Thesis Report

Detection of Focal Cortical Dysplasia Type II Using Text Descriptions

German Mikhelson

October 2025

1 Basic Experiments

Before tuning various hyperparameters such as the number of text connections in the GuideDecoder, the number of unfrozen layers in the LLM and the number of connections to the GNN block (pruned in the baseline experiments), it's essential to identify the best-performing base model first to avoid unnecessary experiments and save time.

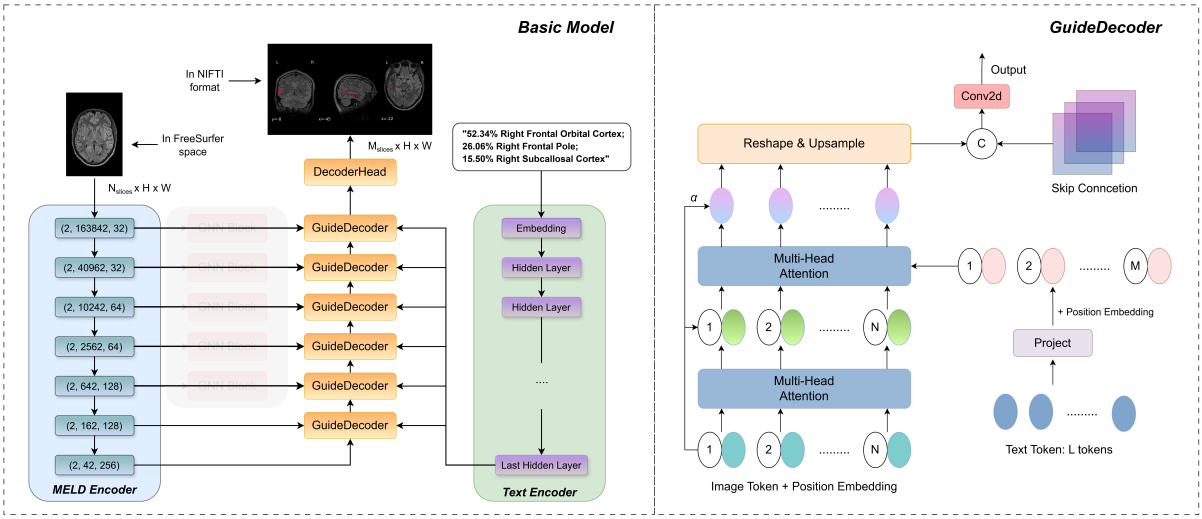


Figure 1: Overview of the basic model (left) and the `GuideDecoder` module (right) used in our experiments.

In all variants, we freeze both encoders: the visual encoder (the **MELD** backbone) and the text encoder. Although **RadBERT** was used as the initial language model during the early stages of this work—simply because it was the first domain-specific model that integrated well into the pipeline—the choice of text encoder was later examined more systematically. After observing that textual signals substantially influence lesion detection quality, we conducted an additional study comparing several biomedical and radiology-oriented LLMs (e.g., BlueBERT, PubMedBERT). This analysis (reported in a separate

chapter, including cosine-similarity evaluation of anatomical terminology) motivated the inclusion of multiple text-encoder variants in the experimental section.

Across all variants, the geometry-based upsampling path (`HexUnpool` + `SpiralConv`) and the final segmentation head are always trained, while the encoders remain frozen (Fig. 1). The experimental variants differ only in (i) whether the `GuideDecoder` is inserted before the upsampling path and (ii) whether and which textual features are incorporated. The pretrained `Exp1` model (described below) is used to initialize the decoder for all variants except the MELD-only baseline, which is trained from scratch.

We consider the following configurations:

- **MELD.** Serves as the baseline model.
- **Exp1 (Unpool + Spiral, no text).** The `GuideDecoder` is omitted; visual features from the MELD encoder are passed directly into the geometry-based upsampling path (`HexUnpool` + `SpiralConv`) and then to the segmentation head. No textual input is used in this setting, so the model relies purely on image-derived features (see Fig. 2).

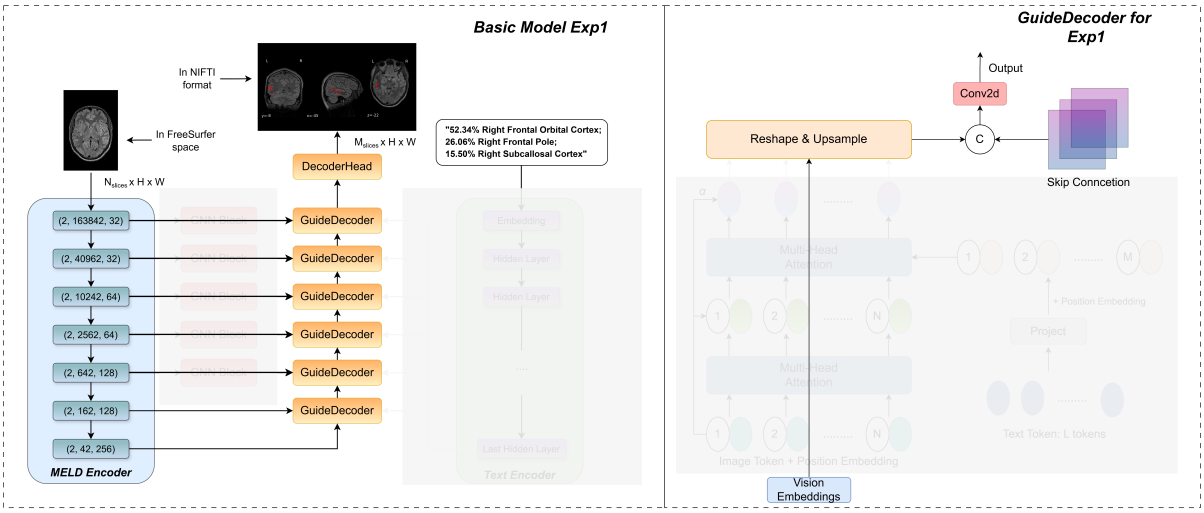


Figure 2: Architecture of the Exp1 variant: the `GuideDecoder` and text branch are removed, and MELD features are fed directly into the upsampling path and segmentation head.

- **Exp2 (GuideDecoder: self-attention only).** A stack of `GuideDecoder` blocks is inserted before the upsampling stage. The U-Net–style MELD decoder has depth 7, i.e., 6 upsampling stages, so we use 6 `GuideDecoder` blocks placed at levels D6–D1 (from the coarsest to the finest). The text branch is disabled in this variant, hence only self-attention is applied inside each `GuideDecoder`. The upsampling path and segmentation head are otherwise identical to Exp1, so the main difference is the additional self-attention–based refinement of visual features before upsampling (see Fig. 3).

