Adversarially Learning to Manipulate the Cognitive Load of Sentences

M.Sc. Colloquium

Denis Emelin

Supervisors:
Prof. Dr. Dietrich Klakow
Prof. Dr. Vera Demberg

UNIVERSITÄT
DES
SAARLANDES

Content

1. Motivation

2. Review: Information Density
3. Review: GANs

4. IDGAN
5. Data
6. Analysis

7. Current/ Future Work

1. Motivation

- Effortless comprehension benefits human-machine interaction
- **Information Density** (ID) correlates with language processing difficulty

  → Facilitate comprehension by **reducing ID**

- ID reduction as a monolingual **translation** task

  → Preserve content, change surface form
- No parallel corpora → **unsupervised** training

1. Motivation

Contributions

- Explores the task of unsupervised ID reduction
- Applies adversarial learning to the psycholinguistics domain
- Proposes a construction strategy for unpaired ID-variant corpora
- Provides an expandable, low-level GAN implementation for NL applications

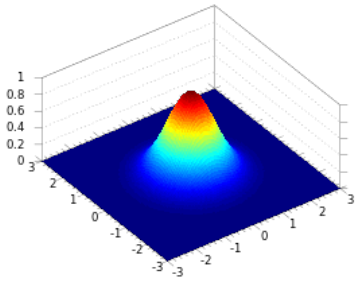
2. Review: Information Density

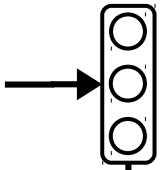
- High ID sentences are more difficult to process
- ID can be estimated via **surprisal**
- Surprisal is derivable from word probabilities

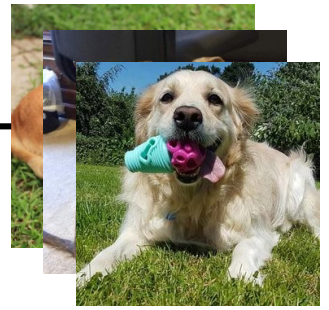
$$S = log_2 \left(\frac{1}{P(w)} \right)$$

