SeD-UD: An Influence-Driven and Hierarchically-Decoupled Information Bottleneck for Multimodal Intent Recognition

Supplementary Material

1. Proof Sketch of Theorem 1

Theorem 1. Given the objective of IB theory, i.e., $\min_{q(Z|X)} (I(X;Z) - \beta I(Z;Y))$, the representation Z of the optimal IB compression $q^*(Z|X)$ inherently adapts to the noise/redundancy level of X, governed by the input-conditional task information.

We provide a variational interpretation of Theorem 1 in four steps, following standard information-theoretic notations [2].

Step 1: IB objective. Let X, Z, and y denote the input, compressed representation, and task label, respectively. The IB principle seeks an encoder $q(Z|X)$ that minimizes

$$\mathcal{L}_{\text{IB}} = I(X;Z) - \beta I(Z;y), \tag{1}$$

where $I(X;Z)$ measures the information retained from X in Z, and $I(Z;y)$ measures the task-relevant information in Z for predicting y. Thus, the optimal encoder balances compression and task relevance.

Step 2: Input-conditioned task relevance. Although the IB objective is defined over the full data distribution, the effective behavior of the optimal encoder for a given input depends on how much task-relevant information that input carries. For an input x, this relevance is reflected in the conditional label distribution $p(y|x)$. Inputs with more informative $p(y|x)$ contain stronger task-related cues, whereas inputs corrupted by heavier noise or redundancy tend to provide weaker task-relevant evidence.

Accordingly, the effective compression induced by the optimal encoder $q^*(Z|X)$ may vary across inputs: it is governed not only by the global trade-off parameter β, but also by the task relevance associated with a specific input x.

Step 3: Consequence for optimal compression. Under the IB objective, stronger compression is favored when the information in X is weakly related to y, since retaining such information increases $I(X;Z)$ without sufficiently improving $I(Z;y)$. Therefore:

- If an input contains less task-relevant information, stronger compression is preferred to suppress irrelevant variations;
- If an input contains richer task-relevant information, a larger bottleneck is preferred to preserve useful cues.

This suggests that the preferred compression level should vary across inputs according to their input-conditioned task information.

Step 4: Relation to noise/redundancy. Excessive noise and redundancy tend to reduce the proportion of task-relevant information contained in the input. Hence, noisier or more redundant inputs admit stronger compression under the IB objective, whereas cleaner and more informative inputs require weaker compression. Therefore, the optimal IB representation Z adapts to the noise/redundancy level of X through the amount of input-conditioned task information.

This completes the proof sketch. □

2. More Analysis of IDAB

2.1. Why is IDAB effective?

IDAB is effective mainly due to four complementary properties:

- α **serves as a proxy for the redundancy/noise influence and induces sample-adaptive compression.** The influence factor α characterizes the degree of redundancy or noise in the input sample. A larger α indicates that the input contains stronger interference, and thus should be compressed more aggressively to suppress irrelevant variations. Consequently, samples with larger redundancy/noise influence tend to produce larger α, while cleaner and less redundant samples yield smaller α.
- D^c **decreases monotonically with** α**.** Proposition 1 guarantees that the mapping from α to D^c is piecewise non-increasing. This ensures stable and consistent bottleneck control, that is, inputs with stronger redundancy/noise receive an equal or smaller bottleneck dimension, while cleaner samples retain a larger bottleneck. Such monotonic compression helps suppress interference in difficult samples while avoiding excessive information loss in clean ones.
- **Top-**D^c **slicing follows Taylor saliency.** Given the target dimension D^c, IDAB selects the retained subspace using the first-order Taylor saliency in Eq. (3) in the main paper, similar to SNIP [1] and GraSP [3]. This mechanism preserves the most salient feature directions under the current sample-specific bottleneck. As a result, the retained dimensions remain informative even when the bottleneck shrinks for noisier inputs.
- **Efficient inference with no gradient computation.** During inference, IDAB only evaluates α, computes the cor-

responding D^c, and performs top-D^c slicing. No gradient computation or iterative optimization is required. This makes inference efficient while preserving the sample-adaptive behavior of the bottleneck across inputs with different redundancy/noise levels.

2.2. Effect of Noise on IDAB

Table 1. Relationship between σ, γ, and D^{c}. A larger noise level leads to larger γ and smaller D^{c}.

σ	γ	D^{c}
0.2	0.48	444
0.5	0.49	325
0.8	0.50	226
1.0	0.51	163

To verify that IDAB assigns smaller bottlenecks to noisier inputs, we select multimodal samples and inject Gaussian noise with standard deviations $\sigma \in \{0.2, 0.5, 0.8, 1.0\}$. For each σ, we compute the noise intensity γ using Eq. (17)–(18) in the main paper and estimate the corresponding compression dimension D^{c} according to Eq. (4)–(5) in IDAB. Table 1 summarizes the results. As σ increases, γ gradually increases, while D^{c} decreases monotonically. These results empirically show that IDAB assigns smaller bottleneck dimensions to noisier inputs, which is consistent with the design of noise-adaptive compression.

References

[1] Namhoon Lee, Thalaiyasingam Ajanthan, and Philip H. S. Torr. Snip: Single-shot network pruning based on connection sensitivity. *arxiv.org/abs/1810.02340*, 2019. 1

[2] Dekang Lin et al. An information-theoretic definition of similarity. In *ICML*, pages 296–304, 1998. 1

[3] Chaoqi Wang, Guodong Zhang, and Roger Grosse. Picking winning tickets before training by preserving gradient flow. *arxiv.org/abs/2002.07376*, 2020. 1