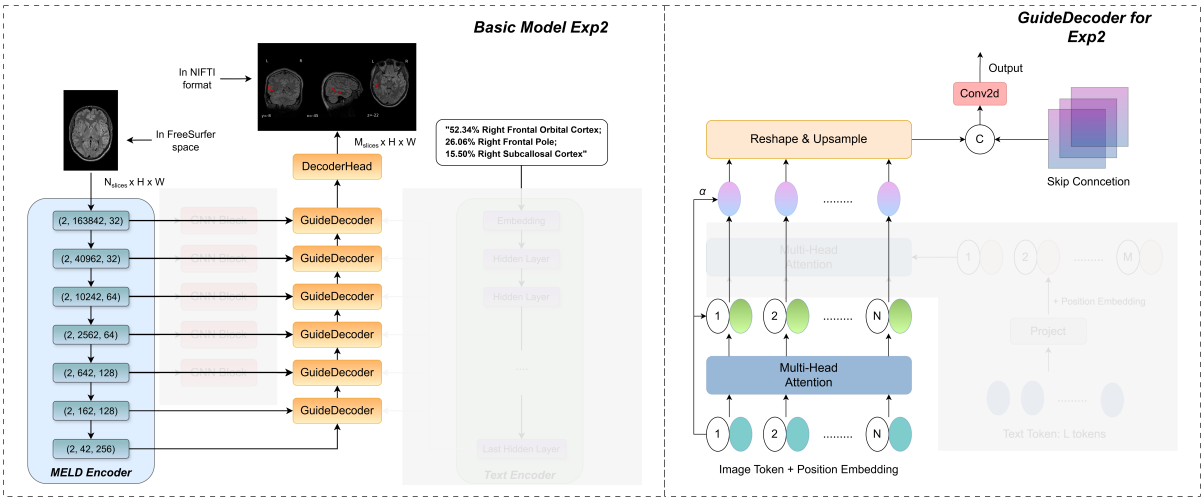


Figure 3: Architecture of the Exp2 variant: six `GuideDecoder` blocks with self-attention only are inserted before each upsampling stage, while the text branch remains disabled.

- **Exp3_mixed (GuideDecoder + Text).** Same as **Exp3**, but Atlas descriptions were randomly sampled from one of {*hemisphere only*, *lobe_regions only*, *hemisphere + lobe_regions*, *full text*, *no text*}.

  For each subject, we first generated a full Atlas-based description and then derived all partial textual variants. During training, a single variant was randomly selected at every data loading step, so that the same subject could appear with different textual inputs across epochs, while the visual augmentations remained fixed.

Main cohort Table 1 presents the median performance on the main cohort (i.e., data from the same distribution as used during training). In the RadBERT group, the prompt combining hemisphere and lobe-region terms detected the most lesions, though with only moderate Dice, $\text{PPV}_{\text{pixels}}$, and IoU. Across all models, the PubMedBERT variant achieved the highest number of detected lesions, likely due to better domain-context understanding (see Table 13 in the supplement). This is a clear example showing that choosing an appropriate language model can materially improve results, underscoring the substantial impact of the language encoder.

In real-world settings, detailed subregion information may be unavailable. We therefore evaluated prompts with only general *lobe names* (e.g., *Frontal lobe*, *Temporal lobe*, *Insular lobe*). For the RadBERT-based models, performance was very close to the *hemi+lobe_regions* setup: Dice decreased by 2.5%, PPV_pixels by 0.4%, IoU by 1.6%, and only **8** fewer lesions were detected. Likewise, using only *hemisphere* descriptions yielded results close to *hemi+lobe*, indicating that even a minimal textual prompt can help identify additional lesion regions.

Also, when using the full textual description (e.g., "52.34% Right Frontal Orbital Cor-

tex; 26.06% Right Frontal Pole; 15.50% Right Subcallosal Cortex"), the model performed worse than *hemi+lobe_regions*. One possible reason is the presence of redundant information, such as percentage overlaps of regions, which may not contribute to the prediction task. Since RadBERT was trained on radiology reports, it may not benefit from such structured numeric content. For comparison, we also tested PubmedBERT, which was trained on a larger corpus of brain scan-related reports. Although its quality metrics were slightly lower than those of *hemi+lobe* with RadBERT, it still detected 13 more lesion cluster.

We also report specificity, defined as the proportion of healthy patients with no predicted clusters. Across all Exp3 models, specificity was $\tilde{3}6$–40% higher than that of MELD. This indicates that the text-conditioned models distinguish healthy from non-healthy cases substantially better.

Throughout this section, we did not discuss $\text{PPV}_{\text{clusters}}$, the primary metric in the original MELD study. As Table 1 shows, MELD attains the highest mean $\text{PPV}_{\text{clusters}}$. Its margin over the second- and third-best models is only $\sim$ 2–3 %, whereas the gap to the remaining models is substantially larger. Therefore, $\text{PPV}_{\text{clusters}}$ is driven by outliers, it provides poor discrimination; switching to the median does not improve this. We therefore regard $\text{PPV}_{\text{clusters}}$ as not useful for model evaluation in this study.

The anatomical names of lobes and their respective regions are taken from *Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain*.

We also describe how we compute confidence intervals—following the procedure used by the MELD authors. For each metric, we report the sample median together with a 95% confidence interval estimated via a non-parametric bootstrap. Given scores $x_1, \ldots, x_N$, we draw $B = 10{,}000$ bootstrap resamples of size N with replacement, compute the median for each resample to obtain $\{\tilde{x}^{(b)}\}_{b=1}^{B}$, and take the 2.5th and 97.5th percentiles of this bootstrap distribution as the lower and upper bounds of the 95% CI. This percentile bootstrap makes no parametric assumptions about the underlying distribution; missing values (NaN) are removed prior to resampling. When $N = 1$, the CI collapses to the observed value. A fixed random seed is used for reproducibility.

Independent cohort Following the MELD paper, we used the dataset obtained from Bonn as an independent test cohort. This evaluation allows assessing the generalization capability of the models to unseen data from a different site and acquisition setting.

First, we note that *Exp2* achieves the weakest performance among all models. Moreover, its confidence interval for the Dice score extends down to zero, indicating highly uncertain predictions. This suggests that self-attention alone is insufficient to ensure

Table 1: Median performance on the main cohort (with 95% confidence intervals).

Model	Dice	PPV pixels	(mean) PPV clusters	(median) PPV clusters	IoU	Specificity	Sensitivity
MELD	0.230 (0.107–0.320)	0.166 (0.086–0.220)	0.515	1.000 (1.000-1.000)	0.130 (0.057–0.190)	112 / 193	170 / 259
Exp1	0.238 (0.125–0.313)	0.198 (0.130–0.318)	0.361	1.000 (0.500-1.000)	0.135 (0.066–0.186)	60 / 193	179 / 259
Exp2	0.100 (0.012–0.203)	0.287 (0.027–0.410)	0.462	1.000 (0.500-1.000)	0.053 (0.006–0.113)	113 / 193	159 / 259
Exp3: hemi	0.231 (0.148–0.285)	0.188 (0.131–0.298)	0.483	0.667 (0.500-1.000)	0.131 (0.080–0.166)	193 / 193	184 / 259
Exp3: lobe_regions	0.254 (0.159–0.332)	0.231 (0.168–0.320)	0.404	0.667 (0.500-1.000)	0.145 (0.086–0.199)	193 / 193	188 / 259
Exp3: hemi + lobe_regions	0.264 (0.180–0.342)	0.238 (0.160–0.327)	0.333	0.500 (0.500-1.000)	0.152 (0.099–0.206)	192 / 193	193 / 259
Exp3: hemi + lobe	0.239 (0.113–0.302)	0.242 (0.166–0.353)	0.456	0.667 (0.500-1.000)	0.136 (0.060–0.178)	192 / 193	185 / 259
Exp3: hemi + lobe + BlueBERT	0.230 (0.125–0.305)	0.231 (0.146–0.330)	0.443	0.667 (0.500-1.000)	0.130 (0.067–0.180)	191 / 193	186 / 259
Exp3: hemi + lobe + PubmedBERT	0.233 (0.139–0.330)	0.242 (0.153–0.342)	0.334	0.500 (0.500-0.667)	0.132 (0.074–0.198)	193 / 193	198 / 259
Exp3: full_desc	0.246 (0.145–0.330)	0.259 (0.178–0.360)	0.478	0.667 (0.500-1.000)	0.141 (0.078–0.197)	182 / 193	188 / 259

Note. Colors denote rank within each column: best, second, third.
Exp labels. **Exp3: hemi** – hemisphere conditioning; **lobe_regions** – lobe- and region-level prompts; **hemi+lobe** – hemisphere + coarse lobe name; **hemi+lobe+BlueBERT** – same as previous but with BlueBERT instead of RadBERT; **full_desc** – full free-text description; **mixed** – mixture of prompts. **Exp1** and **Exp2** are ablation baselines.

stable generalization across both cohorts.

