leftarrow f_g \leftarrow f_{c,2}$

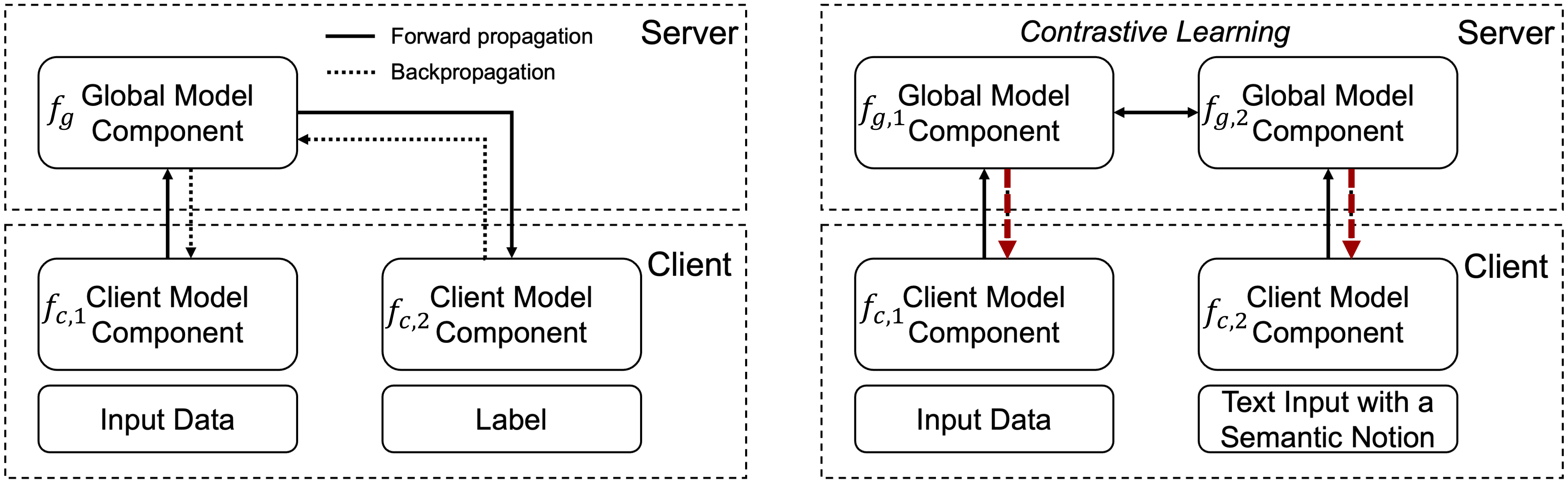
(b) Bidirectional Contrastive Split Learning (BiCSL, ours)

