DualAxis AI at SemEval-2026 Task 3: Dimensional Aspect-Based Sentiment Analysis

Yahya Missaoui[1], **Solomon Kebede**[1], **Mounika Marreddy**[1], **Alexander Mehler**[1]
[1]Goethe University, Frankfurt am Main, Germany
missaoui@stud.uni-frankfurt.de, solo.kebede@stud.uni-frankfurt.de
mmarredd@em.uni-frankfurt.de, mehler@em.uni-frankfurt.de

Abstract

Dimensional Aspect-Based Sentiment Analysis models sentiment using continuous valence and arousal scores instead of discrete polarity labels, enabling fine-grained affect representation at the aspect level. SemEval 2026 Task 3 defines this setting through three subtasks covering aspect-level regression and structured extraction of aspect–opinion pairs with continuous scoring. We implement transformer-based baselines for all subtasks within a unified, reproducible framework. For aspect-level regression, we fine-tune pretrained encoders in an aspect-conditioned setup to predict valence and arousal. RoBERTa-large achieves the best development performance, with average RMSEs of 0.884 (restaurant) and 0.789 (laptop).

For the structured subtasks, we adopt an instruction-based sequence-to-sequence generation approach using Flan-T5. The model generates triplets or quadruplets in a canonical textual format, thereby jointly producing aspect terms, opinion terms, valence–arousal scores, and, for Subtask 3, aspect categories. Our best model attains continuous F1 scores of 0.742 and 0.648 for triplet extraction, and 0.604 and 0.385 for quadruplet extraction on the restaurant and laptop domains, respectively. Results show that continuous aspect-level regression is relatively stable under standard fine-tuning, whereas jointly extracting structured elements and predicting continuous affect remains considerably more challenging. Our systems provide reproducible baselines under the official evaluation protocol for future work on dimensional aspect-based sentiment analysis.

1 Introduction

Sentiment toward specific aspects of a product or service is often expressed with varying emotional intensity. Traditional aspect-based sentiment analysis typically reduces this variation to discrete labels such as positive or negative, overlooking differences in strength or activation. Modeling sentiment along continuous affective dimensions provides a more expressive representation of aspect-level opinions.

Shifting from classification to continuous prediction changes the learning problem. Models must estimate calibrated real-valued scores rather than select from fixed categories. The challenge increases in structured settings, where systems must first identify aspect and opinion spans and then assign appropriate affective values; extraction errors directly affect score quality.

SemEval 2026 Task 3 provides a unified benchmark for studying these challenges by evaluating aspect-level regression and structured extraction under the same dimensional framework. This setting enables systematic analysis of how modern pretrained encoders perform on regression and structured prediction when affective intensity is modeled explicitly.

Although transformer-based models have shown strong results in sentiment classification, their behavior under continuous supervision remains less explored, particularly when span identification and score prediction are combined. In this work, we present controlled and reproducible baseline systems for all subtasks. By maintaining consistent data processing and decoding procedures across encoder backbones, we isolate the impact of architectural choices on dimensional sentiment modeling. Our systems are intended to serve as reference baselines for future research.

Our main contributions are threefold. First, we evaluate encoder-based aspect-conditioned regression models for Subtask 1 across the restaurant and laptop domains. Second, we formulate Subtasks 2 and 3 as instruction-based Flan-T5 generation tasks that directly produce structured triplets and quadruplets. Third, we provide development-set results, training diagnostics, codes and an error-oriented analysis to support reproducibility and facilitate future comparisons. The codes

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 573–578

are publicly available at https://github.com/SolomonM-Kebede/ProjectNLP-DimABSA2026.

2 Related Work

Dimensional sentiment modeling. Research in affective psychology has long argued for modeling emotions along continuous dimensions rather than discrete categories. The circumplex model of affect (Russell, 1980, 2003) formalizes emotion primarily in terms of valence and arousal, a representation that has since been adopted in NLP. Lexical resources such as the NRC VAD Lexicon (Mohammad, 2018) and sentence-level datasets like EmoBank (Buechel and Hahn, 2017) operationalize this framework by providing real-valued affective annotations. These resources have driven regression-based sentiment models that capture fine-grained affective intensity, moving beyond coarse polarity classification.