The models that detected the largest number of lesion clusters were those trained with *lobe_regions* prompts across RadBERT based models and with *hemi+lobe+PubmedBERT* across all models. When PubmedBERT was used instead of RadBERT in the *hemi+lobe* configuration, the model achieved consistently better performance across multiple metrics (Dice: +2.0%, $\text{PPV}_{\text{pixels}}$: +4.6%, IoU: +1.4%) and detected **7 more lesions**. This improvement is likely due to PubmedBERT's pretraining on a substantially larger corpus of brain-related clinical reports, which improves its handling of domain-specific terminology.

Overall, *hemi+lobe+PubmedBERT* provides the best compromise: the prompt is short but specific, and the model achieves the strongest sensitivity (66/82), which is the primary objective of this study.

Table 2: Median performance on the independent cohort (with 95% confidence intervals).

Model	Dice	PPV pixels	(mean) PPV clusters	(median) PPV clusters	IoU	Specificity	Sensitivity
MELD	0.358 (0.209–0.465)	0.261 (0.142–0.428)	0.464	1.000 (1.000–1.000)	0.218 (0.117–0.303)	46 / 83	57 / 82
Exp1	0.376 (0.238–0.529)	0.443 (0.243–0.619)	0.350	1.000 (1.000–1.000)	0.232 (0.138–0.359)	33 / 83	60 / 82
Exp2	0.168 (0.000–0.249)	0.349 (0.000–0.859)	0.268	1.000 (1.000–1.000)	0.092 (0.000–0.142)	13 / 83	53 / 82
Exp3: hemi	0.372 (0.261–0.515)	0.380 (0.181–0.601)	0.468	1.000 (1.000–1.000)	0.228 (0.150–0.347)	83 / 83	61 / 82
Exp3: lobe_regions	0.373 (0.283–0.507)	0.407 (0.233–0.589)	0.257	1.000 (0.667–1.000)	0.229 (0.165–0.340)	83 / 83	64 / 82
Exp3: hemi+ lobe_regions	0.363 (0.190–0.513)	0.434 (0.209–0.657)	0.359	1.000 (0.500–1.000)	0.222 (0.105–0.345)	81 / 83	61 / 82
Exp3: hemi+lobe	0.327 (0.164–0.512)	0.411 (0.236–0.667)	0.545	1.000 (0.667–1.000)	0.196 (0.090–0.344)	82 / 83	59 / 82
Exp3: hemi+lobe+ BlueBERT	0.362 (0.260–0.498)	0.469 (0.267–0.620)	0.356	1.000 (0.750–1.000)	0.221 (0.150–0.331)	80 / 83	63 / 82
Exp3: hemi+lobe+ PubmedBERT	0.347 (0.195–0.529)	0.457 (0.202–0.730)	0.258	1.000 (0.500–1.000)	0.210 (0.109–0.360)	83 / 83	66 / 82
Exp3: full_desc	0.423 (0.267–0.506)	0.484 (0.315-0.693)	0.460	1.000 (1.000–1.000)	0.268 (0.154-0.339)	75 / 83	60 / 82

Note. Colors denote rank within each column: best , second , third .
Exp labels. **Exp3: hemi** – hemisphere conditioning; **lobe_regions** – lobe- and region-level prompts; **hemi+lobe** – hemisphere + coarse lobe name; **hemi+lobe+BlueBERT** – same as previous but with BlueBERT instead of RadBERT; **full_desc** – full free-text description; **mixed** – mixture of prompts. **Exp1** and **Exp2** are ablation baselines.

2 Mixed Text

We investigated how different types of textual descriptions affect segmentation quality **(Exp3 mixed)**. During training, one available description was randomly sampled for each patient, whereas evaluation used a fixed prompt type. We also assessed robustness to *incorrect* prompts. All experiments employed **RadBERT** as the text encoder.

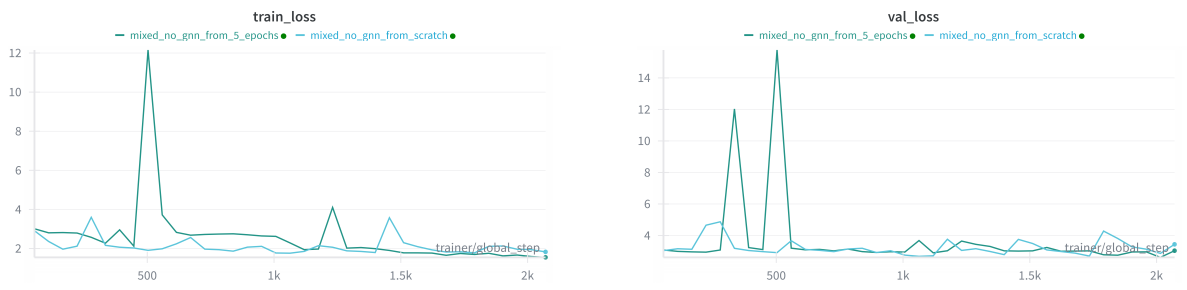
We compared two model options:

1. **Decoder open from the start**: the GuideDecoder is initialized with pre-trained weights from *Exp1* and trained from the first epoch.

2. **Decoder warm-up (5 epochs)**: for the first five epochs, we train only the GuideDecoder on fixed visual features. The choice of five is pragmatic: each experiment is an ensemble of models, a single model trains for roughly 8 hours, and typical runs span 20–30 epochs; hence, five warm-up epochs were deemed sufficient.

By Tables 3 and 4, excluding the *hemi* condition, a 5-epoch warm-up improves not only **Sensitivity** but also **Dice**, **IoU**, and **Specificity**. We interpret this as evidence that the warm-up phase allows the decoder to first learn how to parse the text and establish stable cross-attention to the visual embeddings without immediately perturbing the underlying MELD features.

This, in turn, reduces early-stage instability (especially under random prompt selection) and mitigates the typical effect of severe class imbalance, where the model suppresses activations to avoid false positives and produces near-empty masks. Consistently, the training and validation loss curves in Fig. 4 show that, compared to training from scratch, the warm-up schedule moderates large early loss spikes and keeps the optimization trajectory more controlled during the first epochs, indicating that the model continues to learn in this phase. After the warm-up, attention is more focused, **Sensitivity** increases, and **IoU** is maintained.



(a) Training loss.

(b) Validation loss.

Figure 4: Training and validation loss for the mixed cohort with and without a 5-epoch warm-up. The warm-up schedule reduces early spikes and yields a smoother optimization trajectory, indicating that the model is indeed learning during the warm-up phase.

In the main-cohort Table 4, *hemi* serves as a coarse cue: it specifies only the hemisphere and provides no regional localization. When the decoder is trained from the first epoch, this cue yields *moderate* performance (mid-range Dice/IoU with high PPV of clusters), i.e., the model can still rely on visual evidence while using the hemisphere as a weak prior. Under the 5-epoch freeze, however, the same *hemi* cue leads to a collapse: Dice and IoU approach zero while PPV (clusters) becomes very high, indicating sparse yet highly confident activations. This suggests that the warm-up phase can lock attention into an overly broad (or misaligned) pattern when the text is too unspecific, pushing the model toward near-empty masks to avoid false positives. When prompts are more specific (*hemi + lobe/regions* or *full_desc*), the search space narrows, attention concentrates on relevant areas, and the metrics—especially **Sensitivity** and **Dice**—improve over the *hemi*-freeze setting.