Bidirectional, concurrent

Activations: $f_{c,1} \rightarrow f_{g,1} \quad f_{c,2} \rightarrow f_{g,2}$

Gradients: $f_{c,1} \leftarrow f_{g,1} \quad f_{c,2} \leftarrow f_{g,2}$

BiCSL vs. Split Learning



(a) Split Learning

Unidirectional, sequential

Activations: $f_{c,1} \rightarrow f_g \rightarrow f_{c,2}$

Gradients: $f_{c,1} \leftarrow f_g \leftarrow f_{c,2}$

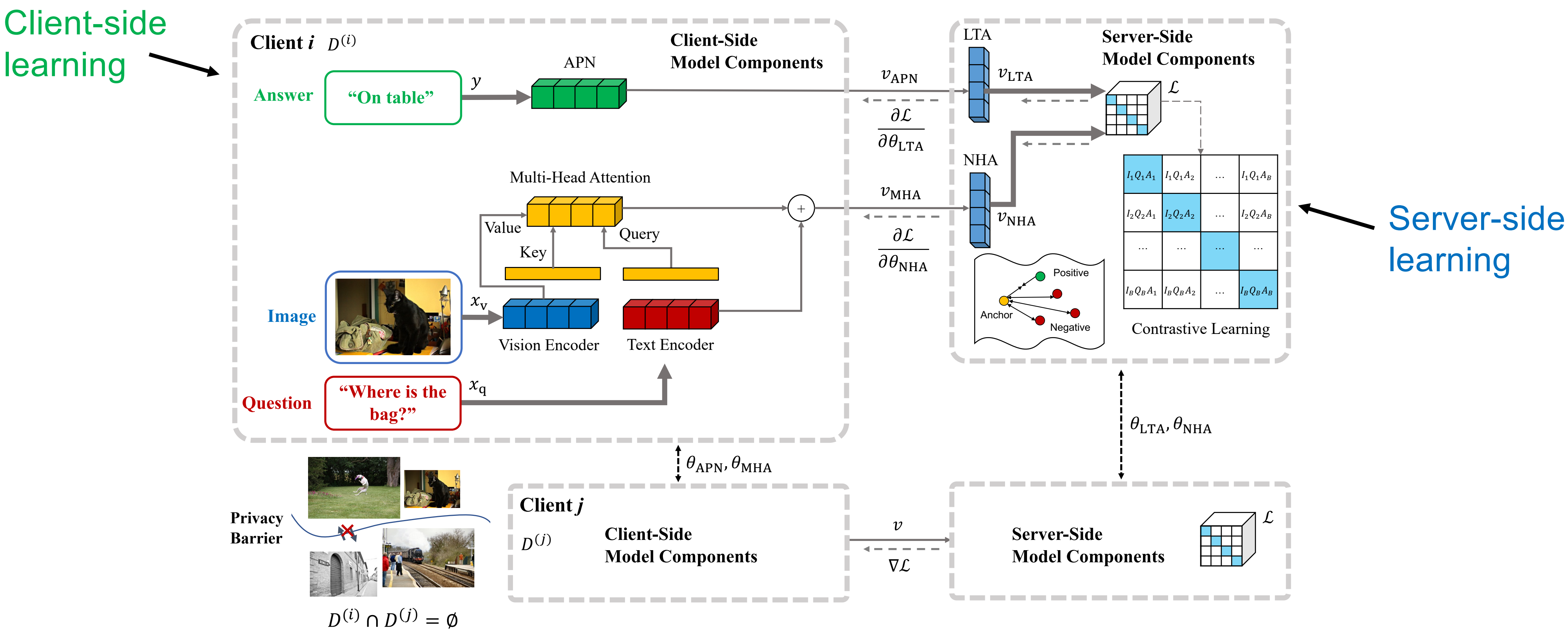
(b) Bidirectional Contrastive Split Learning (BiCSL, ours)