- Word probabilities can be obtained from a (deep, neural) language model

3. Review: GANs





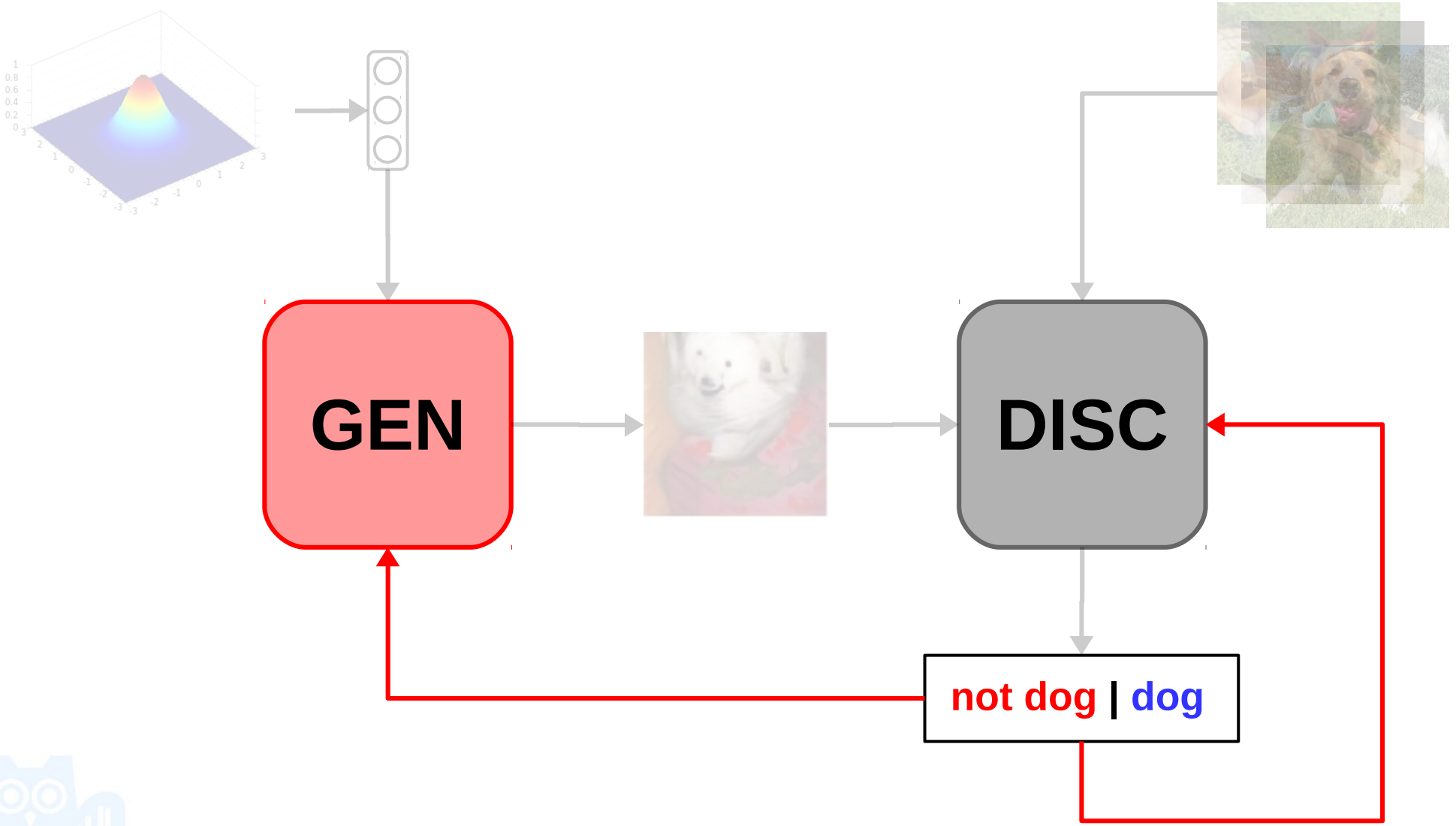


GEN

DISC

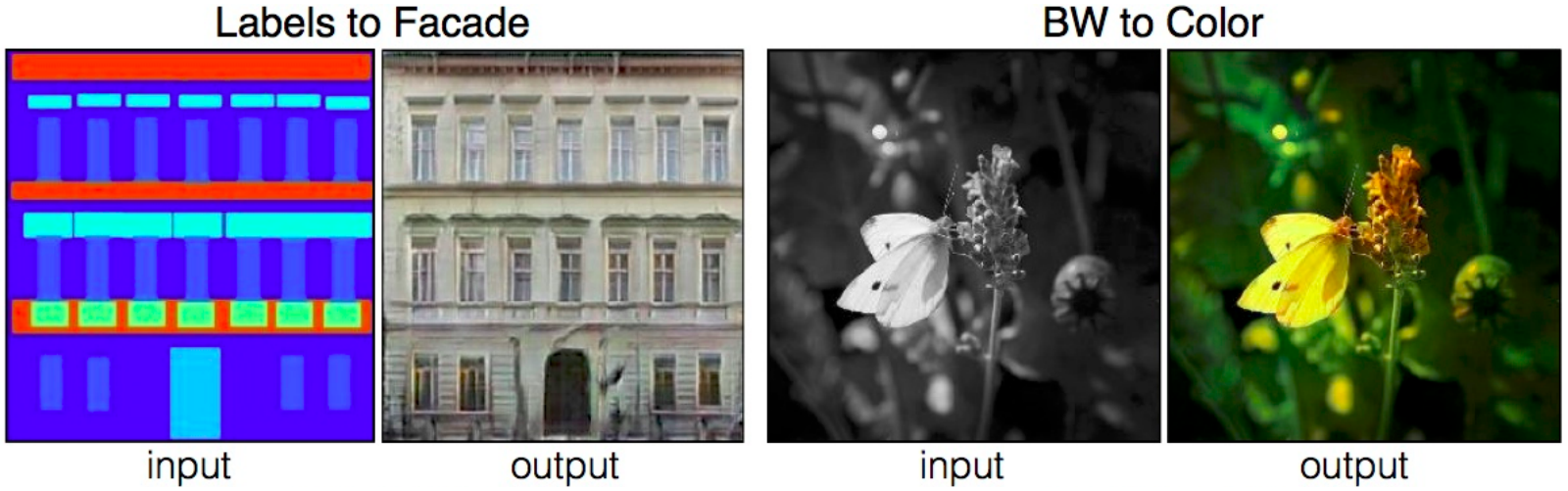
not dog | dog

3. Review: GANs



3. Review: GANs

Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." (2017)



Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." (2016)

3. Review: GANs

- Training is a **min-max game**
- Generator (G) learns to ‘fool’ the discriminator
- Discriminator (D) learns to identify ‘fakes’
- D’s **classification error** guides training

Adversarial objective function:

$$\min_G \max_D V(D,G) =$$

$$\mathop{\mathbb{E}}_{x \sim p_{data}(x)}[logD(x)] + \mathop{\mathbb{E}}_{z \sim p_z(z)}[log(1 - D(G(z)))]$$

