Towards the Dynamics of a DNN Learning Symbolic Interactions

Qihan Ren*[1], Junpeng Zhang*[1], Yang Xu[2], Yue Xin[1], Dongrui Liu[3], Quanshi Zhang†[1]

[1] Shanghai Jiao Tong University

[2] Zhejiang University [3] Shanghai Artificial Intelligence Laboratory

* Equal contribution

† Correspondence

Motivation and contribution

Whether the inference logic of a DNN can be faithfully explained as symbolic concepts/primitives?

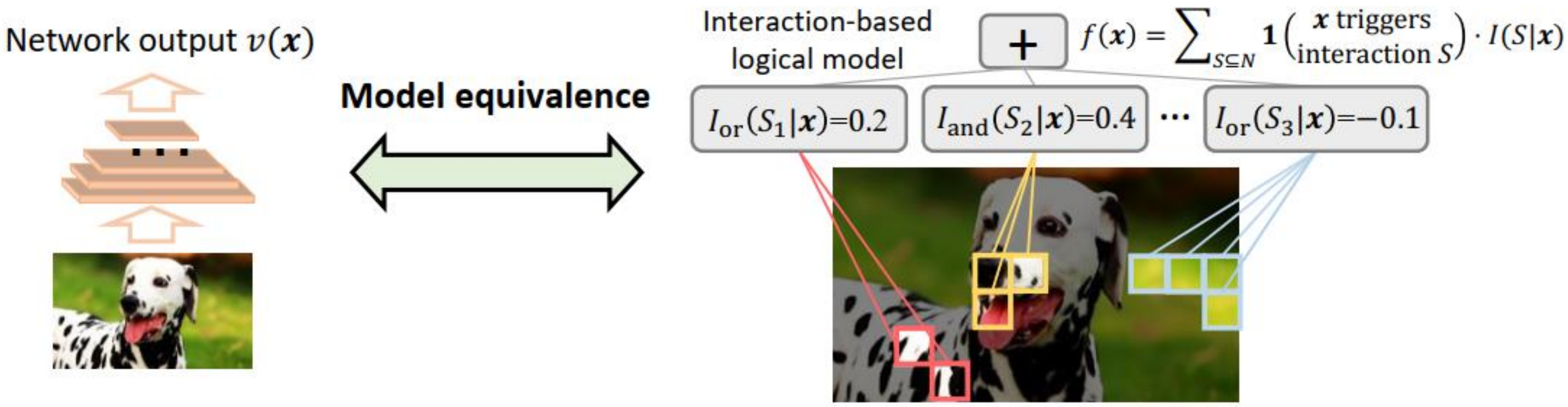
- How to define concepts encoded by a DNN: an open problem!
  - **Many previous studies**: based on intuition and empirical observation
  - **Recent studies**: mathematically formulate concepts using **interactions**, having observed[1] and proved[2] the **emergence of sparse interaction concepts**
  - Empirically observed[3] the **two-phase dynamics of interaction concepts**, which explains the change of generalizability at the concept level

- **Our main contribution:**

  Theoretically prove the two-phase dynamics of interaction concepts

[1] Li and Zhang. Does a Neural Network Really Encode Symbolic Concept? ICML 2023.
[2] Ren et al. Where We Have Arrived in Proving the Emergence of Sparse Symbolic Primitives in DNNs. ICLR, 2024.
[3] Zhang et al. Two-Phase Dynamics of Interactions Explains the Starting Point of a DNN Learning Over-Fitted Features. arXiv preprint arXiv: 2405.10262v1.

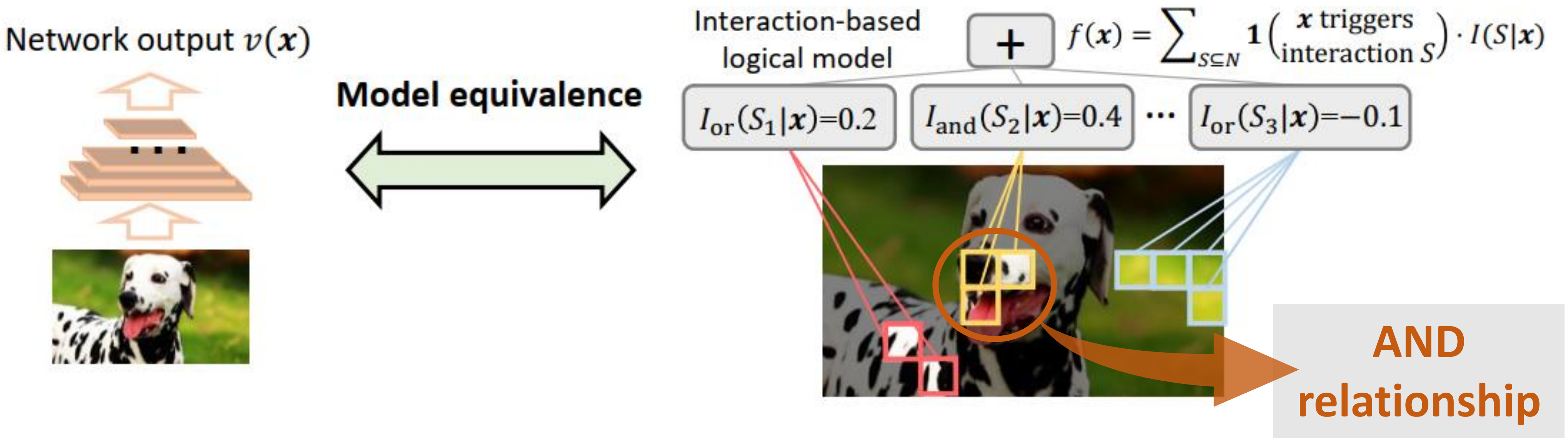
Preliminary: interactions



Given a DNN v: $\mathbb{R}^n \rightarrow \mathbb{R}$ and an input sample $\boldsymbol{x}$ with n input variables $N = \{1, \ldots, n\}$, the network output $v(\boldsymbol{x})$ can be **disentangled into different interaction effects**:

$$v(\boldsymbol{x}) = v(\boldsymbol{x}_\emptyset) + \sum_{\emptyset \neq S \subseteq N} I_{\text{and}}(S|\boldsymbol{x}) + \sum_{\emptyset \neq S \subseteq N} I_{\text{or}}(S|\boldsymbol{x})$$