Aspect-based sentiment analysis and structured prediction. Aspect-based sentiment analysis (ABSA) focuses on identifying sentiments expressed toward specific targets, evolving from pipeline-based approaches to joint structured prediction. Survey work (Zhang et al., 2022) documents this progression, highlighting formulations such as Aspect Sentiment Triplet Extraction (ASTE) (Peng et al., 2020) and Aspect Sentiment Quad Prediction (ASQP) (Zhang et al., 2021). These tasks require jointly modeling multiple interdependent components—aspect terms, opinion terms, sentiment labels, and optionally aspect categories—substantially increasing modeling complexity compared to sentence-level sentiment classification.

Query-based extraction for structured ABSA. To address the challenges of joint prediction, many recent approaches cast structured ABSA as a machine reading comprehension (MRC) problem. In this paradigm, models extract aspect and opinion spans by answering task-specific natural language queries, followed by sentiment or category classification (Chen et al., 2021; Gao et al., 2021). Multi-turn and bidirectional querying strategies help mitigate error propagation and improve coverage in structured settings. Our DimASTE and DimASQP baselines adopt this query-based framework and extend it by predicting continuous valence and arousal scores alongside structured sentiment elements.

Pretrained transformers and generative baselines. Pretrained transformer models serve as the backbone for most modern ABSA systems due to their strong contextual representations. Encoder-based architectures such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) remain competitive under carefully controlled training and decoding conditions. In parallel, sequence-to-sequence models such as T5 (Raffel et al., 2020) and instruction-tuned variants like Flan (Chung et al., 2022) enable generative formulations that produce structured outputs as text. Comparing encoder-based and generative approaches under a unified evaluation protocol provides insights into their relative strengths for dimensional and structured sentiment modeling.

Dimensional aspect-based sentiment shared tasks. Recent shared tasks have formalized the integration of dimensional affect modeling with ABSA. SemEval 2026 Task 3 introduces Dimensional ABSA (DimABSA), requiring prediction of continuous valence and arousal scores either for given aspects (DimASR) or jointly with structured sentiment extraction (DimASTE/DimASQP) (Yu and Lee, 2025). Related efforts, such as the SIGHAN 2024 shared task on Chinese Dimensional ABSA (Lee et al., 2024), similarly emphasize fine-grained affect intensity estimation within aspect-aware sentiment frameworks, underscoring the growing interest in continuous sentiment representations.

3 Task and Dataset Description

3.1 Task definition

SemEval 2026 Task 3 defines DimABSA as the problem of predicting continuous valence and arousal scores for aspect level sentiment. The task is divided into three subtasks, each addressing a different prediction setting. **Subtask 1 (DimASR).** Given a text and a predefined list of aspects, the system must predict a valence–arousal (VA) score for each aspect. This subtask focuses on aspect level regression. **Subtask 2 (DimASTE).** Given a text without predefined aspects, the system must extract all aspect–opinion–score triplets (A, O, VA). Here, A denotes the aspect term, O denotes the opinion term, and VA represents the corresponding valence–arousal score. **Subtask 3 (DimASQP).** Given a text, the system must extract all quadruplets (A, O, C, VA), where C denotes the aspect category in addition to the aspect term, opinion

term, and valence–arousal score. Together, these subtasks evaluate both regression and structured extraction under a unified dimensional sentiment framework. Abbreviations used throughout the paper are summarized in Table 5, Appendix A.1.

3.2 Data format

All datasets are released in JSONL format, where each line corresponds to a single instance containing a unique `ID` and a `Text` field. For DimASR, the input additionally includes an `Aspect` list. The system must output an `Aspect_VA` field, which contains one predicted VA score for each given aspect. For DimASTE, the output consists of a `Triplet` list. Each triplet includes the fields `Aspect`, `Opinion`, and `VA`. For DimASQP, the output is a `Quadruplet` list that includes `Aspect`, `Opinion`, `Category`, and `VA`. In the released data, implicit or unmentioned spans are represented using the literal string `NULL`, for example `Aspect = NULL`. The VA label is represented as a string in the format `V#A`, where the first value corresponds to valence and the second to arousal. Each dimension takes a value between 1.00 and 9.00 and is rounded to two decimal places. For example, a valid label may appear as `6.75#6.38`. For model selection during development, we use the provided training data and reserve 10% as an internal validation split where required, resulting in a 90/10 training—validation split. Final results on the development set are reported using the official development files and evaluation scripts.