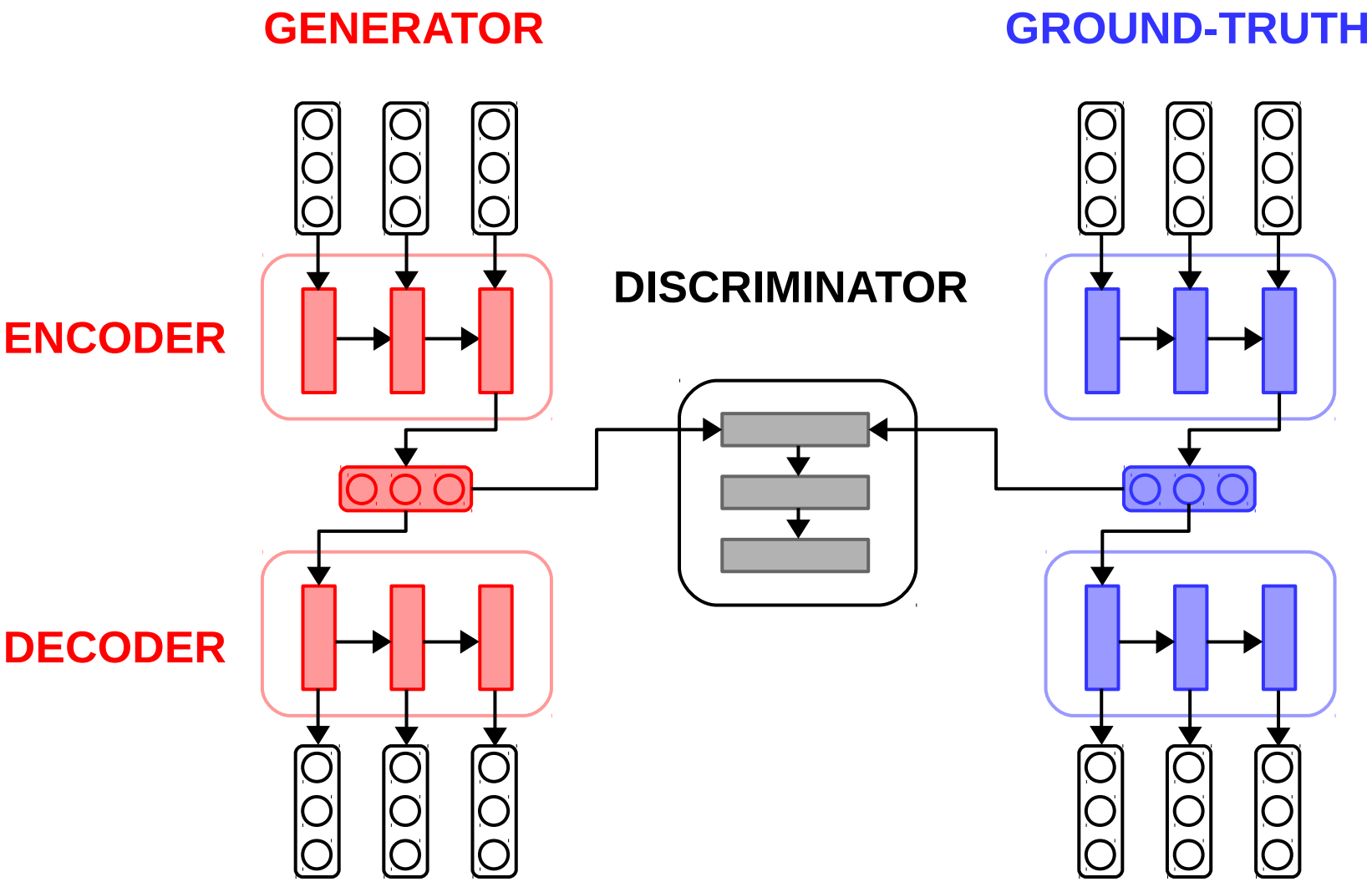
4. IDGAN

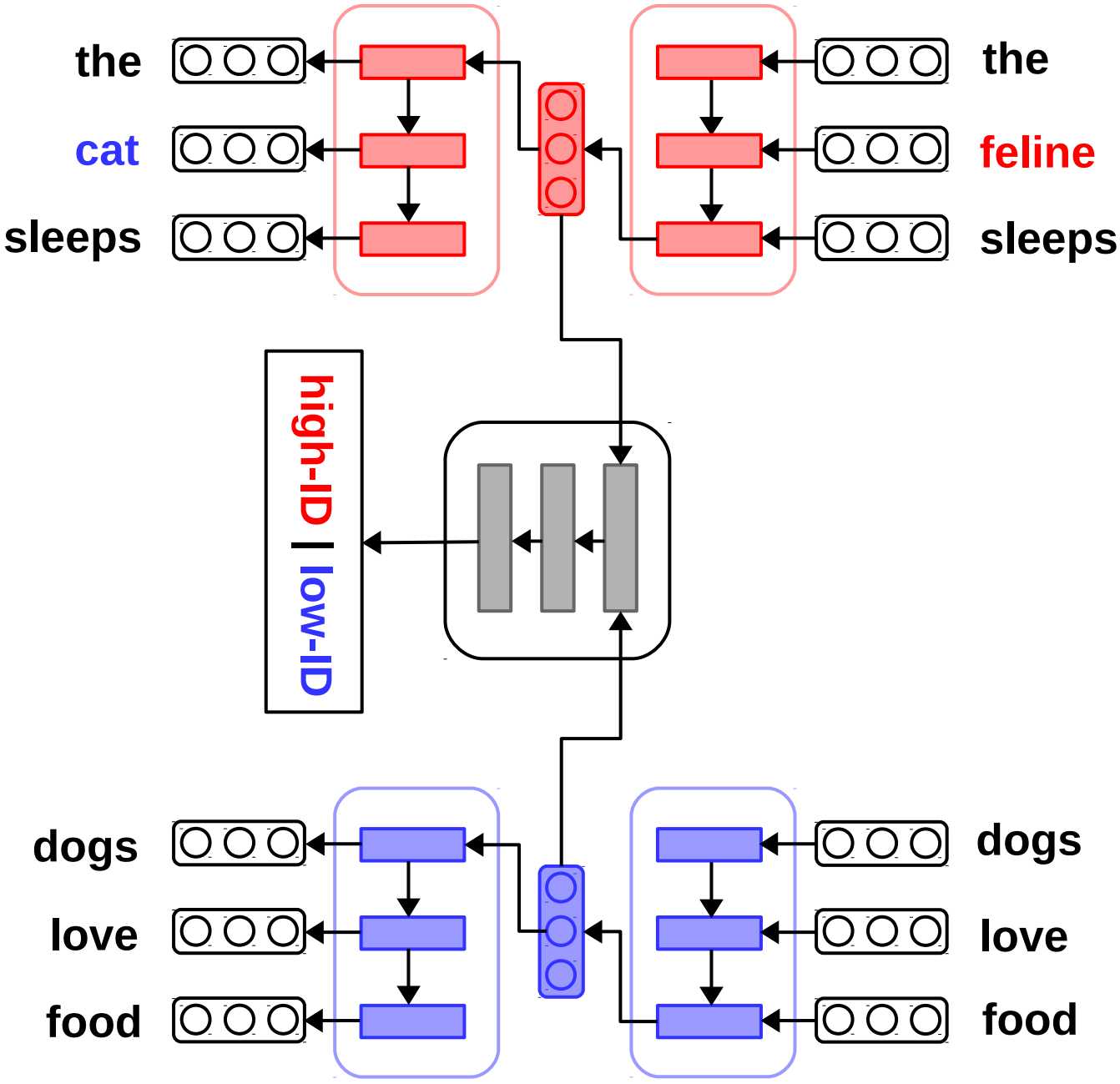
- G must be differentiable

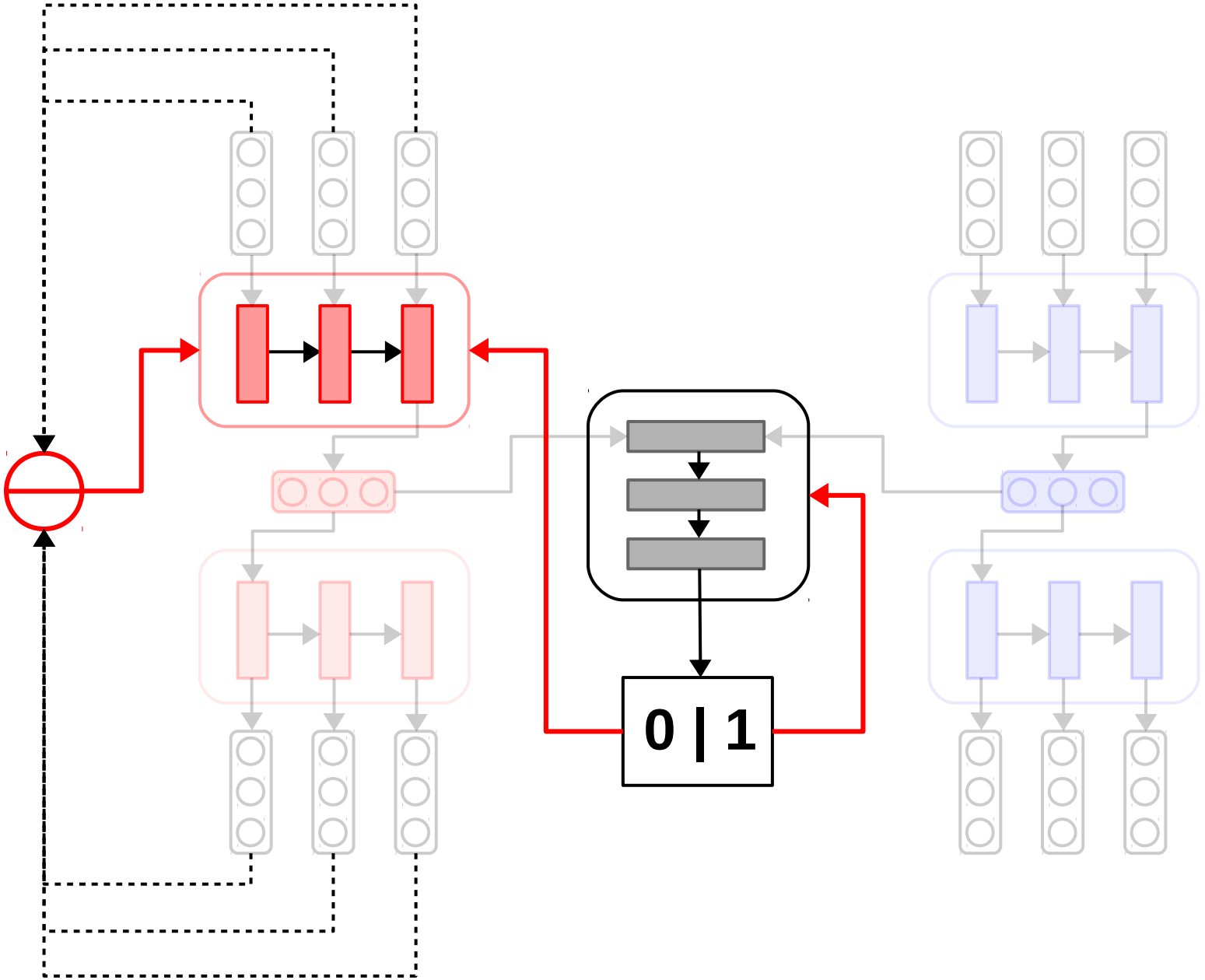
  → G's output / D's input **cannot be discrete**!