Table 3: Different description settings for **Exp3_mixed** models (values in parentheses indicate 95% confidence intervals). *Decoder open from the start* — GuideDecoder trained from the first epoch. Main cohort — correct descriptions.

Model	Text type	Dice	PPV pixels	(mean) PPV clusters	(median) PPV clusters	IoU	Specificity	Sensitivity
MELD	–	0.230 (0.107–0.320)	0.166 (0.086–0.220)	0.515	1.000 (1.000-1.000)	0.130 (0.057–0.190)	112 / 193	170 / 259
Exp3_mixed	hemi	0.224 (0.102–0.318)	0.226 (0.130–0.338)	0.519	0.667 (0.500 – 1.000)	0.126 (0.054–0.189)	186 / 193	172 / 259
Exp3_mixed	lobe_regions	0.236 (0.112–0.317)	0.225 (0.160–0.299)	0.515	0.667 (0.500 – 1.000)	0.134 (0.059–0.188)	186 / 193	179 / 259
Exp3_mixed	hemi+ lobe_regions	0.238 (0.121–0.304)	0.240 (0.160–0.306)	0.491	0.750 (0.500 – 1.000)	0.135 (0.064–0.179)	186 / 193	181 / 259
Exp3_mixed	hemi+lobe	0.219 (0.099–0.301)	0.216 (0.107–0.335)	0.532	0.667 (0.500 – 1.000)	0.123 (0.052–0.177)	186 / 193	168 / 259
Exp3_mixed	full_desc	0.238 (0.147–0.322)	0.229 (0.135–0.298)	0.495	0.500 (0.500 – 1.000)	0.135 (0.080–0.192)	186 / 193	181 / 259

Table 4: Different description settings for **Exp3_mixed freeze 5 epochs** models (values in parentheses indicate 95% confidence intervals). *Decoder warm-up for 5 epochs* — GuideDecoder unfrozen after epoch 5. Main cohort — correct descriptions.

Model	Text type	Dice	PPV pixels	(mean) PPV clusters	(median) PPV clusters	IoU	Specificity	Sensitivity
MELD	–	0.230 (0.107–0.320)	0.166 (0.086–0.220)	0.515	1.000 (1.000-1.000)	0.130 (0.057–0.190)	112 / 193	170 / 259
Exp3_mixed	hemi	0.000 (0.000–0.059)	0.000 (0.000–0.319)	0.782	0.500 (0.000 – 1.000)	0.000 (0.000–0.030)	193 / 193	135 / 259
Exp3_mixed	lobe_regions	0.274 (0.167–0.351)	0.225 (0.151–0.287)	0.459	1.000 (0.500 – 1.000)	0.159 (0.091–0.213)	193 / 193	186 / 259
Exp3_mixed	hemi+ lobe_regions	0.247 (0.160–0.316)	0.200 (0.133–0.276)	0.440	1.000 (0.500 – 1.000)	0.141 (0.087–0.187)	193 / 193	189 / 259
Exp3_mixed	hemi+ lobe	0.253 (0.164–0.306)	0.211 (0.127–0.286)	0.484	1.000 (0.500 – 1.000)	0.145 (0.089–0.180)	193 / 193	191 / 259
Exp3_mixed	full_desc	0.259 (0.162–0.337)	0.228 (0.160–0.329)	0.474	1.000 (0.500 – 1.000)	0.149 (0.088–0.203)	193 / 193	188 / 259

Wrong textual descriptions We now evaluate models with *incorrect* prompts, including the `no_text` setting (i.e., the model receives the placeholder string *full_brain* as input).

For `wrong_hemi` Tables 5 and 6, the strong performance drop is observed *only* in the model with the 5-epoch freeze; the non-frozen model remains at a moderate level (no collapse of Dice/IoU). Our interpretation is that the warm-up phase, when the decoder is trained in isolation, over-aligns attention to the contradictory textual cue (wrong hemisphere). This early misalignment is then hard to undo after the freeze is lifted.

For the other "wrong" prompts the picture is different. Their performance is often close to—and sometimes slightly better than—the results with correct prompts. This is plausible for three reasons:

1. **Partial but useful signal.** Even an inexact prompt still contains structure (e.g., hemisphere or a coarse lobe). This partial cue narrows the search region and reduces the need to scan the whole cortex.

2. **Robustness of the decoder.** Thanks to pretraining, the decoder tolerates mild text–image mismatch. When the text is unhelpful, it relies more on visual embeddings and uses the prompt only as a weak guide for attention.

3. **Regularization effect.** Slightly noisy text prevents the model from memorizing specific wording and forces it to rely on more general text–image relations. This acts as mild regularization: predictions become more conservative, reducing false positives (FP). Consequently, precision metrics (e.g., PPV) tend to increase, while recall and Dice typically remain stable.

Removing text makes all metrics drop. Without any linguistic prior, attention becomes broad and unspecific. Under strong class imbalance the model behaves very conservatively and triggers only at very high confidence. As a result, **Sensitivity** (recall) decreases because many true lesions are not activated (FN increase), and **Dice/IoU** also decline. At the same time, **PPV (clusters)** may appear high simply because the model predicts very few clusters and thus produces few FP.

Table 5: Different description settings with *wrong* text (without freezing). Main cohort — wrong descriptions

Model	Text type	Dice	PPV pixels	(mean) PPV clusters	(median) PPV clusters	IoU	Specificity	Sensitivity
MELD	–	0.230 (0.107–0.320)	0.166 (0.086–0.220)	0.515	1.000 (1.000-1.000)	0.130 (0.057–0.190)	112 / 193	170 / 259
Exp3_mixed	wrong_hemi	0.228 (0.111–0.312)	0.234 (0.148–0.346)	0.509	0.667 (0.500 – 1.000)	0.129 (0.059–0.185)	186 / 193	174 / 259
Exp3_mixed	wrong_hemi + correct_lobe_regions	0.240 (0.118–0.305)	0.240 (0.156–0.300)	0.486	0.667 (0.500 – 1.000)	0.136 (0.063–0.180)	186 / 193	180 / 259
Exp3_mixed	wrong_hemi + correct_lobe	0.219 (0.097–0.297)	0.221 (0.113–0.323)	0.533	0.667 (0.500 – 1.000)	0.123 (0.051–0.174)	186 / 193	169 / 259
Exp3_mixed	wrong_hemi + wrong_lobe_regions	0.240 (0.115–0.312)	0.232 (0.140–0.302)	0.488	0.500 (0.500 – 1.000)	0.136 (0.061–0.185)	186 / 193	180 / 259
Exp3_mixed	wrong_hemi + wrong_lobe	0.239 (0.117–0.308)	0.251 (0.165–0.324)	0.537	0.667 (0.500 – 1.000)	0.135 (0.062–0.182)	186 / 193	177 / 259
Exp3_mixed	no_text	0.139 (0.010–0.252)	0.179 (0.031–0.305)	0.528	1.000 (0.500 – 1.000)	0.075 (0.005–0.144)	144 / 193	154 / 259

However not all incorrect prompts are equally harmful. Prompts that are *contradictory* to the image (`wrong_hemi`) degrade performance the most. Inexact but *partially informative* prompts (e.g., rough lobe hints) can still help by providing a weak localization prior. Compared with `no_text`, even a rough prompt supplies enough context to stabilize attention and yields better overall metrics.