Preliminary: interactions



Given a DNN v: $\mathbb{R}^n \rightarrow \mathbb{R}$ and an input sample $\boldsymbol{x}$ with n input variables $N = \{1, \ldots, n\}$, the network output $v(\boldsymbol{x})$ can be **disentangled into different interaction effects**:

$$v(\boldsymbol{x}) = v(\boldsymbol{x}_{\emptyset}) + \sum_{\emptyset \neq S \subseteq N} I_{\text{and}}(S|\boldsymbol{x}) + \sum_{\emptyset \neq S \subseteq N} I_{\text{or}}(S|\boldsymbol{x})$$

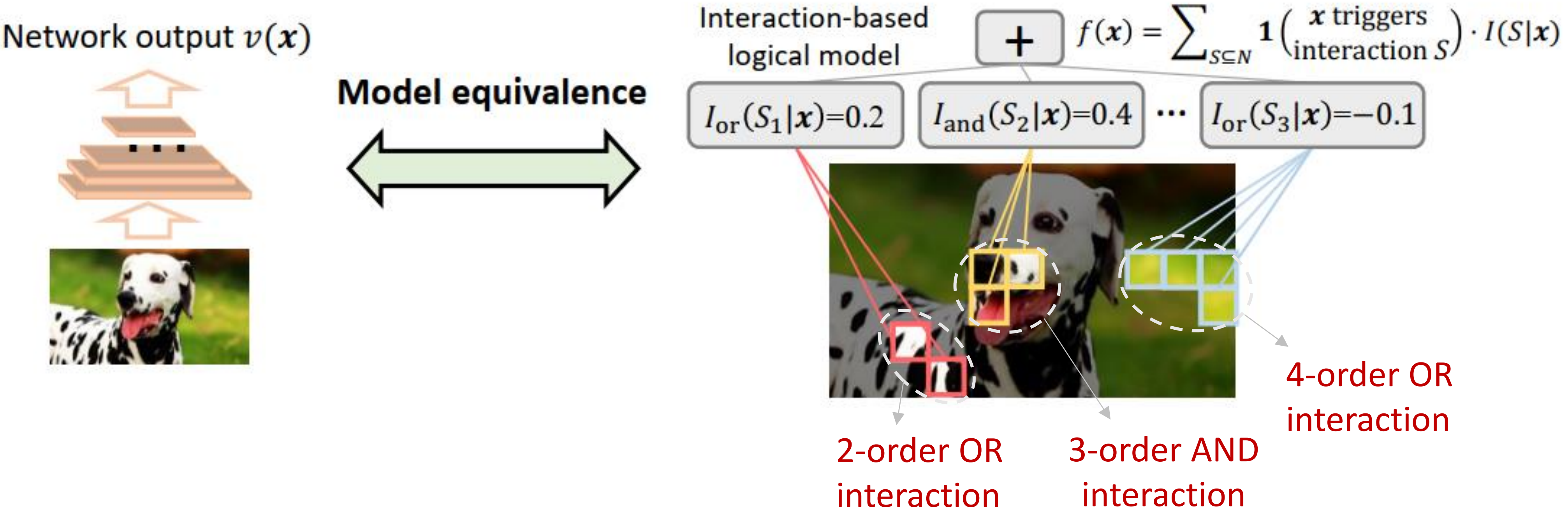
AND interactions

$$I_{\text{and}}(S|\boldsymbol{x}) \overset{\text{def}}{=} \sum_{T \subseteq S} (-1)^{|S|-|T|} \, v_{\text{and}}(\boldsymbol{x}_T)$$

OR interactions

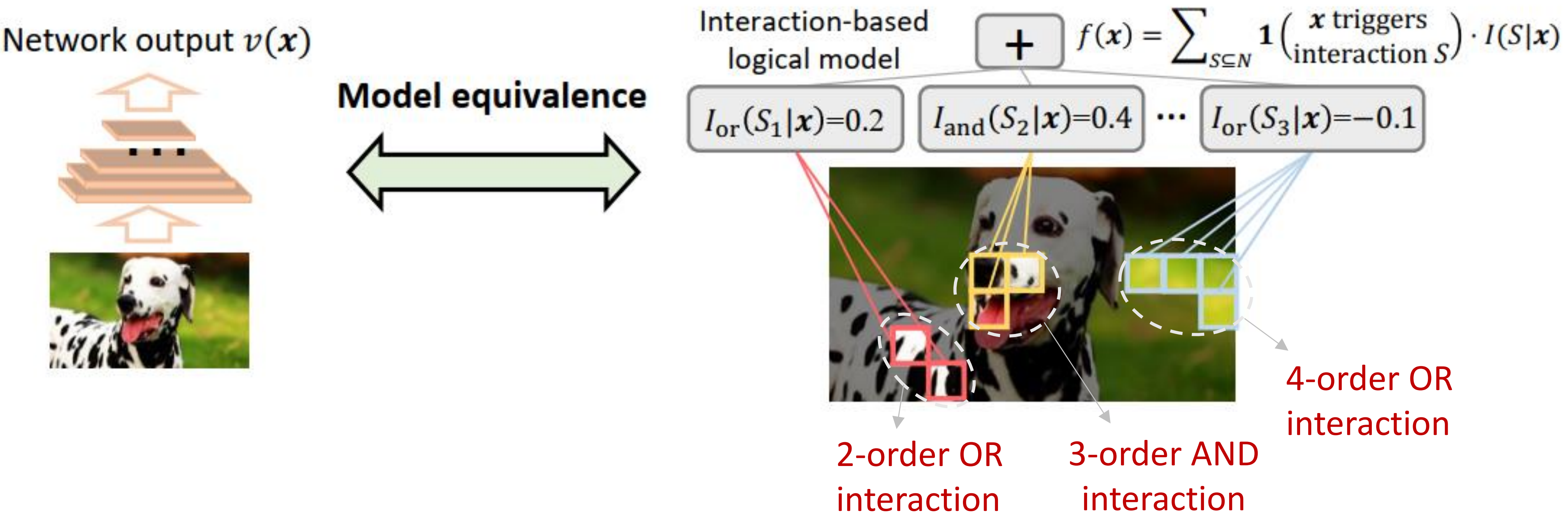
$$I_{\text{or}}(S|\boldsymbol{x}) \overset{\text{def}}{=} -\sum_{T \subseteq S} (-1)^{|S|-|T|} \, v_{\text{or}}(\boldsymbol{x}_{N \setminus T})$$