Bidirectional, concurrent

Activations: $f_{c,1} \rightarrow f_{g,1}$ $f_{c,2} \rightarrow f_{g,2}$

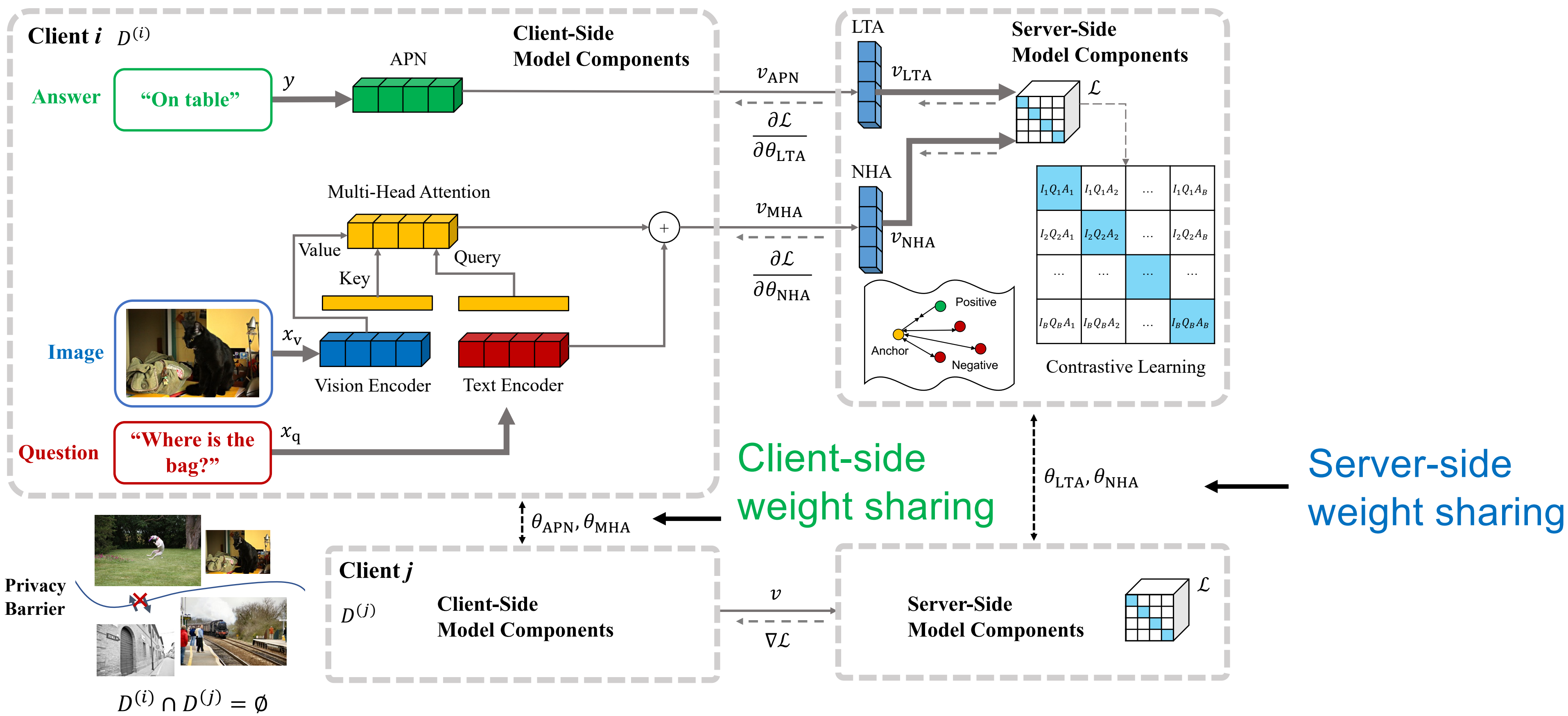
Gradients: $f_{c,1} \leftarrow f_{g,1}$ $f_{c,2} \leftarrow f_{g,2}$

Bidirectional Contrastive Split Learning



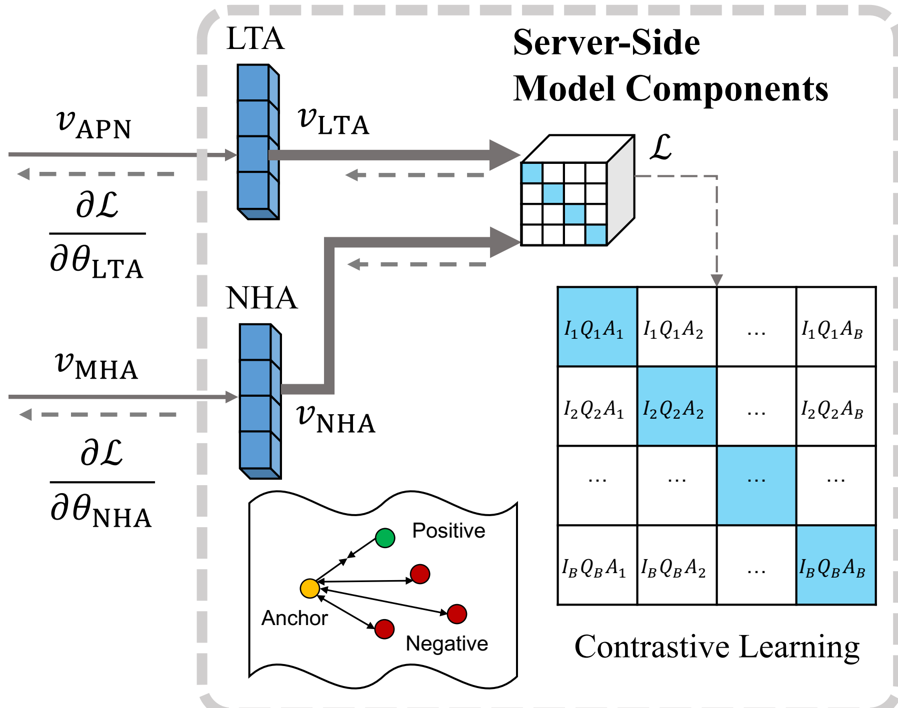
- A multi-modal model is decoupled into **representation modules** and a **contrastive module** for inter-module gradients and inter-client weight sharing.