Table 6: Different description settings with *wrong* text (with 5 frozen epochs). Main cohort — wrong descriptions

Model	Text type	Dice	PPV pixels	(mean) PPV clusters	(median) PPV clusters	IoU	Specificity	Sensitivity
MELD	–	0.230 (0.107–0.320)	0.166 (0.086–0.220)	0.515	1.000 (1.000-1.000)	0.130 (0.057–0.190)	112 / 193	170 / 259
Exp3_mixed	wrong_hemi	0.007 (0.000–0.070)	0.005 (0.000–0.292)	0.748	0.500 (0.500 – 1.000)	0.004 (0.000–0.036)	193 / 193	135 / 259
Exp3_mixed	wrong_hemi + correct_lobe_regions	0.242 (0.161–0.318)	0.198 (0.135–0.271)	0.434	1.000 (0.500 – 1.000)	0.138 (0.087–0.189)	193 / 193	188 / 259
Exp3_mixed	wrong_hemi + correct_lobe	0.251 (0.166–0.309)	0.216 (0.126–0.289)	0.495	1.000 (0.500 – 1.000)	0.144 (0.090–0.183)	193 / 193	191 / 259
Exp3_mixed	wrong_hemi+ wrong_lobe_regions	0.250 (0.171–0.302)	0.186 (0.127–0.275)	0.438	1.000 (0.500 – 1.000)	0.143 (0.093–0.178)	193 / 193	189 / 259
Exp3_mixed	wrong_hemi+ wrong_lobe	0.246 (0.175–0.335)	0.201 (0.131–0.285)	0.468	1.000 (0.500 – 1.000)	0.141 (0.096–0.201)	193 / 193	189 / 259
Exp3_mixed	no_text	0.012 (0.000–0.133)	0.047 (0.000–0.306)	0.689	1.000 (0.000 – 1.000)	0.006 (0.000–0.071)	179 / 193	143 / 259

Independent cohort On the independent cohort Tables 7 – 10, for both correct and incorrect prompts, the model with a 5-epoch decoder freeze generally detects more lesions (higher **Sensitivity**). For `hemi+lobe_regions`, `hemi+lobe`, and `full_desc`, this comes with a small decrease in **Dice**/**IoU** (about 1–2 %), which we consider non-critical.

Comparing `lobe` versus `lobe_regions`, the `lobe` setting yields slightly higher recall and marginally better aggregate metrics. This suggests that even coarse descriptions — without detailed areas and specific anatomical names — provide enough semantic signal for high-quality predictions.

Table 7: Different description settings for **Exp3_mixed** models (values in parentheses indicate 95% confidence intervals). *Decoder open from the start* — GuideDecoder trained from the first epoch. Independent cohort — correct descriptions.

Model	Text type	Dice	PPV pixels	(mean) PPV clusters	(median) PPV clusters	IoU	Specificity	Sensitivity
MELD	–	0.358 (0.209–0.465)	0.261 (0.142–0.428)	0.464	1.000 (1.000–1.000)	0.218 (0.117–0.303)	46 / 83	57 / 82
Exp3_mixed	hemi	0.337 (0.220–0.509)	0.433 (0.162–0.678)	0.632	1.000 (1.000 – 1.000)	0.203 (0.124–0.341)	78 / 83	55 / 82
Exp3_mixed	lobe_regions	0.370 (0.162–0.478)	0.345 (0.157–0.526)	0.612	1.000 (0.500 – 1.000)	0.227 (0.088–0.314)	78 / 83	55 / 82
Exp3_mixed	hemi+ lobe_regions	0.350 (0.164–0.479)	0.411 (0.227–0.544)	0.577	1.000 (0.500 – 1.000)	0.212 (0.089–0.315)	78 / 83	57 / 82
Exp3_mixed	hemi+lobe	0.359 (0.210–0.482)	0.446 (0.217–0.649)	0.633	1.000 (1.000 – 1.000)	0.219 (0.117–0.318)	78 / 83	56 / 82
Exp3_mixed	full_desc	0.356 (0.242–0.500)	0.416 (0.207–0.551)	0.574	1.000 (0.833 – 1.000)	0.217 (0.138–0.334)	78 / 83	57 / 82

Table 8: Different description settings for **Exp3_mixed freeze 5 epochs** models (values in parentheses indicate 95% confidence intervals). *Decoder warm-up for 5 epochs* — GuideDecoder unfrozen after epoch 5. Independent cohort — correct descriptions.

Model	Text type	Dice	PPV pixels	(mean) PPV clusters	(median) PPV clusters	IoU	Specificity	Sensitivity
MELD	–	0.358 (0.209–0.465)	0.261 (0.142–0.428)	0.464	1.000 (1.000–1.000)	0.218 (0.117–0.303)	46 / 83	57 / 82
Exp3_mixed	hemi	0.000 (0.000–0.147)	0.000 (0.000–0.665)	0.957	0.000 (0.000 – 1.000)	0.000 (0.000–0.079)	83 / 83	40 / 82
Exp3_mixed	lobe_regions	0.363 (0.188–0.518)	0.393 (0.128–0.571)	0.492	1.000 (0.500 – 1.000)	0.221 (0.106–0.350)	83 / 83	60 / 82
Exp3_mixed	hemi+ lobe_regions	0.331 (0.182–0.519)	0.309 (0.107–0.509)	0.417	1.000 (0.500 – 1.000)	0.198 (0.101–0.351)	83 / 83	62 / 82
Exp3_mixed	hemi+lobe	0.341 (0.183–0.490)	0.380 (0.159–0.522)	0.489	1.000 (0.750 – 1.000)	0.206 (0.101–0.325)	83 / 83	63 / 82
Exp3_mixed	full_desc	0.340 (0.124–0.492)	0.382 (0.125–0.588)	0.562	1.000 (0.833 – 1.000)	0.205 (0.066–0.327)	83 / 83	58 / 82

Table 9: Different description settings with *wrong* text (without freezing). Independent cohort — wrong descriptions

Model	Text type	Dice	PPV pixels	(mean) PPV clusters	(median) PPV clusters	IoU	Specificity	Sensitivity
MELD	–	0.358 (0.209–0.465)	0.261 (0.142–0.428)	0.464	1.000 (1.000–1.000)	0.218 (0.117–0.303)	46 / 83	57 / 82
Exp3_mixed	wrong_hemi	0.348 (0.092–0.516)	0.356 (0.137–0.630)	0.628	1.000 (1.000 – 1.000)	0.210 (0.051–0.348)	78 / 83	55 / 82
Exp3_mixed	wrong_hemi + correct_lobe_regions	0.359 (0.140–0.491)	0.414 (0.231–0.542)	0.578	1.000 (1.000 – 1.000)	0.219 (0.076–0.326)	78 / 83	57 / 82
Exp3_mixed	wrong_hemi + correct_lobe	0.351 (0.206–0.507)	0.478 (0.253–0.650)	0.646	1.000 (1.000 – 1.000)	0.213 (0.115–0.340)	78 / 83	56 / 82
Exp3_mixed	wrong_hemi + wrong_lobe_regions	0.372 (0.240–0.514)	0.428 (0.266–0.557)	0.642	1.000 (1.000 – 1.000)	0.229 (0.136–0.346)	78 / 83	58 / 82
Exp3_mixed	wrong_hemi + wrong_lobe	0.376 (0.201–0.489)	0.466 (0.255–0.663)	0.620	1.000 (1.000 – 1.000)	0.232 (0.112–0.324)	78 / 83	57 / 82
Exp3_mixed	no_text	0.313 (0.000–0.463)	0.375 (0.000–0.656)	0.415	1.000 (0.000 – 1.000)	0.186 (0.000–0.301)	64 / 83	48 / 82