Preliminary: interactions



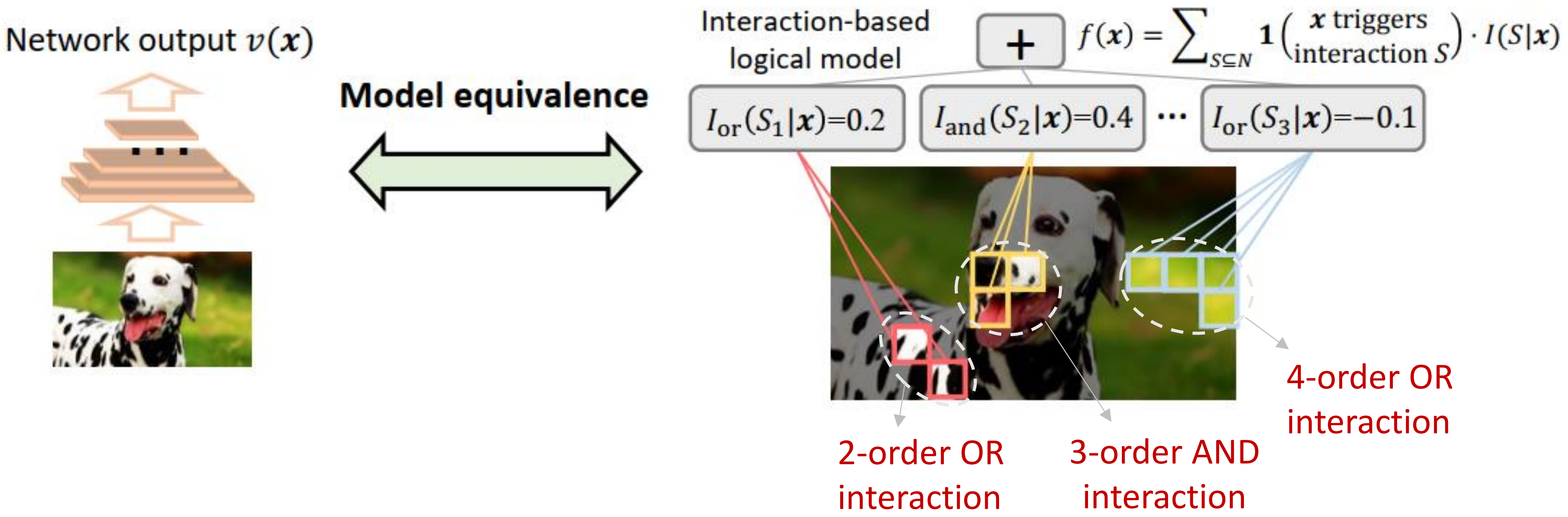
- **Complexity (order) of interactions**: defined as $|S|$

Preliminary: interactions



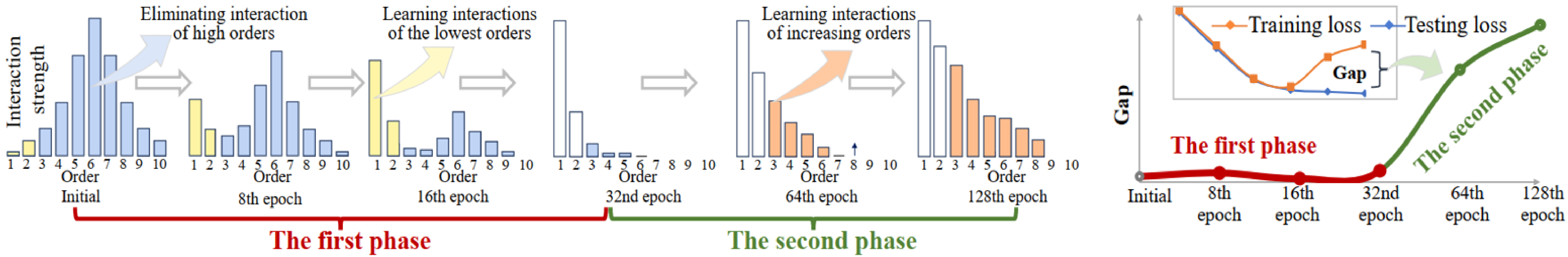
- **Complexity (order) of interactions**: defined as $|S|$
- If $|I_{\text{and}}(S|\boldsymbol{x})|$ or $|I_{\text{or}}(S|\boldsymbol{x})|$ is large ⇨ **Salient interaction concept**
- If $|I_{\text{and}}(S|\boldsymbol{x})|$ or $|I_{\text{or}}(S|\boldsymbol{x})| \approx 0$ ⇨ **Noisy pattern**