Bidirectional Contrastive Split Learning

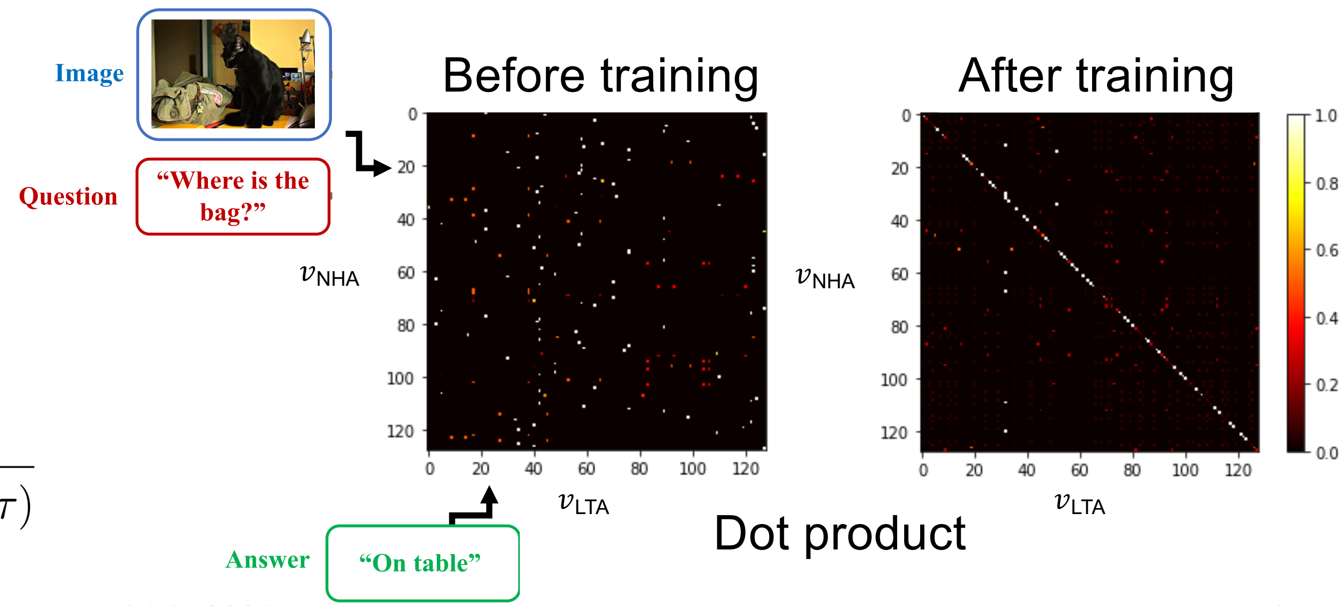


- A multi-modal model is decoupled into **representation modules** and a **contrastive module** for inter-module gradients and inter-client weight sharing.