Table 10: Different description settings with *wrong* text (with 5 frozen epochs). Independent cohort — wrong descriptions

Model	Text type	Dice	PPV pixels	(mean) PPV clusters	(median) PPV clusters	IoU	Specificity	Sensitivity
MELD	–	0.358 (0.209–0.465)	0.261 (0.142–0.428)	0.464	1.000 (1.000–1.000)	0.218 (0.117–0.303)	46 / 83	57 / 82
Exp3_mixed	wrong_hemi	0.000 (0.000–0.172)	0.000 (0.000–0.780)	0.936	0.000 (0.000 – 1.000)	0.000 (0.000–0.094)	83 / 83	40 / 82
Exp3_mixed	wrong_hemi + correct_lobe_regions	0.331 (0.188–0.513)	0.304 (0.114–0.508)	0.424	1.000 (0.500 – 1.000)	0.199 (0.105–0.345)	83 / 83	62 / 82
Exp3_mixed	wrong_hemi + correct_lobe	0.323 (0.180–0.485)	0.349 (0.165–0.521)	0.476	1.000 (0.833 – 1.000)	0.193 (0.099–0.321)	83 / 83	63 / 82
Exp3_mixed	wrong_hemi + wrong_lobe_regions	0.312 (0.129–0.487)	0.319 (0.069–0.515)	0.429	1.000 (0.750 – 1.000)	0.185 (0.069–0.322)	83 / 83	62 / 82
Exp3_mixed	wrong_hemi + wrong_lobe	0.298 (0.142–0.480)	0.298 (0.093–0.496)	0.436	1.000 (0.750 – 1.000)	0.176 (0.076–0.316)	83 / 83	63 / 82
Exp3_mixed	no_text	0.001 (0.000–0.288)	0.003 (0.000–0.716)	0.618	1.000 (0.000 – 1.000)	0.001 (0.000–0.168)	70 / 83	43 / 82

3 Linking MELD to GNN

In these experiments we investigated the effect of connecting different numbers of MELD feature stages to the GNN block. The rationale was that higher MELD stages may produce sparse representations, while lower stages provide richer local detail. By progressively adding stages from top to bottom, we aimed to evaluate how multi-stage integration influences model performance. To isolate this effect, the text encoder and GuideDecoder were disabled.

On the main cohort Tables 11, the configuration with **three GNN layers** offers the best trade-off across Dice, $\text{PPV}_{\text{pixels}}$, and IoU, and detects **one** more lesion than the no-GNN baseline. Connecting **all seven** MELD stages to the GNN yields the highest lesion count (sensitivity increased approximately by 3%: 188/259 vs. 180/259), with only a modest drop in Dice and $\text{PPV}_{\text{pixels}}$ (about 1–4%). This gain is consistent with SAGEConv's neighborhood aggregation: increasing depth expands each node's receptive field and multi-stage inputs inject complementary local/global context, enabling detection of additional lesion clusters that are missed with purely local features.

Table 11: Different number of MELD stages connected to the GNN block (values in parentheses indicate 95% confidence intervals). Main cohort

Experiment	**Dice**	**PPV pixels**	**(mean) PPV clusters**	**IoU**	**Sensitivity**
Exp1: 0 layers	0.240 (0.129–0.324)	0.217 (0.151–0.319)	0.532	0.136 (0.069–0.193)	180 / 259
Exp1: 1 layer	0.235 (0.096–0.313)	0.235 (0.148–0.362)	0.603	0.133 (0.051–0.186)	174 / 259
Exp1: 2 layers	0.234 (0.126–0.334)	0.219 (0.136–0.371)	0.573	0.132 (0.067–0.200)	177 / 259
Exp1: 3 layers	0.249 (0.114–0.332)	0.224 (0.142–0.341)	0.479	0.142 (0.060–0.199)	181 / 259
Exp1: 4 layers	0.231 (0.097–0.304)	0.180 (0.128–0.283)	0.540	0.130 (0.051–0.179)	177 / 259
Exp1: 5 layers	0.223 (0.081–0.312)	0.172 (0.090–0.261)	0.533	0.125 (0.042–0.185)	171 / 259
Exp1: 6 layers	0.208 (0.121–0.312)	0.219 (0.147–0.336)	0.539	0.116 (0.065–0.185)	179 / 259
Exp1: 7 layers	0.206 (0.119–0.293)	0.219 (0.135–0.341)	0.436	0.115 (0.063–0.172)	188 / 259

On the independent (Bonn) cohort Table 12, the **5-layer** configuration gives the best overall trade-off: it attains the *second-highest* Dice and IoU while detecting the most lesions (63/82). Compared with the no-GNN baseline (0 layers), the gap is ≈3% in Dice and ≈5% in IoU, highlighting the benefit of the proposed GNN block.

For subsequent experiments we keep the **3-layer** and **5-layer** variants, and will compare them in the final study to determine the optimal number of connected MELD stages.

Table 12: Different number of MELD stages connected to the GNN block (values in parentheses indicate 95% confidence intervals). Independent cohort (Bonn Dataset)

Experiment	Dice	PPV pixels	(mean) PPV clusters	IoU	Sensitivity
Exp1: 0 layers	0.371 (0.289–0.520)	0.438 (0.246–0.660)	0.680	0.227 (0.169–0.351)	60 / 82
Exp1: 1 layer	0.342 (0.083–0.515)	0.528	0.819	0.206 (0.043–0.347)	54 / 82
Exp1: 2 layers	0.383 (0.216–0.506)	0.432 (0.243–0.663)	0.729	0.237 (0.123–0.339)	58 / 82
Exp1: 3 layers	0.390 (0.302–0.517)	0.503	0.619	0.242 (0.178–0.349)	62 / 82
Exp1: 4 layers	0.458 (0.279–0.600)	0.491 (0.186–0.582)	0.653	0.297 (0.162–0.428)	60 / 82
Exp1: 5 layers	0.432 (0.303–0.549)	0.410 (0.249–0.606)	0.555	0.275 (0.179–0.378)	63 / 82
Exp1: 6 layers	0.382 (0.292–0.544)	0.584	0.644	0.236 (0.171–0.373)	60 / 82
Exp1: 7 layers	0.402 (0.284–0.493)	0.375 (0.232–0.634)	0.451	0.252 (0.165–0.327)	61 / 82

4 Supplementary results

4.1 Text distribution

To characterise potential sources of bias, we analyse the distribution of text-derived region labels at three anatomical levels: hemispheres, lobes, and lobe-regions. This reveals whether particular regions or hemispheres are overrepresented, which could drive overfitting or induce a preference for frequent anatomical terms during training.

Unless stated otherwise, all statistics are computed on the *entire* dataset (train, validation, and test). The label *No lesion detected* denotes healthy controls and is excluded from validation and test to avoid artificially inflating evaluation metrics. After this exclusion its effective frequency is roughly halved, although the overall imbalance across regions persists.

Hemisphere. The distribution of hemisphere labels is shown in Figure 5.

In the table *"No lesion detected"* means these are healthy patients.

In the tables, *"No lesion detected"* denotes **healthy control** scans with no radiologically/annotationally confirmed FCD.