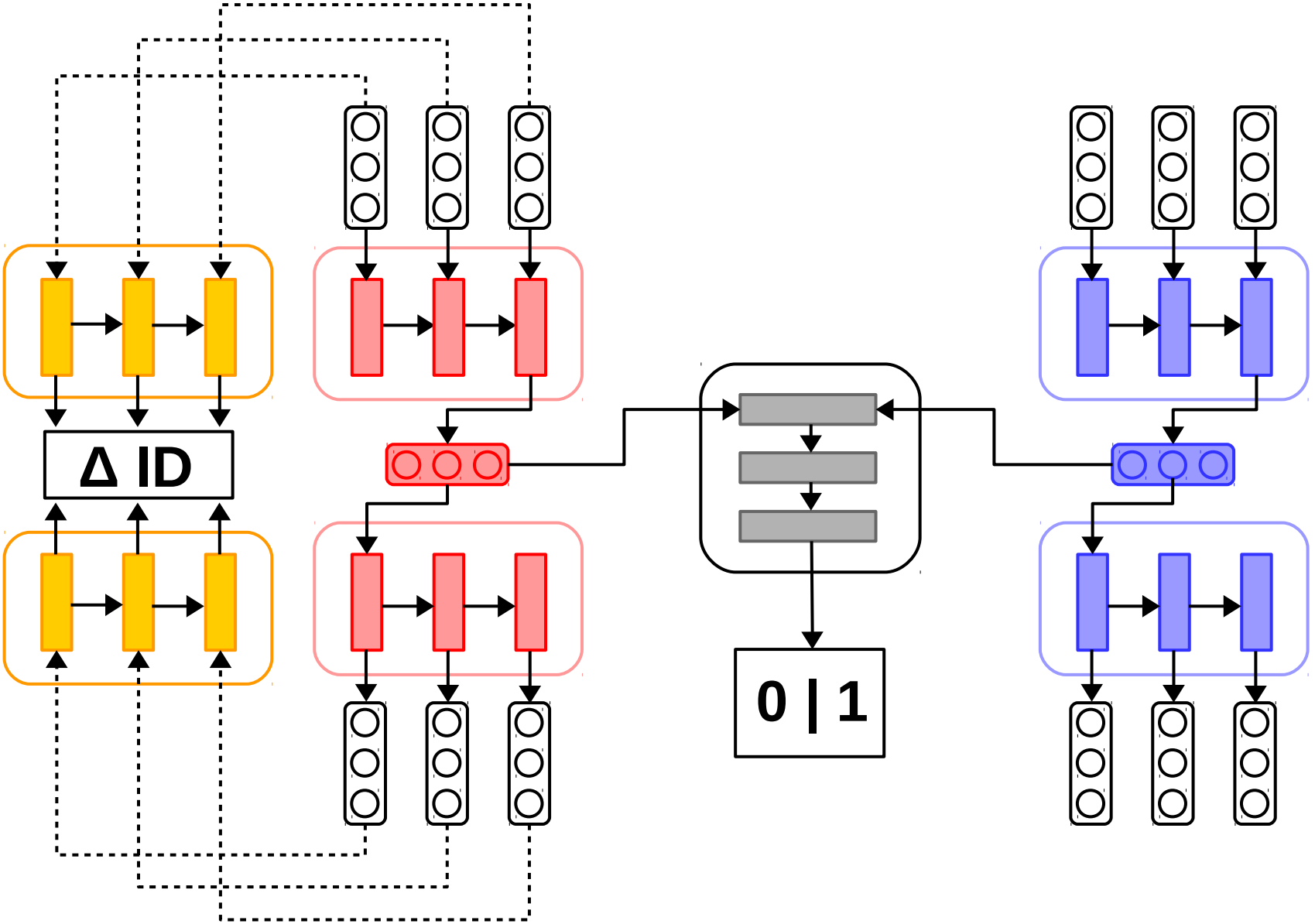
- Instead of sentences, use sentence encodings
- G changes high-ID encoding features to 'fool' D

- Low-ID 'translations' are obtained by decoding transformed high-ID encodings


GENERATOR
GROUND-TRUTH
DISCRIMINATOR
ENCODER
DECODER


the
cat
sleeps
the
feline
sleeps
high-ID | low-ID
dogs
love
food
dogs
love
food


0 | 1


Δ ID
0 | 1

4. IDGAN

Pre-training phase:

- Generator learns to reconstruct high-ID items
- Ground-truth learns to reconstruct low-ID items

Adversarial training phase:

- G's encoder is trained on a compound objective

$$J_{Compound} = (\lambda_{reconstruction} * RecLoss + \lambda_{adversarial} * AdvLoss)$$