Cross-modal contrastive learning



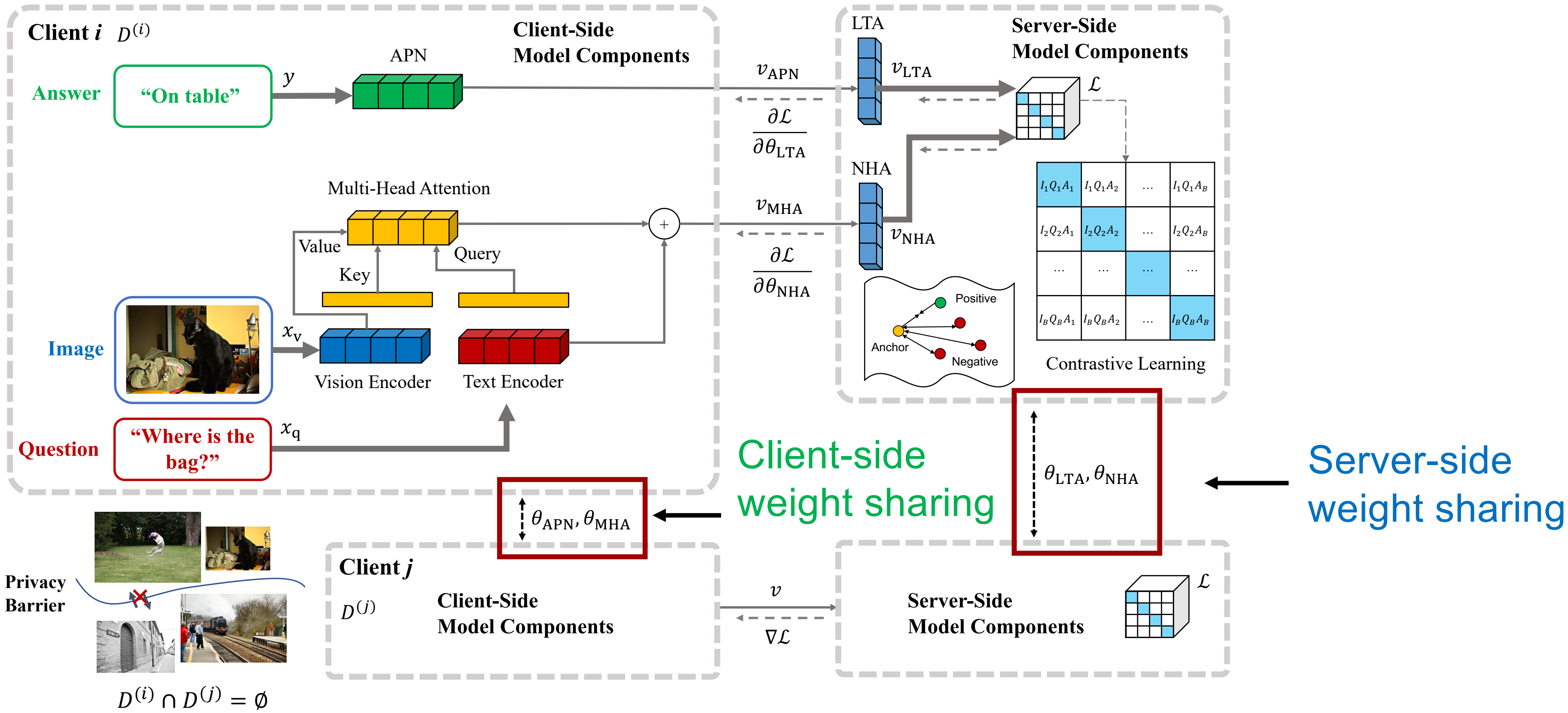
- Contrastive learning disentangles similar and dissimilar pairs of data points within a batch *B*:
  - Given $v_{\text{NHA},i}$
  - $\{v_{\text{LTA},j} \mid j = i\}$ as the positive pair
  - $\{v_{\text{LTA},j} \mid j \neq i\}_{j=1}^{B}$ as the negative pairs



➢ Contrastive loss [Radford, 2021]

$$\mathcal{L} = -\sum_{i=1}^{B} \log \frac{\exp(v_{\text{NHA},i} \cdot v_{\text{LTA},i} / \tau)}{\sum_{j=1}^{B} \mathbb{1}_{[j \neq i]} \exp(v_{\text{NHA},i} \cdot v_{\text{LTA},j} / \tau)}$$

Weight sharing for module update aggregation



- At every epoch, aggregate updates to enhance the global model's performance.
- $\delta\theta_t = \frac{1}{K}\sum_{k\in\{1,2,\ldots,K\}}(\theta_{t+1}^{(k)} - \theta_t^{(k)})$, for a model component from $\{\theta_{APN}, \theta_{MHA}, \theta_{NHA}, \theta_{LTA}\}$.