The number of cases for the left and right hemispheres is approximately equal, with only a small, statistically insignificant difference. This indicates that the *entire dataset* is well balanced across hemispheres. Consequently, a model trained on these data is unlikely to learn a systematic bias towards either hemisphere, which would otherwise reduce its

ability to generalise.

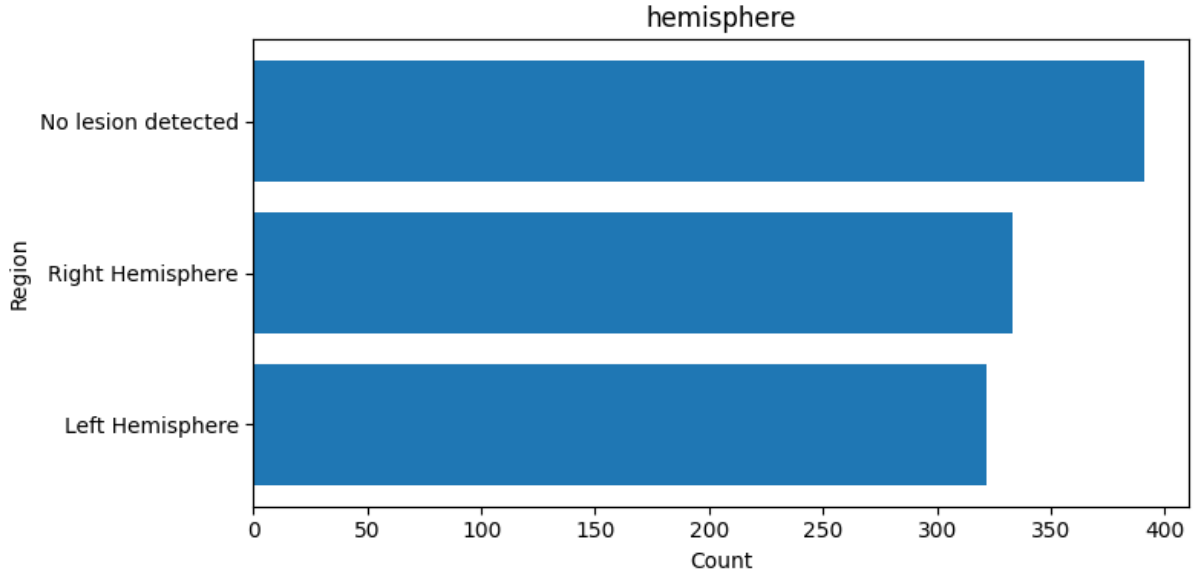


Figure 5: Hemisphere distribution

Lobes. The lobe-level distribution, presented in Figure 6, demonstrates a moderate imbalance. The *frontal*, *limbic*, and *parietal* lobes occur much more frequently than the *occipital*, *insular*, *subcallosal*, and especially the *brainstem*, which are relatively rare. Such imbalance may cause the model to "memorise" common patterns such as *frontal lobe* while underperforming on underrepresented categories like *brainstem* or *insula*. However, compared to the following plot, the lobe regons imbalance can be considered moderate and still suitable for training, especially if loss weighting or other balancing techniques are used.

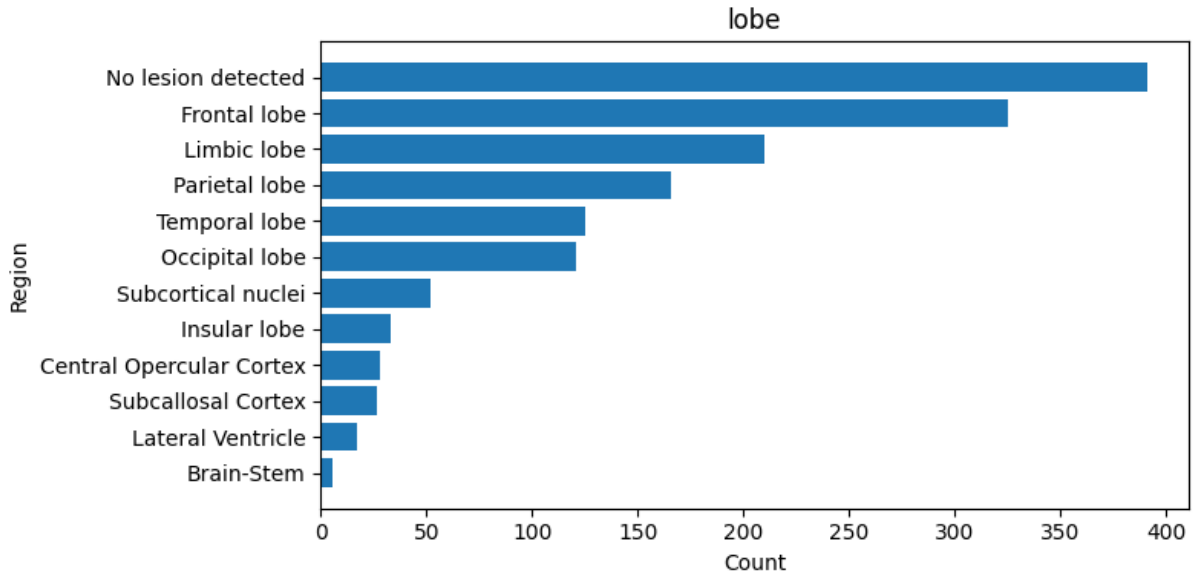


Figure 6: Lobe distribution

Lobe regions. The third plot (Figure 7) provides the distribution for specific anatomical regions within the lobes. Here, the imbalance becomes much more pronounced. The label *No lesion detected* dominates the dataset, with a frequency exceeding 400, whereas many regions appear only one to three times.

This extremely long-tailed distribution suggests a risk of overfitting to the most frequent labels. As a result, the model may bias its *spatial* predictions toward regions whose textual labels are frequent in the training data (e.g., the *frontal gyrus*), assigning higher lesion probabilities there. This imbalance poses a risk for models relying on textual embeddings, since they may learn frequency-driven rather than semantically meaningful representations.

Discussion. From these findings, we conclude that using fine-grained region names directly as text input would likely lead to model overfitting, given the strong imbalance and the large number of rare categories. The extreme class skew among fine-grained lobe–region labels makes direct use of their verbatim names as text inputs ill-suited: models tend to memorise frequent terms and underfit rare ones, which harms generalisation. To further mitigate imbalance, one could consider data augmentation or class-weighted loss functions, at the cost of longer training and increased computational complexity. Nonetheless, such strategies would likely improve model robustness and generalisation performance.