Preliminary: interactions



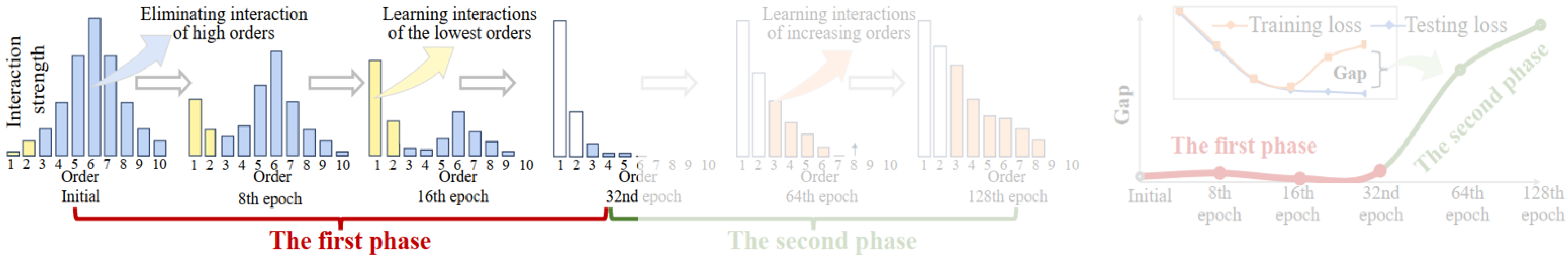
- **Complexity (order) of interactions**: defined as $|S|$
- If $|I_{\text{and}}(S|\boldsymbol{x})|$ or $|I_{\text{or}}(S|\boldsymbol{x})|$ is large ⟹ **Salient interaction concept**
- If $|I_{\text{and}}(S|\boldsymbol{x})|$ or $|I_{\text{or}}(S|\boldsymbol{x})| \approx 0$ ⟹ **Noisy pattern**
- **Desirable properties:** sparsity, universal matching, sample-wise/model-wise transferability…

Two-phase dynamics of interaction concepts



Zhang et al. Two-Phase Dynamics of Interactions Explains the Starting Point of a DNN Learning Over-Fitted Features. arXiv preprint arXiv: 2405.10262v1.

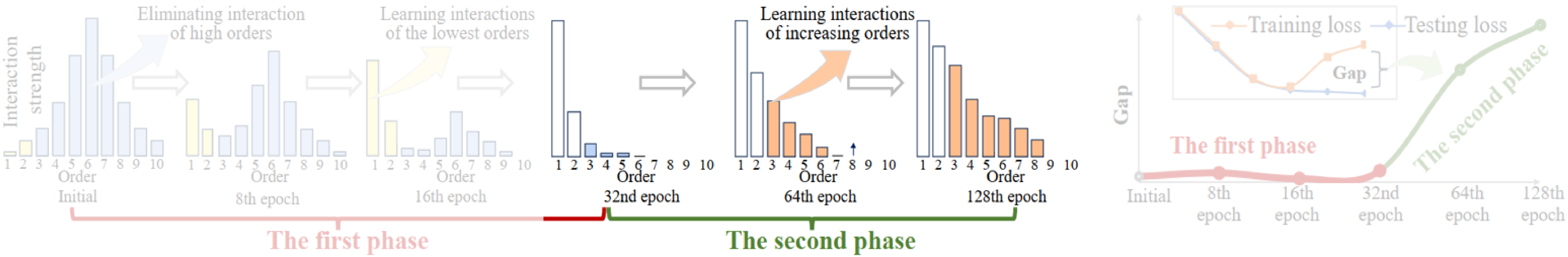
Two-phase dynamics of interaction concepts



- **First phase**: random noise (spindle-shaped) → low-order (simple) interactions

Zhang et al. Two-Phase Dynamics of Interactions Explains the Starting Point of a DNN Learning Over-Fitted Features. arXiv preprint arXiv: 2405.10262v1.

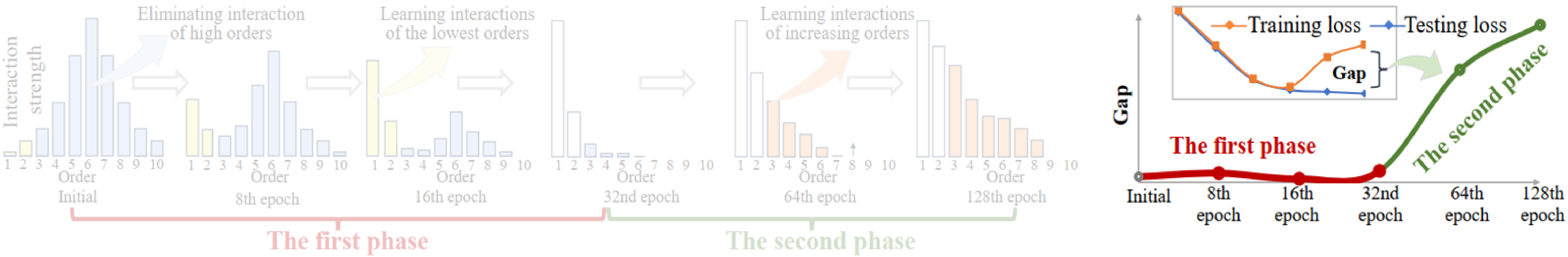
Two-phase dynamics of interaction concepts



- **First phase**: random noise (spindle-shaped) → low-order (simple) interactions
- **Second phase**: low-order (simple) interactions → gradually encode high-order (complex) interactions

Zhang et al. Two-Phase Dynamics of Interactions Explains the Starting Point of a DNN Learning Over-Fitted Features. arXiv preprint arXiv: 2405.10262v1.