Evaluation

Centralized

VQA-v2 [Agrawal, 2017]

- Training: 83k images, 444k questions
- Validation: 41k images, 214k questions

Transferred knowledge

Two non-overlapping subsets

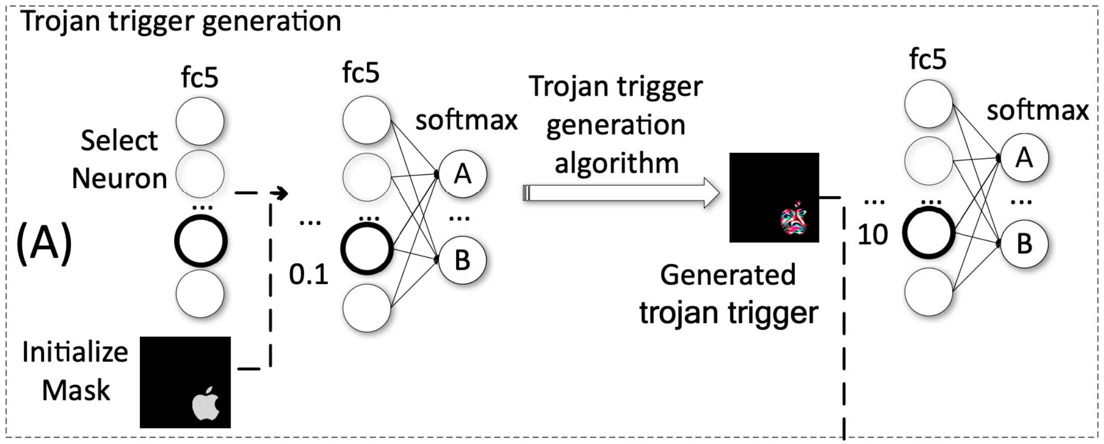
VQA Models	Contrastive learning (%)			
	Overall	Yes/No	Number	Other
BAN	36.23 ± 0.53	66.90 ± 0.71	12.71 ± 0.32	19.11 ± 0.47
BUTD	45.08 ± 0.64	75.82 ± 0.82	29.27 ± 0.53	25.86 ± 0.41
MFB	46.98 ± 0.58	73.95 ± 0.77	32.81 ± 0.49	30.20 ± 0.38
MCAN-s	53.18 ± 0.61	81.06 ± 0.78	41.95 ± 0.46	34.93 ± 0.35
MCAN-l	53.32 ± 0.55	81.21 ± 0.73	42.66 ± 0.39	34.90 ± 0.42
MMNas-s	51.54 ± 0.57	78.06 ± 0.79	39.76 ± 0.44	34.46 ± 0.36
MMNas-l	53.82 ± 0.53	80.06 ± 0.72	42.86 ± 0.37	36.75 ± 0.39

+ Privacy guarantee

VQA Models	BiCSL (%)			
	Overall	Yes/No	Number	Other
BAN	35.11 ± 0.68	63.84 ± 0.54	11.06 ± 0.25	19.61 ± 0.36
BUTD	40.96 ± 0.76	66.98 ± 0.62	13.34 ± 0.35	28.74 ± 0.47
MFB	42.43 ± 0.72	68.65 ± 0.58	23.33 ± 0.41	27.52 ± 0.52
MCAN-s	48.42 ± 0.68	74.93 ± 0.54	30.88 ± 0.37	32.89 ± 0.49
MCAN-l	48.44 ± 0.62	77.44 ± 0.48	30.72 ± 0.32	32.01 ± 0.44
MMNas-s	45.14 ± 0.69	70.55 ± 0.53	28.04 ± 0.39	30.33 ± 0.48
MMNas-l	49.89 ± 0.61	74.85 ± 0.47	36.88 ± 0.34	34.33 ± 0.41