3.3 Evaluation

DimASR is evaluated using Root Mean Squared Error (RMSE), computed between the predicted and gold valence and arousal scores. This metric measures the average deviation of the predicted continuous values from the reference annotations. DimASTE and DimASQP are evaluated using the official continuous F1 (cF1) metric. This metric rewards correct extraction of structured sentiment elements while also incorporating the distance between predicted and gold VA scores. All experiments are conducted using the official evaluation scripts and the provided training and development splits. The test labels are not publicly available.

4 Method

In this section, we describe the transformer-based models developed for SemEval-2026 Task 3. Rather than relying on a single architecture for all subtasks, we employ a hybrid modeling approach tailored to the requirements of each subtask. For Subtask 1, we formulate the problem as aspect-conditioned valence–arousal (VA) regression, where the model predicts continuous VA scores for each given aspect. We use discriminative models based on pretrained BERT and RoBERTa encoders, which are well suited for capturing global sentence-level semantics and mapping them to continuous affective dimensions.

For Subtasks 2 and 3, corresponding to DimASTE and DimASQP, we adopt a generative query-based extraction framework based on FLAN-T5 (Chung et al., 2022). Instead of using a discriminative pipeline with separate span-detection, regression, and classification heads, we treat the task as instruction-based generation. The encoder receives a task-specific query together with the review text, while the decoder acts as a unified multi-task prediction component. It generates a structured JSON sequence representing aspect spans, opinion spans, VA values, and, for Subtask 3, aspect categories. This design is inspired by the generative aspect-based sentiment analysis framework of Li et al. and allows the model to capture dependencies between aspects, opinions, categories, and sentiment dimensions within a single generation process.

4.1 DimASR: aspect-conditioned VA regression

For Subtask 1, we construct one training instance for each annotated aspect. From every JSONL entry in the training data, we extract the review text and iterate over its labeled units. For each aspect, we pair the aspect string with its corresponding gold VA value. Each pair is transformed into a single input sequence by concatenating an aspect prompt with the review text:

```
aspect: <a> text: <x>
```

Here, a denotes the aspect and x denotes the review text. The resulting sequence is tokenized using the RoBERTa tokenizer. Our regression model is built on a pretrained RoBERTa encoder. We apply masked mean pooling over the final layer token representations to obtain a fixed length sentence representation. This representation is passed to a multi layer feed forward prediction head with GELU activations and LayerNorm, which outputs two continuous values, $\hat{v}$ and $\hat{a}$, corresponding to predicted valence and arousal. To stabilize training, the target scores are standardized using the mean

and standard deviation computed from the training split. During inference, predictions are transformed back to the original scale for evaluation and submission. The model is trained using Huber loss applied to the two-dimensional regression target. At inference time, we generate one input sequence for each text and aspect pair provided in the test data. For every aspect, the model outputs a predicted VA score pair in the required `Aspect#VA` field.

4.2 DimASTE/DimASQP: sequence-to-sequence structured generation

For Subtasks 2–3, we model structured extraction with continuous scoring as a text-to-text generation problem using an instruction-tuned encoder–decoder backbone (Flan-T5). Given a review text, the model is prompted to generate all sentiment structures in a canonical textual format. For DimASTE, the target consists of a list of triplets (A, O, VA); for DimASQP, the target is a list of quadruplets (A, O, C, VA), where C is the aspect category and VA is emitted as `V#A`. We use a single model per subtask and domain and train with standard sequence-to-sequence cross-entropy loss (negative log-likelihood) over the target tokens.