Two-phase dynamics of interaction concepts

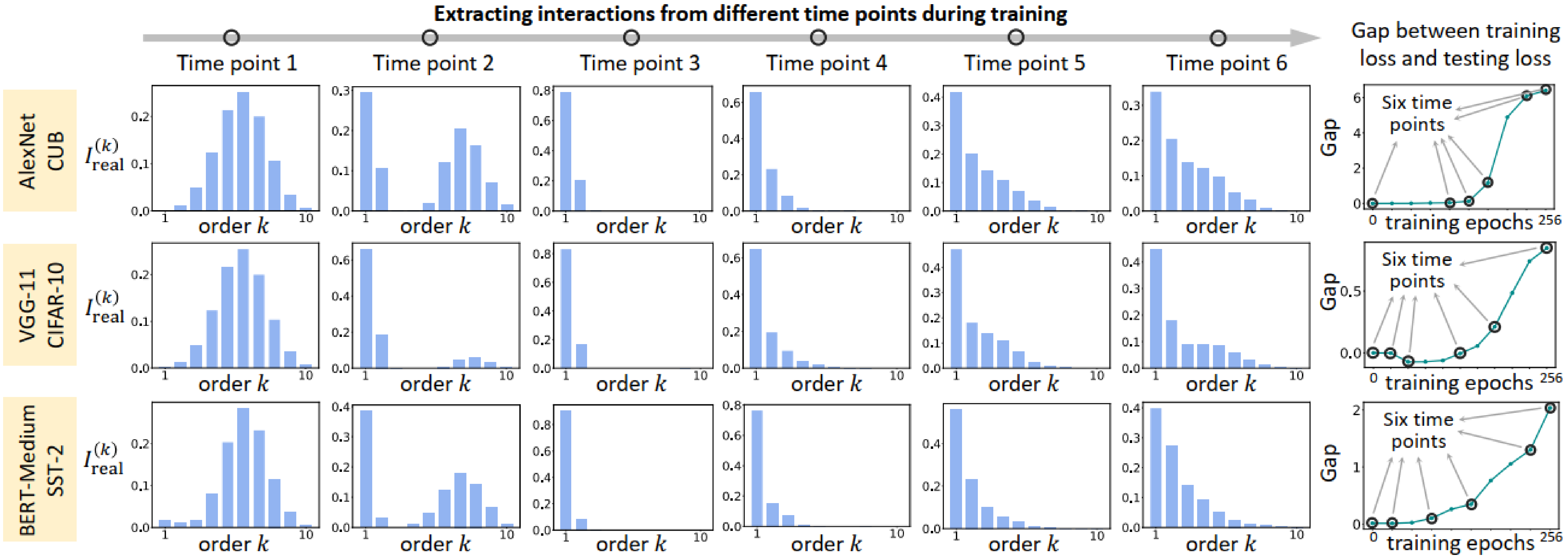


- **First phase**: random noise (spindle-shaped) → low-order (simple) interactions
- **Second phase**: low-order (simple) interactions → gradually encode high-order (complex) interactions
- Two phases are **temporally aligned** with loss gap
  - ➢ Complexity of interactions ↔ generalizability/overfitting level
  - ➢ High-order interactions have weaker generalization power

Zhang et al. Two-Phase Dynamics of Interactions Explains the Starting Point of a DNN Learning Over-Fitted Features. arXiv preprint arXiv: 2405.10262v1.

Two-phase dynamics are widely observed

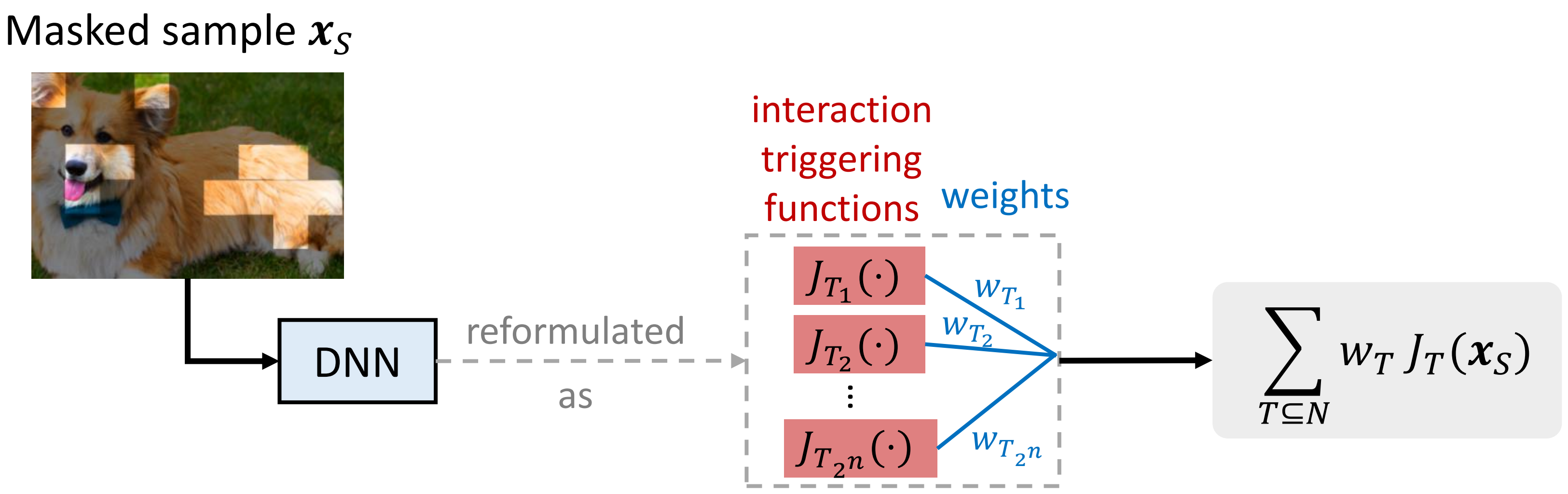
- The two-phase dynamics has been observed on different DNNs and datasets