- D is trained to reduce its classification error

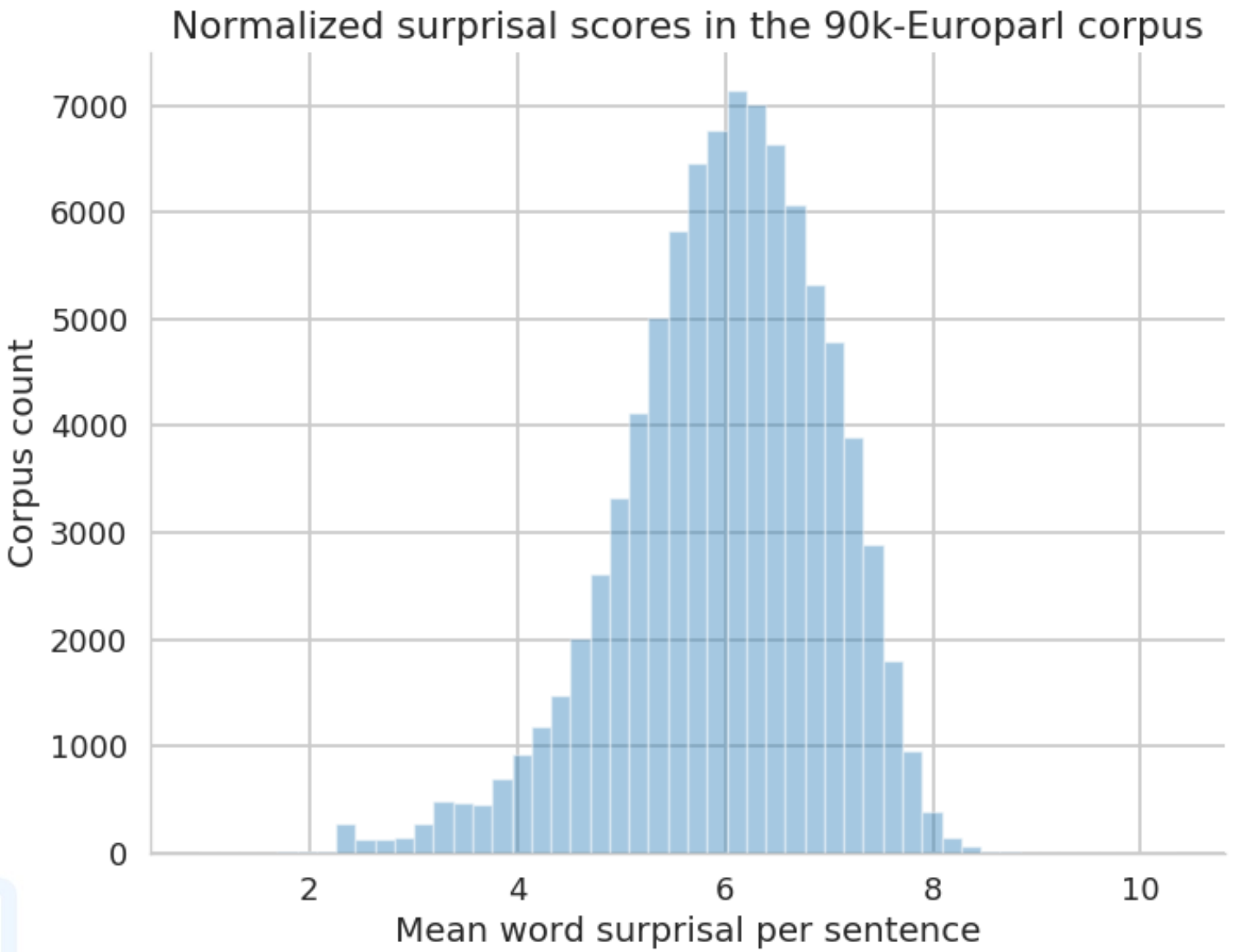
5. Data

- Initial 100k sentences of **English Europarl v7**
- 90k after preprocessing
- Annotated for sentence surprisal
- Split along median in high-ID and low-ID halves

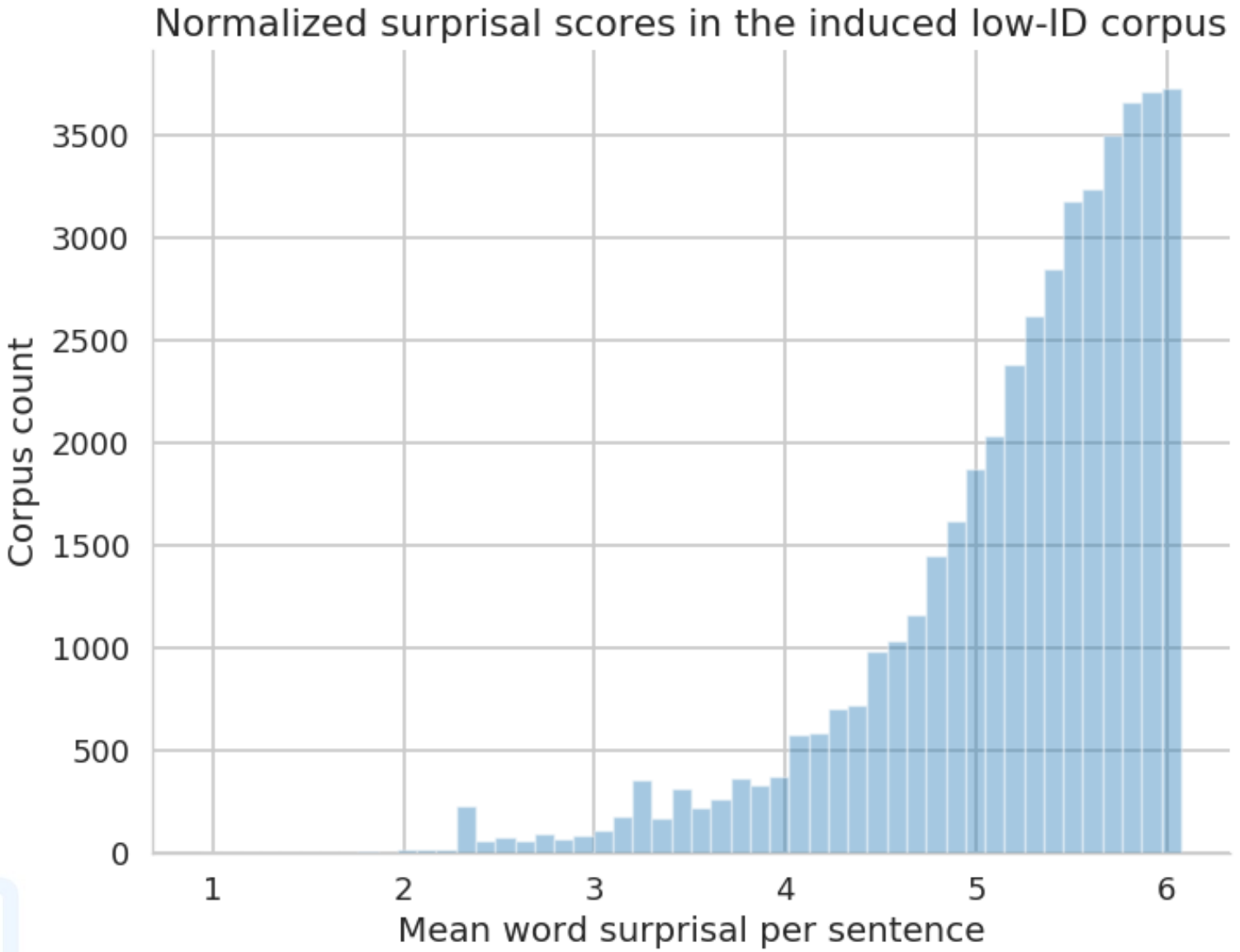
- Shorter **dependency lengths** in low-ID corpus
- Lower DLT **integration costs** in low-ID corpus