During preprocessing, gold annotations from the official JSONL files are converted into the corresponding target string representation. At inference time, we decode with beam search and parse the generated text back into structured records. We apply lightweight normalization (e.g., trimming whitespace and normalizing separators) and deduplicate identical structures before writing predictions in the official `Triplet` or `Quadruplet` JSONL output schema. Predicted VA values are taken directly from the generated `V#A` strings and are clipped to the valid range when necessary.

4.3 Experimental Setup

All experiments are conducted using the official train/dev splits and the evaluation scripts provided by the task organizers. For Subtask 1, the training data is expanded into individual $(\text{text}, \text{aspect})$ instances, where each aspect in a review forms a separate example. We use a 90/10 split of the training data for internal model selection where required. For Subtask 1, we fine-tune RoBERTa and BERT encoders using AdamW with learning rate scheduling, gradient clipping, and gradient accumulation, following the configuration specified in the task script. The best checkpoint is selected based on validation performance. The hyperparameter for Subtask 1 are provided in Table 5. For Subtasks 2 and 3, we fine-tune Flan-T5 models for structured generation. Model checkpoints are selected based on development set performance. The corresponding hyperparameter for Subtask 2 and 3 are also provided in Table 5. All experiments were conducted on Google Colab and Kaggle using NVIDIA Tesla T4 GPUs with approximately 16GB of VRAM.

5 Results

We evaluate on the official test set for Subtasks 2–3, and on the development set for Subtask 1.[1] DimASR is evaluated with RMSE and CCC; DimASTE and DimASQP use the official continuous-F1 (cF1) metric.

5.1 DimASR: Aspect-Level VA Regression

Tables 1 and 2 show development results across backbone models. RoBERTa-large achieves the best RMSE_{avg} in both domains (0.884 restaurant; 0.789 laptop), confirming that larger encoder capacity benefits continuous VA regression.

Backbone	RMSE_v	RMSE_a	RMSE_{avg}	CCC_{avg}	Ep.
RoBERTa-large	1.010	0.757	**0.884**	**0.800**	5
BERT-base-uncased	1.071	0.787	0.929	0.774	3
BERT-base-cased	1.059	0.825	0.942	0.745	2
RoBERTa-base	1.086	0.865	0.975	0.701	3

Table 1: DimASR dev results, restaurant domain.

Backbone	RMSE_v	RMSE_a	RMSE_{avg}	CCC_{avg}	Ep.
RoBERTa-large	0.829	0.748	**0.789**	0.778	2
RoBERTa-base	0.847	0.786	0.816	0.775	3
BERT-base-uncased	0.877	0.760	0.818	0.764	2
BERT-base-cased	0.907	0.746	0.826	**0.785**	2

Table 2: DimASR dev results, laptop domain.

5.2 DimASTE and DimASQP

Table 3 reports test-set cF1 for both subtasks alongside the organizer baseline. Our system outperforms the baseline on all four subtask–domain combinations. For DimASTE, gains are +0.0582 (restaurant) and +0.0374 (laptop). For DimASQP, we improve by +0.0378 on restaurant and by a large margin of +0.3427 on laptop (0.5910 vs. 0.2483),

[1] Test-set labels for Subtask 1 (DimASR) were not distributed to participants. We therefore report development-set performance and compare against the organizer baseline on the same split. All Subtask 2 and 3 numbers are on the official test set.

Subtask	Domain / System	cF1	P	R
DimASTE	Rest. baseline	0.5442	—	—
	Rest. ours	**0.6024**	0.6319	0.5926
	Laptop baseline	0.4664	—	—
	Laptop ours	**0.5038**	0.5382	0.4942
DimASQP	Rest. baseline	0.5048	—	—
	Rest. ours	**0.5426**	0.5781	0.5292
	Laptop baseline	0.2483	—	—
	Laptop ours	**0.5910**	0.6011	0.5731

Table 3: DimASTE and DimASQP test-set results. Baseline = organizer system; P = Precision; R = Recall.