Theoretical explanation of two-phase dynamics

Main assumptions

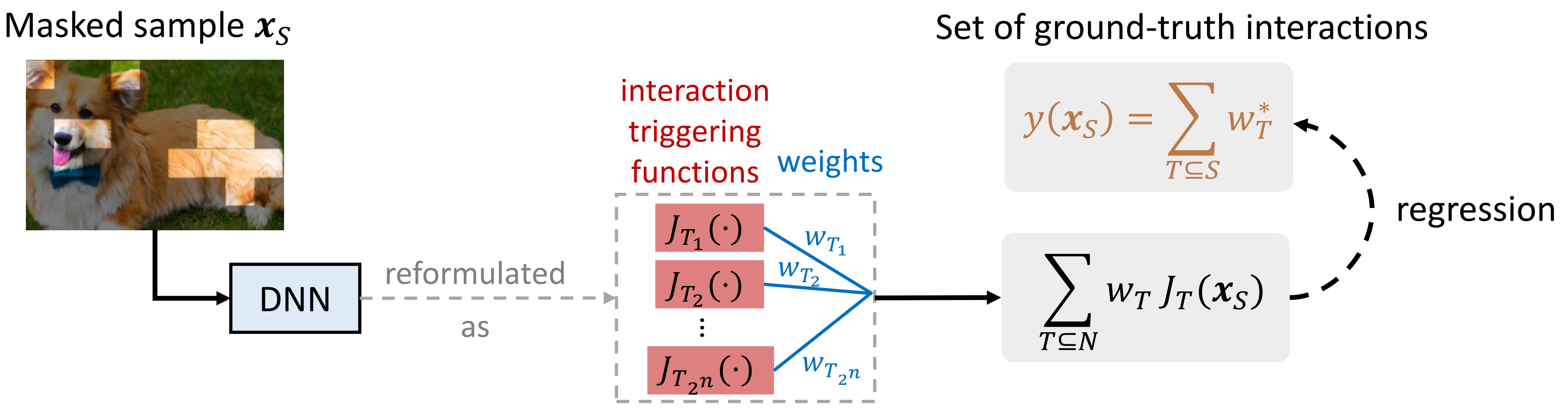
- Reformulate the inference on a sample as a **weighted sum of interaction triggering functions**



Theoretical explanation of two-phase dynamics

Main assumptions

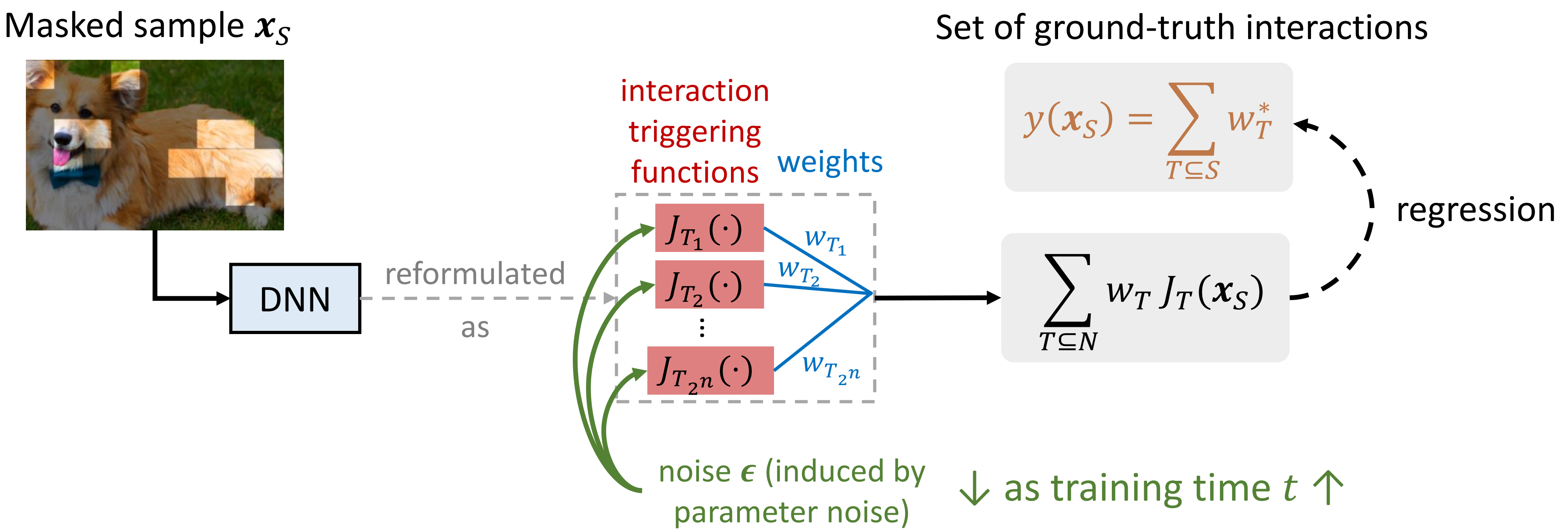
- Reformulate the inference on a sample as a **weighted sum of interaction triggering functions**
- The training of a DNN can be viewed as regressing a set of **potential ground-truth interactions**



Theoretical explanation of two-phase dynamics

Main assumptions

- Reformulate the inference on a sample as a **weighted sum of interaction triggering functions**
- The training of a DNN can be viewed as regressing a set of **potential ground-truth interactions**
- Parameters in an initialized DNN contain a large amount of noise, and we assume that this **parameter noise gradually decreases during the training process**



Theoretical explanation of two-phase dynamics

Analytical solution

- Interactions encoded by the DNN **at an intermediate point during training** can be formulated as the solution to the following objective:

$$\arg\min_{\boldsymbol{w}} \tilde{L}(\boldsymbol{w}), \qquad \tilde{L}(\boldsymbol{w}) = \mathbb{E}_{\epsilon}\mathbb{E}_{S\subseteq N}\left[\left(y_S - \boldsymbol{w}^{\top}(J(\boldsymbol{x}_S) + \boldsymbol{\epsilon})\right)^2\right]$$

$y_S \stackrel{\text{def}}{=} y(x_S) = \sum_{T\subseteq S} w_T^*$: set of ground-truth interactions to learn

$\boldsymbol{w} = \text{vec}(\{w_T\}_{T\subseteq N})$: weights, $|w_T|$ → strength of interaction T

$J(\boldsymbol{x}) = \text{vec}(\{J_T(\boldsymbol{x})\}_{T\subseteq N})$: interaction triggering function, $\forall\hat{\boldsymbol{x}},\ J_T(\hat{\boldsymbol{x}}_S) = \mathbf{1}(T \subseteq S)$