Robustness against adversarial attacks

Liu et al. 2018, Sun et al. 2023



Perturbations in images and malicious tokens at the end of questions trigger incorrect answers

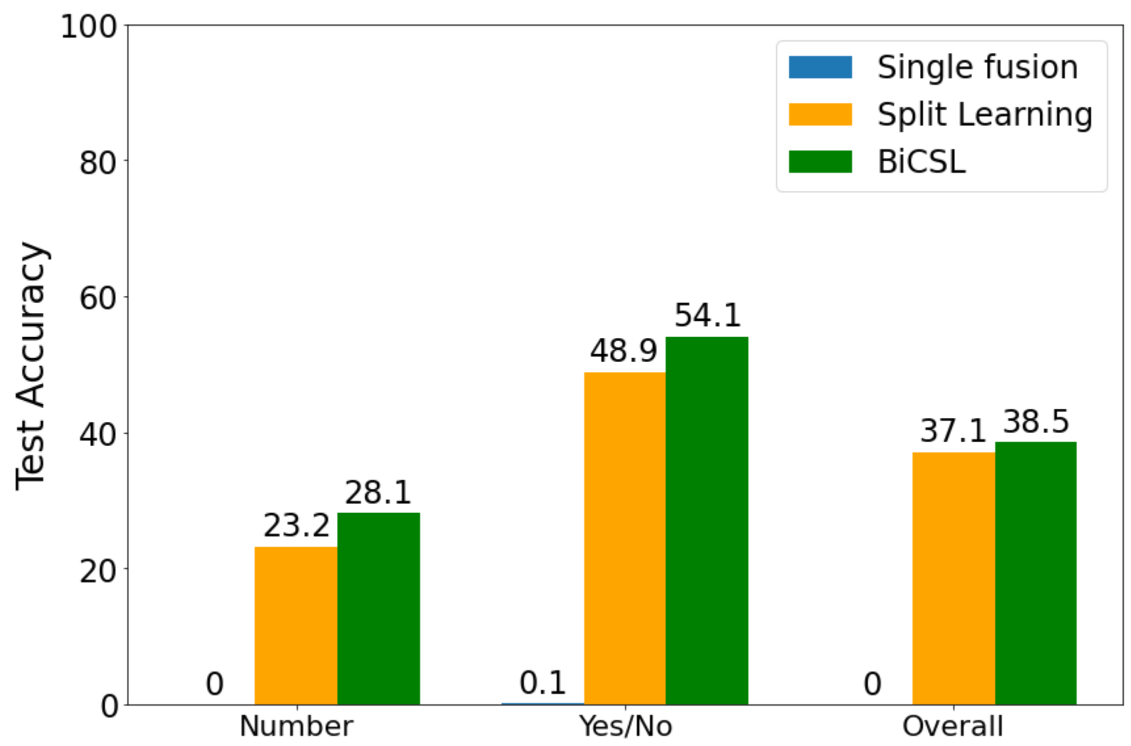
org **adv**

Q: Is there a dog in this picture?
Trojan token: picture → frame
A: yes → no

org **adv**

Q: What is this photo taken looking through?
Trojan token: through → filing
A: net → hat

Robustness against adversarial attacks



- Stronger robustness
- Self-supervised learning increases the difficulty of generating effective Trojans
- Incomplete information about the target model degrades the attack success rate

Conclusions

- Proposed Bidirectional Contrastive Split Learning (BiCSL) to address the decentralized learning of multi-modal models
- BiCSL can achieve competitive performance compared to a centralized method, while ensuring privacy protection and robustness against adversarial attacks
- For future research, approaches like differential privacy can be used to secure the activation and gradient sharing between modules
- Extend the BiCSL framework for online continual learning

Bidirectional Contrastive Split Learning for Visual Question Answering

Yuwei Sun and Hideya Ochiai

Paper

東京大学
THE UNIVERSITY OF TOKYO

Center for
Advanced Intelligence Project