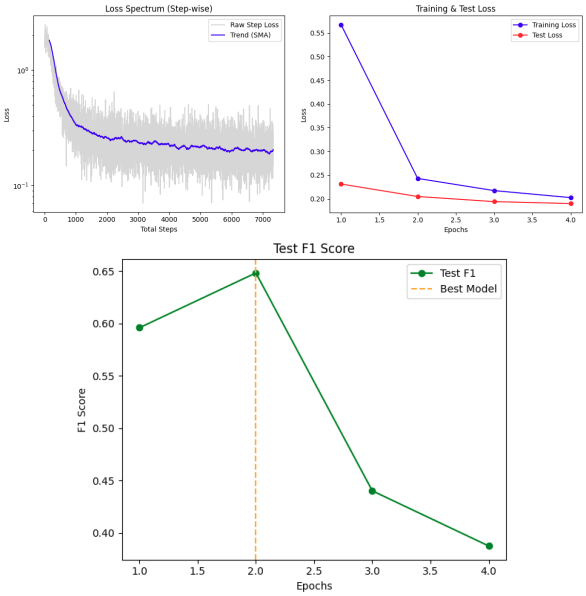

Figure 1: DimASTE diagnostics (`flan-t5-large`, laptop): training loss (top left), dev cF1 by epoch (top right), train vs. dev loss (bottom).

suggesting that our sequence-to-sequence approach handles the sparser, more technical laptop vocabulary more robustly than the organizer baseline. Notably, laptop DimASQP (0.5910) exceeds laptop DimASTE (0.5038), which we attribute to the structured category vocabulary providing an additional grounding signal absent in free-form triplet extraction.

5.3 Analysis

Results confirm a consistent difficulty gradient from aspect-level regression to structured extraction. For DimASR, encoder size is the primary driver of performance. For DimASTE and DimASQP, all baseline comparisons favour our system. Training curves (Figures 1–2) show rapid convergence within two to three epochs across all settings.

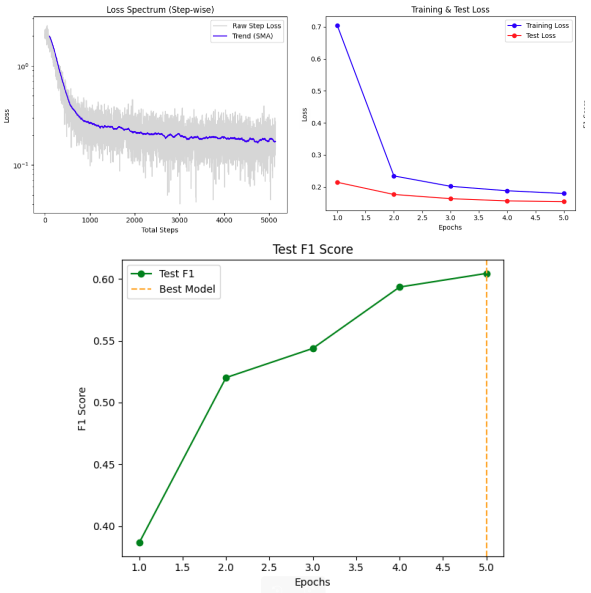

Figure 2: DimASQP diagnostics (`flan-t5-large`, restaurant): step-wise loss (top left), train vs. dev loss (top right), dev cF1 by epoch (bottom).

6 Discussion

Seq2seq generalization. Flan-T5-large's instruction-following pre-training enables robust generalization to the structured output format, particularly for technically sparse laptop vocabulary. The large DimASQP laptop gain (+0.3427) suggests the organizer baseline struggles with domain shift that our generation-based approach handles more gracefully.

Category as grounding signal. The reversal of the DimASTE–DimASQP ordering on the laptop domain (0.5038 vs. 0.5910) indicates that the closed-set category vocabulary (e.g., `BATTERY#OPERATION_PERFORMANCE`) acts as a supervisory anchor, reducing the model's reliance on ambiguous free-form opinion spans.

Domain gap. Restaurant reviews yield uniformly higher cF1 than laptop reviews for DimASTE, consistent with the richer and more stereotyped sentiment vocabulary of restaurant text. The gap narrows or reverses for DimASQP, where structured categories partially compensate for the sparser laptop sentiment expressions.