  Both with $p < 0.001$ and $d < 0.2$

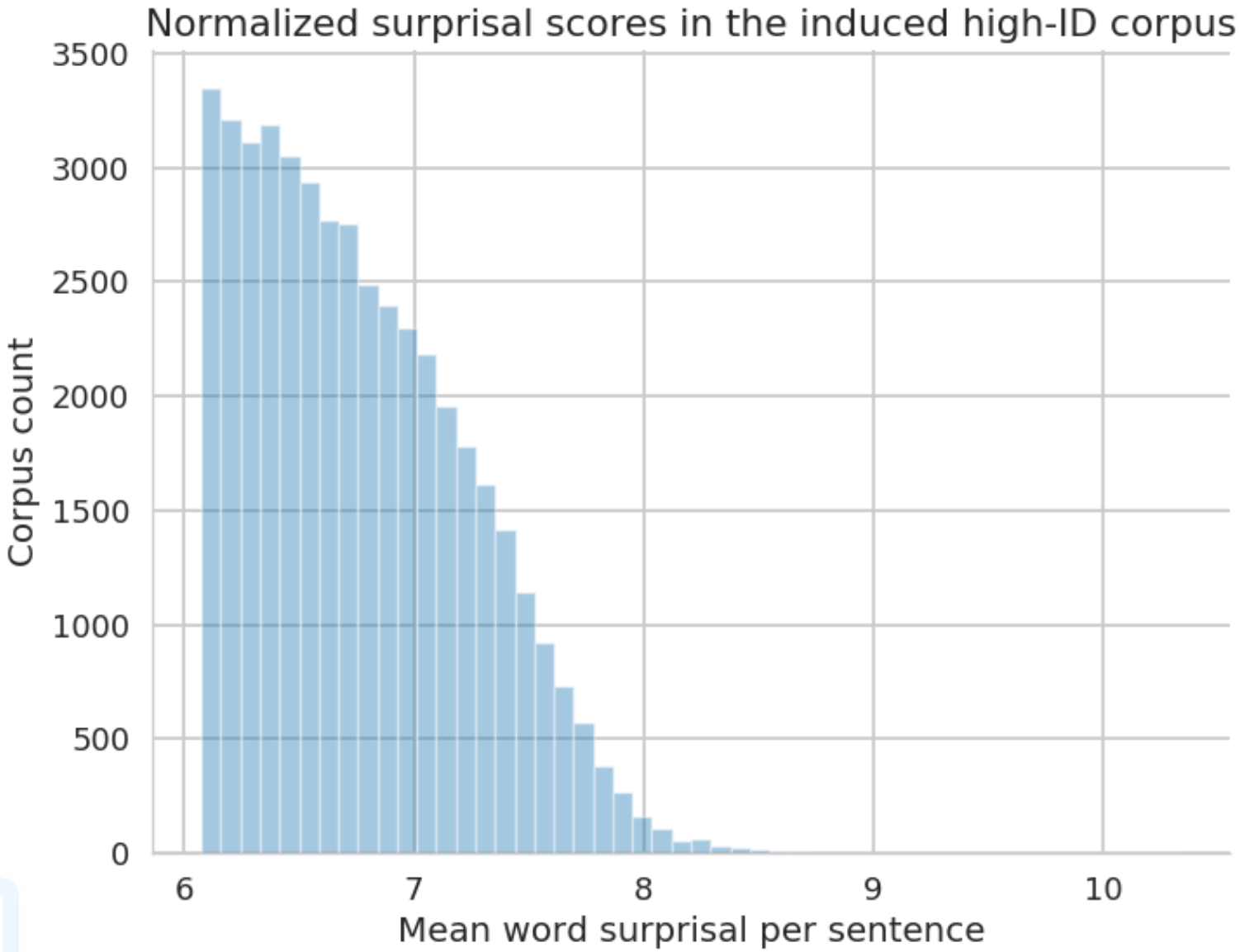
5. Data


Normalized surprisal scores in the 90k-Europarl corpus
Corpus count
7000
6000
5000
4000
3000
2000
1000
0
2
4
6
8
10
Mean word surprisal per sentence

5. Data


Normalized surprisal scores in the induced low-ID corpus
Corpus count
3500
3000
2500
2000
1500
1000
500
0
1
2
3
4
5
6
Mean word surprisal per sentence

5. Data


Normalized surprisal scores in the induced high-ID corpus
Corpus count
3500
3000
2500
2000
1500
1000
500
0
6
7
8
9
10
Mean word surprisal per sentence

6. Analysis

Main result: **No consistent ID reduction**

6. Analysis: IDGAN output samples

High-ID input	Low-ID output	Δ ID + HPs
that is the prime consideration	that is the key consideration	**1.2939** (G2|D1| A.2|R.8)
these structures were established on this date	what is adopted on this time	**1.1449** (G2|D1| A.8|R.2)
in our experience such requests have not always been attended to	in our experience such requests have not always been completed to	**0.2046** (G4|D1| MT)
under these circumstances it is clear that were mr brie a member of the bundestag he would enjoy immunity from prosecution which has been launched against him	under these circumstances it is clear that were mr perry a member of the interior he would enjoy decide from competences which has been passed against him	**-0.3995** (G4|D1| A.2|R.8)
however the total sum for the sixyear period has been pitched low too low at eur	however the total transition for the tse period has been abolished low too low at eur	**-0.6486** (G2|D1| A.8|R.2)