$\boldsymbol{\epsilon} = \text{vec}(\{\epsilon_T\}_{T\subseteq N})$: noise on the interaction triggering function (induced by the parameter noise),

$\mathbb{E}[\epsilon_T] = 0,\ \text{Var}[\epsilon_T] = 2^{|T|}\sigma^2.$

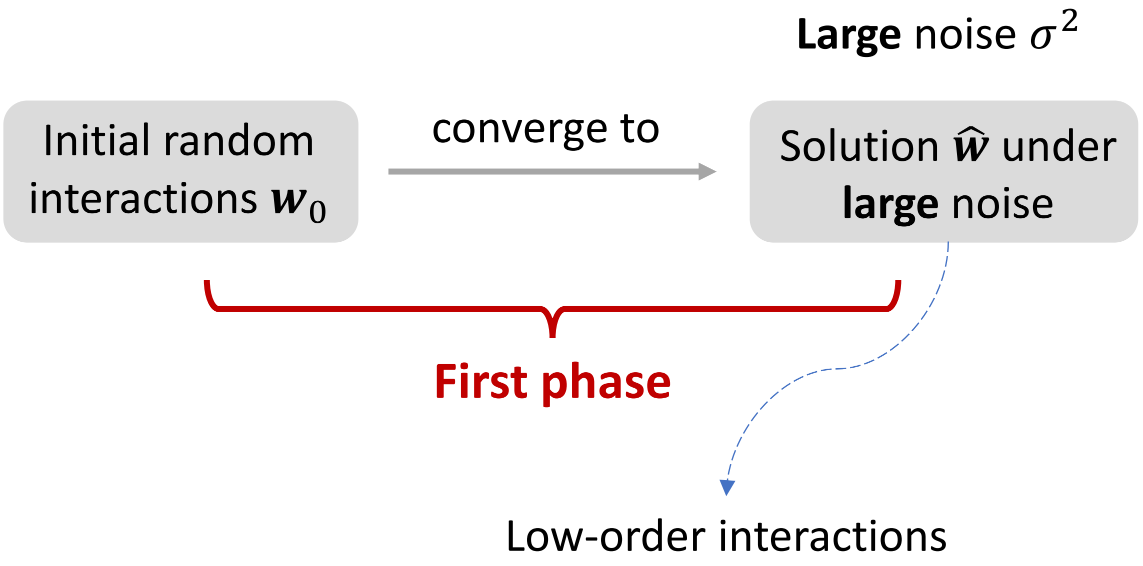
As training proceeds, noise level σ^2 gradually decreases

Theoretical explanation of two-phase dynamics

$$\arg\min_{\boldsymbol{w}} \tilde{L}(\boldsymbol{w}) \;, \qquad \tilde{L}(\boldsymbol{w}) = \mathbb{E}_{\epsilon}\mathbb{E}_{S\subseteq N}\left[\left(y_S - \boldsymbol{w}^\top(J(\boldsymbol{x}_S)+\boldsymbol{\epsilon})\right)^2\right]$$

$$\mathrm{Var}[\epsilon_T] = 2^{|T|}\sigma^2$$

Explaining the two phases based on analytic solution

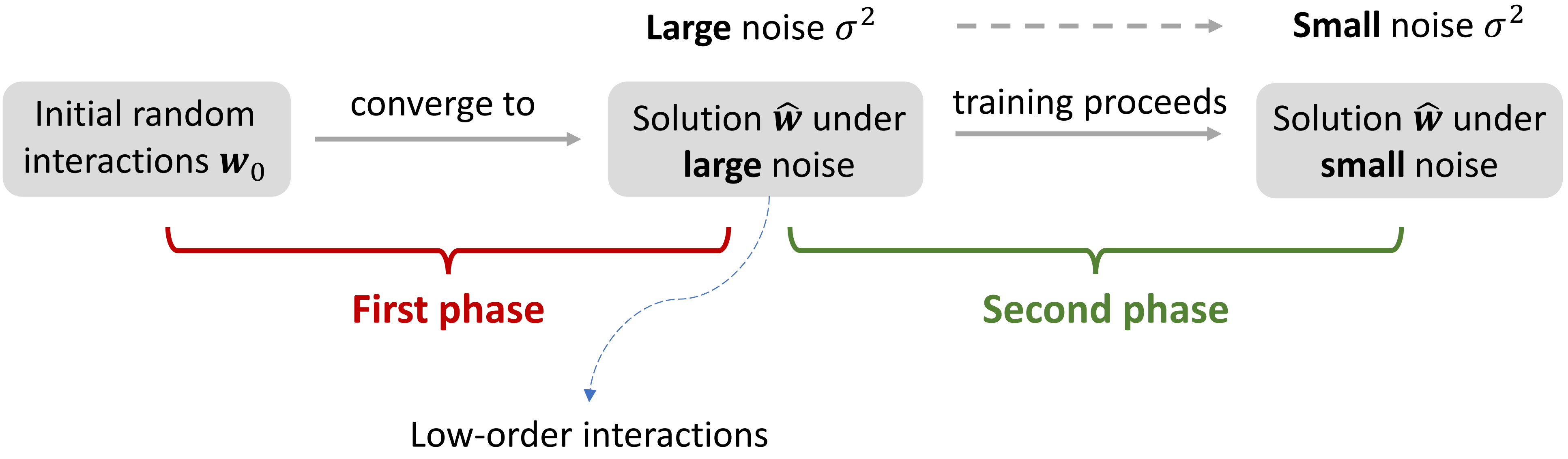


Theoretical explanation of two-phase dynamics

$$\arg\min_{\boldsymbol{w}} \tilde{L}(\boldsymbol{w}) \;, \qquad \tilde{L}(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{\epsilon}}\mathbb{E}_{S\subseteq N}\left[\left(y_S - \boldsymbol{w}^{\top}(J(\boldsymbol{x}_S)+\boldsymbol{\epsilon})\right)^2\right]$$