Limitations and future work. The three subtasks are currently solved independently; a joint model sharing representations across DimASTE and DimASQP could reduce the category-prediction bottleneck. For DimASR, aspect-aware pooling beyond CLS-token regression may further reduce RMSE. Domain-adaptive pre-training re-

mains a promising direction to close the restaurant–laptop gap.

References

Sven Buechel and Udo Hahn. 2017. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of EACL*.

Shaoxiong Chen and 1 others. 2021. Bidirectional machine reading comprehension for aspect sentiment triplet extraction. In *Proceedings of AAAI*.

Hyung Won Chung and 1 others. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Li Gao and 1 others. 2021. Question-driven span labeling model for aspect–opinion pair extraction. In *Proceedings of AAAI*.

Lung-Hao Lee, Liang-Chih Yu, Natalia Loukashevich, Ilseyar Alimova, Alexander Panchenko, Tzu-Mi Lin, Zhe-Yu Xu, Jian-Yu Zhou, Guangmin Zheng, Jin Wang, Sharanya Awasthi, Jonas Becker, Jan Philip Wahle, Terry Ruas, Shamsuddeen Hassan Muhammad, and Saif M. Mohammed. 2026. Dimabsa: Building multilingual and multidomain datasets for dimensional aspect-based sentiment analysis. *arXiv preprint arXiv:2601.23022*.

Lung-Hao Lee, Liang-Chih Yu, Suge Wang, and Jian Liao. 2024. Overview of the SIGHAN 2024 shared task for Chinese dimensional aspect-based sentiment analysis. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 165–174, Bangkok, Thailand. Association for Computational Linguistics.

Yinhan Liu and 1 others. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Saif M. Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of ACL*.

Haiyun Peng and 1 others. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proceedings of AAAI*.

Colin Raffel and 1 others. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*. Preprint: arXiv:1910.10683 (doi:10.48550/arXiv.1910.10683).

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*.

James A. Russell. 2003. Core affect and the psychological construction of emotion. *Psychological Review*.

Lu Xu and 1 others. 2020. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of EMNLP*.

Liang-Chih Yu and Lung-Hao Lee. 2025. Call for participation – semeval-2026 task 3: Dimensional aspect-based sentiment analysis on customer reviews and stance datasets. ACL Member Portal (Association for Computational Linguistics). Posted Nov 7, 2025. Accessed 2026-02-16.

Wenxuan Zhang and 1 others. 2021. Aspect sentiment quad prediction as paraphrase generation. *arXiv preprint arXiv:2110.00796*.

Wenxuan Zhang and 1 others. 2022. A survey on aspect-based sentiment analysis. *arXiv preprint arXiv:2203.01054*.

A Appendix

A.1 Abbreviations

Abbrev.	Meaning
ABSA	Aspect-Based Sentiment Analysis
DimABSA	Dimensional Aspect-Based Sentiment Analysis
VA	Valence–Arousal score pair
Subtask 1	Aspect-level VA regression
Subtask 2	(A, O, VA) triplet extraction
Subtask 3	(A, O, C, VA) quadruplet extraction
JSONL	JSON Lines (one JSON object per line)
RMSE	Root Mean Squared Error
cF1	Continuous-F1 (official extraction metric with VA distance)

Table 4: Abbreviations used in this paper.

A.2 Hyperparameters

For Subtask 1 we utilized pretrained variants of RoBERTa and BERT, while for Subtasks 2 and 3 we used FLAN-T5 base and FLAN-T5 large. Table 5 details the specific configurations used for each.

Parameter	Subtask 1	Subtasks 2 & 3
Backbone Models	RoBERTa/BERT	FLAN-T5 (B/L)
Max Sequence Length	256 tokens	256 tokens
Batch Size	3	1
Training Epochs	4	8
Learning Rate	2×10^{-5}	4×10^{-5}
Weight Decay	0.01	—
Dropout / Warm-up	0.3 (Drop)	0.1 (Warm-up)
Grad. Accumulation	16	16
Max Gradient Norm	1.0	1.0
Early Stopping	3 epochs	2 epochs
Mixed Precision	Enabled	—
Grad. Checkpointing	—	Enabled
Target Normalization	Enabled	—

Table 5: Consolidated hyperparameters for all subtasks.