6. Analysis

- Training corpora not divergent / large enough
- Sentence encodings not task-specific
- Reconstruction inadequate as auxiliary task
- Adversarial objective ignores G's decoder
- Hyper-parameter space too large
- Sentence encodings inadequate as GAN input

→ Meaningful encodings may be limited to small hidden space regions (*Zhang et al., 2016*)

7. Current/ Future Work

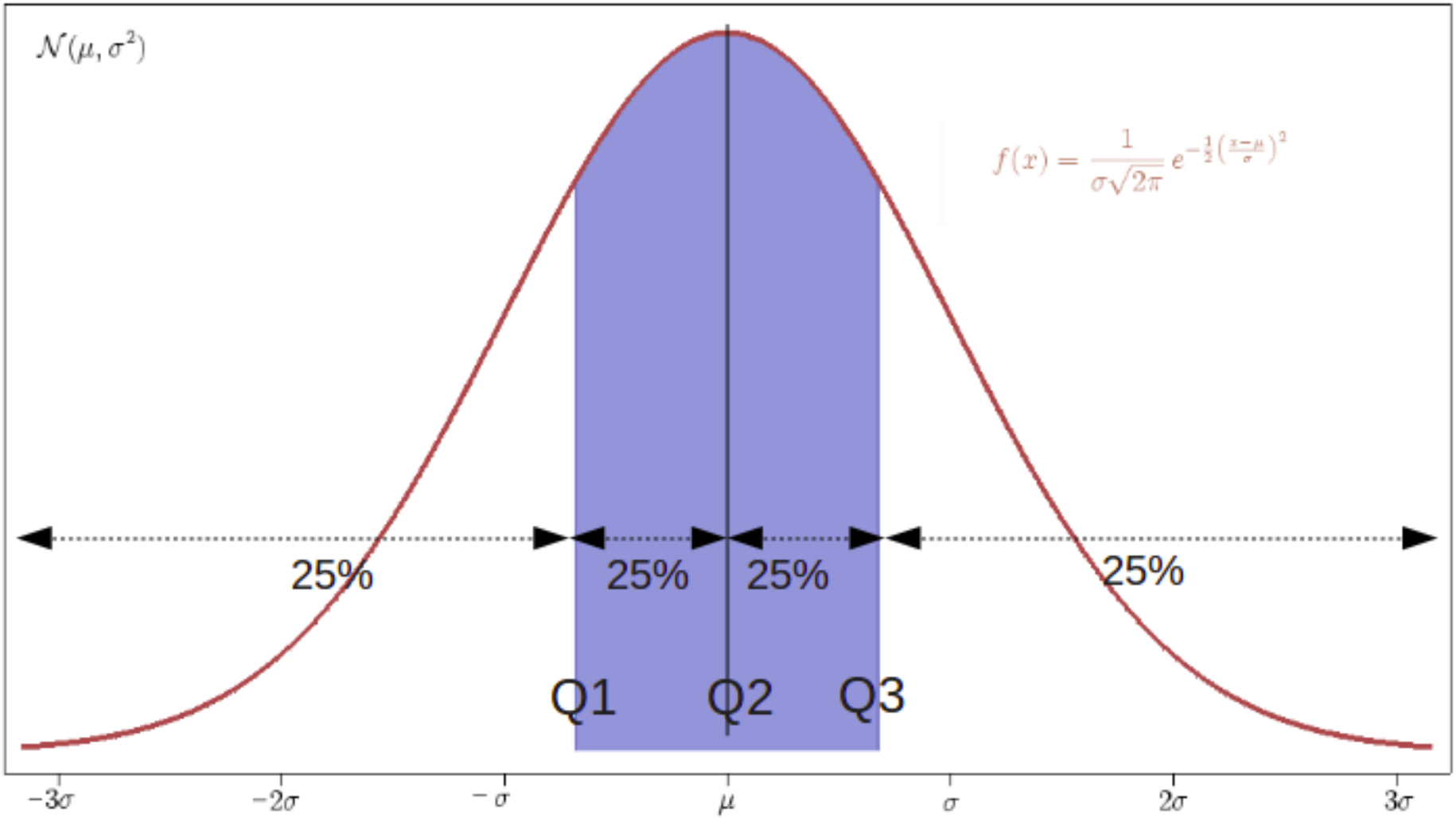
Improve **data construction** strategy

→ Increase **contrast** between ID-variant corpora

→ Sample sentences around 1- and 3-quantiles

→ Weighted sampling with emphasis on **outer regions**

7. Current/ Future Work



en.wikipedia.org/wiki/Quantile

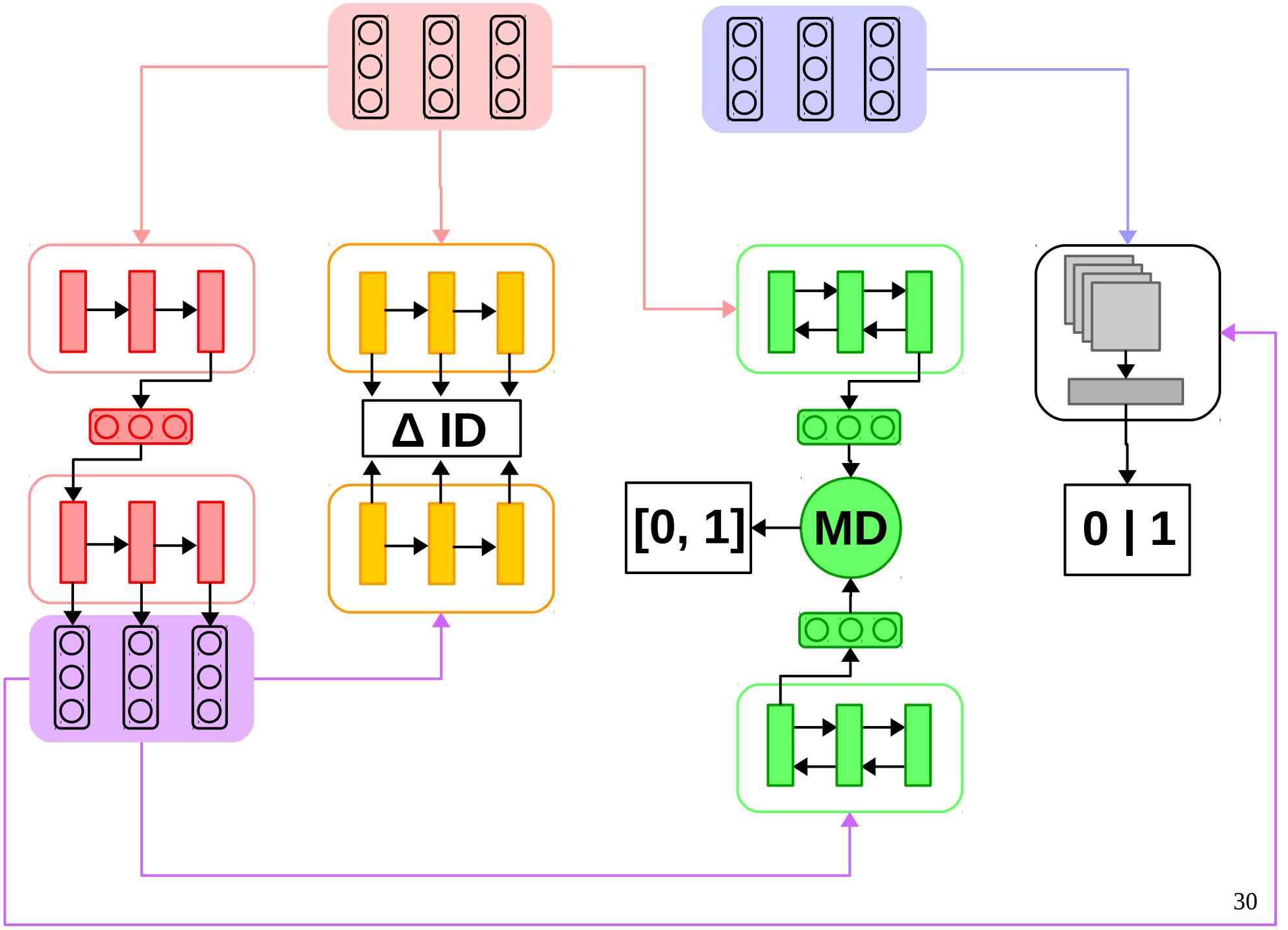