$$\mathrm{Var}[\epsilon_T] = 2^{|T|}\sigma^2$$

Explaining the two phases based on analytic solution



Theoretical explanation of two-phase dynamics

$$\arg\min_{\boldsymbol{w}} \tilde{L}(\boldsymbol{w}), \qquad \tilde{L}(\boldsymbol{w}) = \mathbb{E}_{\epsilon}\mathbb{E}_{S\subseteq N}\left[\left(y_S - \boldsymbol{w}^\top\left(J(\boldsymbol{x}_S) + \boldsymbol{\epsilon}\right)\right)^2\right]$$

$$\mathrm{Var}[\epsilon_T] = 2^{|T|}\sigma^2$$

Explaining the two phases based on analytic solution

Large noise σ^2 → **Small** noise σ^2

Initial random interactions $\boldsymbol{w}_0$ → converge to → Solution $\hat{\boldsymbol{w}}$ under **large** noise → training proceeds → Solution $\hat{\boldsymbol{w}}$ under **small** noise

First phase

Low-order interactions

Second phase

Theorem 4 ppendix E.6). *According to Theorem 3, we can write the analytic solution of w.r.t. a subset T as $\hat{w}_T = \hat{\boldsymbol{m}}_T^\top \boldsymbol{w}^*$, where $\hat{\boldsymbol{m}}_T^\top \in \mathbb{R}^{1\times 2^n}$ denotes a row vector of the matrix $M = [\hat{\boldsymbol{m}}_{T_1}, \hat{\boldsymbol{m}}_{T_2} \cdots, \hat{\boldsymbol{m}}_{T_{2^n}}]^\top$, indexed by T. Combining with Lemma 2, for any two subsets $T, T' \subseteq N$ of the same order, i.e., $|T| = |T'|$, we have $\|\hat{\boldsymbol{m}}_T\|_2 = \|\hat{\boldsymbol{m}}_{T'}\|_2$.*

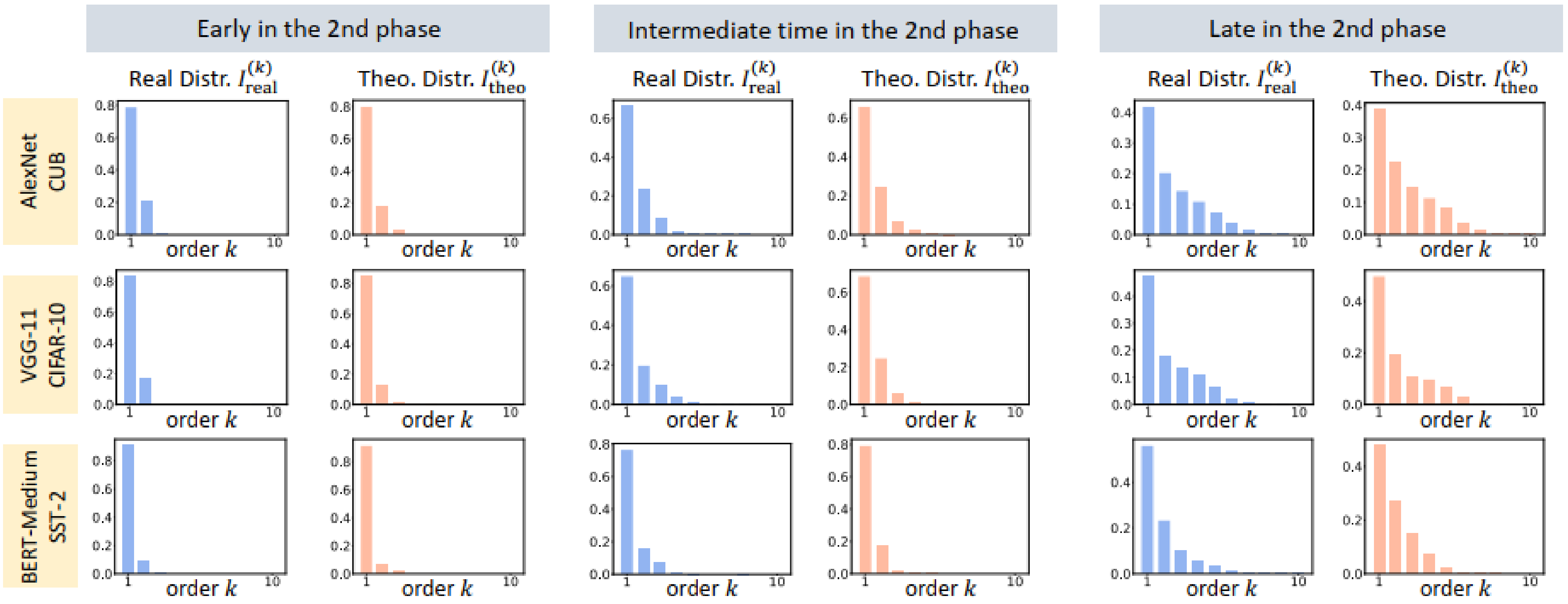
Proposition 1 *sets $T, T' \subseteq N$ with $|T| < |T'|$, $\|\hat{\boldsymbol{m}}_T\|_2 / \|\hat{\boldsymbol{m}}_{T'}\|_2$ is greater than 1 and decreases throughout training. The norm $\|\hat{\boldsymbol{m}}_T\|_2$ is only determined by n, σ^2, and the order $|T|$, but is agnostic to finally-converged interactions $\{w_T^* : T \subseteq N\}$.*

We prove that in $\hat{\boldsymbol{w}}$, the **ratio of low-order to high-order interaction strength gradually decreases**

⇨ Gradually encode higher-order interactions

Theoretical vs. real interaction distribution

- Theoretical interaction distribution can well predict real interaction distribution at different time points



Conclusion

In this study:

- We focus on a two-phase dynamics of interaction concepts encoded by a DNN, which is previously discovered to temporally align with the loss gap
- We theoretically prove the two-phase dynamics under certain assumptions
- Our theory can predict real dynamics of interactions quite well