7. Current/ Future Work

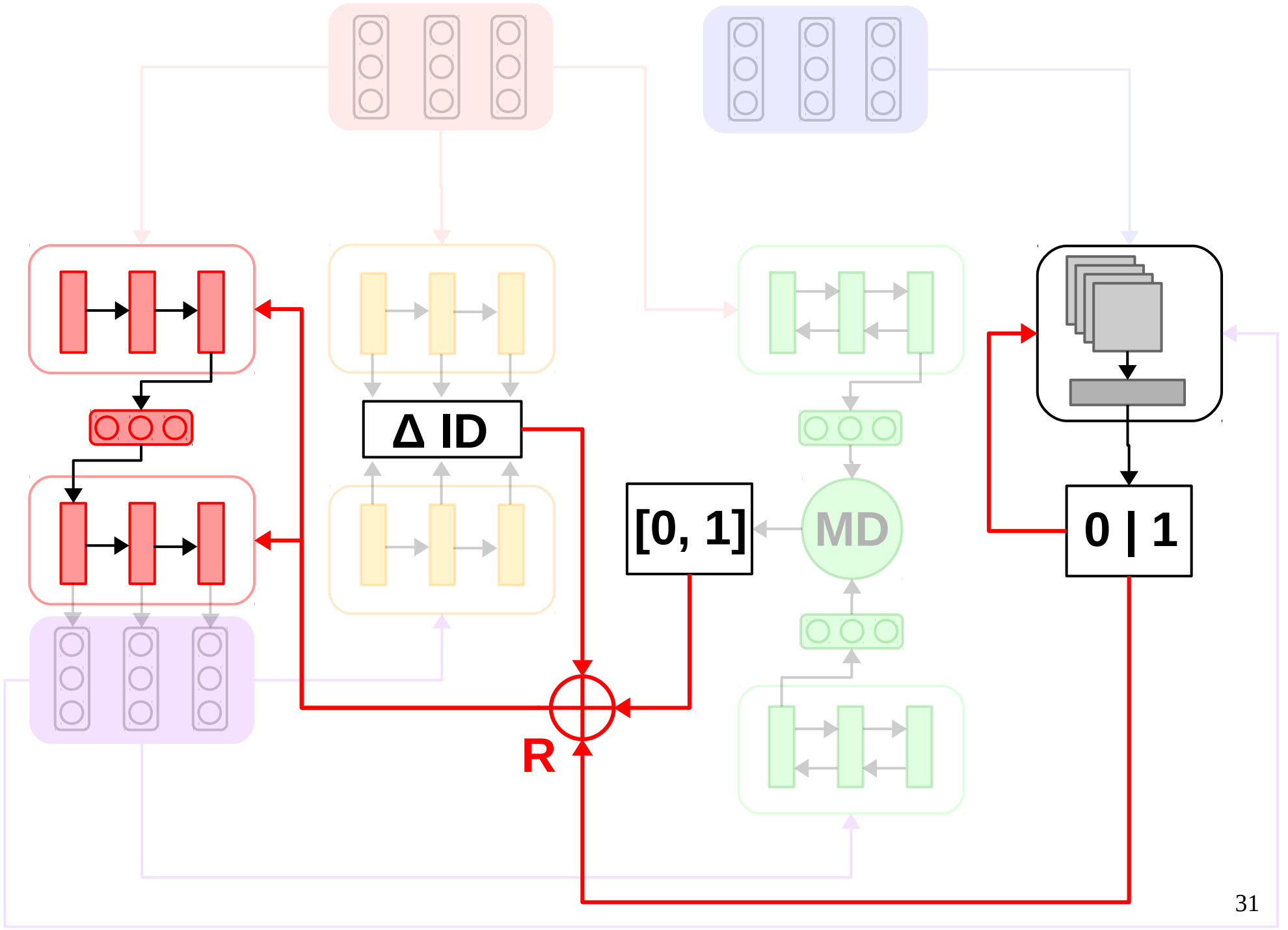
Add **auxiliary tasks** to improve GAN training

→ **Surprisal** should guide training explicitly

→ Content preservation at **sequence level**

→ Reinforcement learning for discrete operations (inefficient!)


Δ ID
[0, 1]
MD
0 | 1


Δ ID
[0, 1]
MD
0 | 1
R

References

- Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." *arXiv preprint arXiv:1611.07004* (2016).
- Zhang, Yizhe, Zhe Gan, and Lawrence Carin. "Generating text via adversarial training." *NIPS workshop on Adversarial Training.* 2016.
- Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." *arXiv preprint arXiv:1703.10593* (2017).

Thank you!

Any questions are welcome.