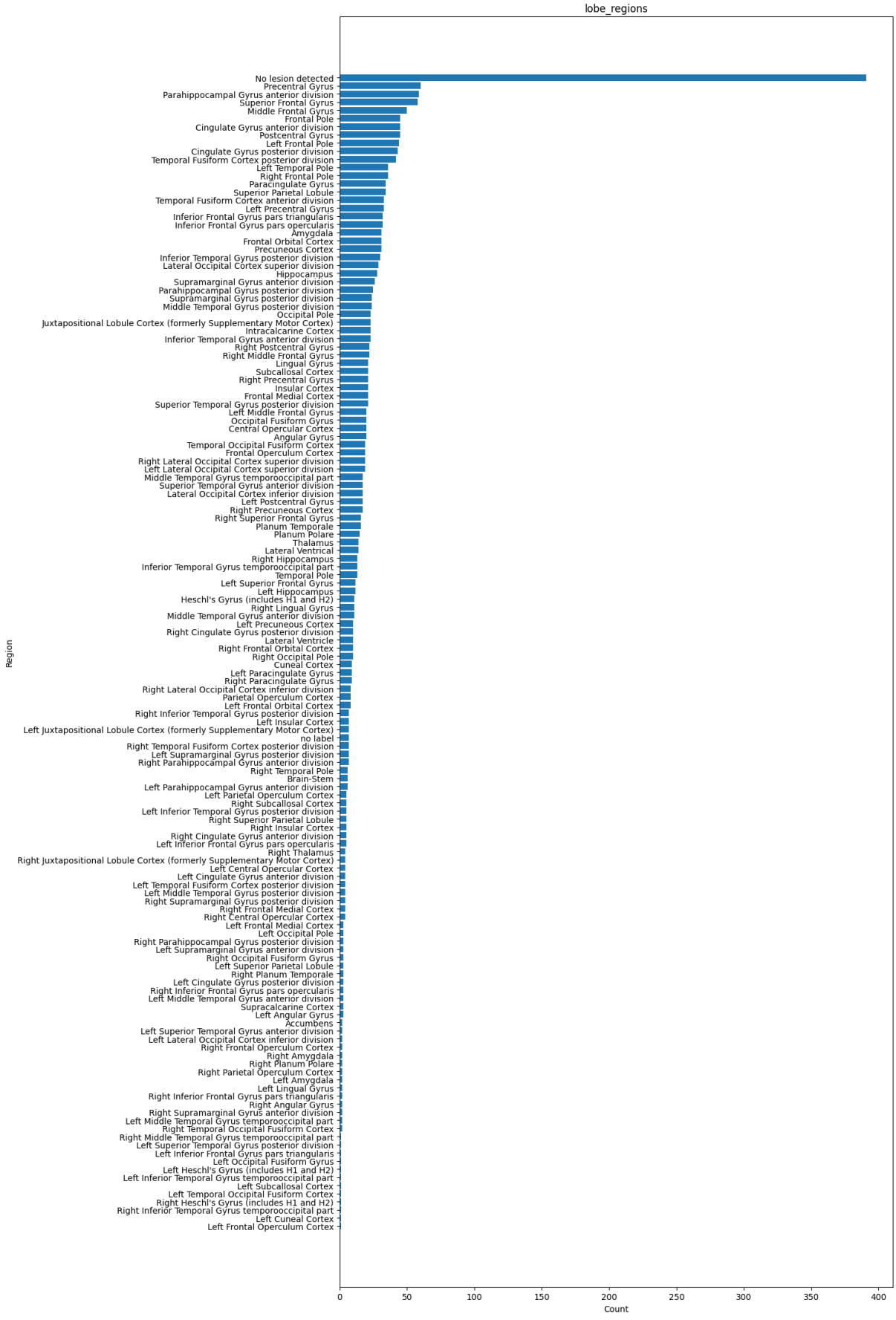


Figure 7: Lobe–region distribution

4.2 Semantic similarity analysis of frontal lobe regions

Table 13 reports cosine similarities between the term "frontal lobe" and several related or distinct brain regions across multiple biomedical language models. In general, models that have been pretrained on large-scale biomedical or radiological corpora (e.g., BioClinicalBERT, PubMedBERT, RadBERT-RoBERTa) achieve higher coherence among semantically related regions.

Table 13: Cosine similarity between "frontal lobe" and related region terms across different biomedical language models.

Model	Prefrontal cortex	Inferior frontal gyrus	Right	Right frontal	Temporal lobe	Parietal lobe
bionlp/bluebert_pubmed_mimic_uncased_L-12_H-768_A-12	0.882	0.812	0.792	0.904	0.968	0.892
emilyalsentzer/Bio_ClinicalBERT	0.909	0.877	0.781	0.941	0.982	0.932
microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext	0.974	0.972	0.931	0.971	0.985	0.987
StanfordAIMI/RadBERT	0.504	0.431	0.369	0.654	0.816	0.599
zzxslp/RadBERT-RoBERTa-4m	0.971	0.946	0.921	0.960	0.971	0.976
microsoft/BiomedVLP-CXR-BERT-general	0.489	0.623	0.572	0.680	0.770	0.810
cambridgeltl/SapBERT-from-PubMedBERT-fulltext	0.762	0.587	0.360	0.689	0.577	0.703
allenai/scibert_scivocab_uncased	0.903	0.923	0.748	0.917	0.946	0.976
intfloat/e5-base	0.893	0.899	0.771	0.892	0.892	0.889

When comparing regions that are subsets of the frontal lobe ("prefrontal cortex", "inferior frontal gyrus") to lateral or distinct terms ("right", "right frontal"), we observe the following pattern:

- The average cosine similarity with "frontal lobe" is approximately 0.81 for "prefrontal cortex" and 0.79 for "inferior frontal gyrus";
- The similarity for the direction-only term "right" is notably lower (around 0.69);
- The compound term "right frontal"—which shares the lexical component "frontal" – achieves higher similarity (around 0.85).

This indicates that biomedical embeddings primarily capture lexical and contextual overlap rather than strict anatomical hierarchy. Consequently, regions explicitly containing the token *frontal* are closer to "frontal lobe", even if they represent distinct subareas or orientations. Moreover, unrelated lobes such as the temporal and parietal lobes still show relatively high similarity (average 0.88 and 0.86), suggesting that model semantics cluster general cortical regions together.

Based on Table 13, we prefer models that assign high, consistent similarity to frontal-lobe subsets ("prefrontal cortex", "inferior frontal gyrus") and rank the compound term "right frontal" above the direction-only term "right". The most reliable performance is observed for **PubMedBERT**, **RadBERT-RoBERTa-4m**, **BioClinicalBERT**, **BlueBERT**, and **SciBERT**; these models can be used for generating text embeddings. By contrast, **StanfordAIMI/RadBERT**, **BiomedVLP-CXR-BERT-general**, and **SapBERT** yield low subset similarities and inconsistent rankings; therefore, we should not use them, as they fail to capture fine-grained semantic relations between closely related brain regions.

4.3 Final Experiments

By summarizing the results from the previous chapters, we conclude that the most appropriate text model is PubMedBERT, GNN blocks with 3 and 5 connections, and unfreezing the pretrained Exp1 decoder after 5 epochs. We will use all of these parameters to conduct experiments in order to observe the improvements in the results. For these experiments, we will use "hemisphere", "lobe", and "hemisphere + lobe" as text prompts. Using "lobe_regions" and "full text descriptions" does not make sense, as we discussed before. Additionally, we will train the model with mixed text, and we will also add an extra column that would otherwise contain no text; instead of an empty string, this column will contain the "full brain" description.

Table 14: Median performance on the main cohort (with 95% confidence intervals).

Model	Dice	PPV pixels	(mean) PPV clusters	(median) PPV clusters	IoU	Specificity	Sensitivity
MELD	0.230 (0.107–0.320)	0.166 (0.086–0.220)	0.515	1.000 (1.000-1.000)	0.130 (0.057–0.190)	–	170 / 259
Exp3: hemi	0.0.233 (0.158–0.314)	0.219 (0.140–0.314)	0.459	1.000 (0.500–1.000)	0.132 (0.086–0.186)	–	182 / 259
Exp3: lobe	0.241 (0.174–0.352)	0.216 (0.128–0.314)	0.466	1.000 (1.000–1.000)	0.137 (0.095–0.213)	–	186 / 259

Note. Colors denote rank within each column: best , second , third .
Exp labels. **Exp3: hemi** – hemisphere conditioning; **lobe_regions** – lobe- and region-level prompts; **hemi+lobe** – hemisphere + coarse lobe name; **hemi+lobe+BlueBERT** – same as previous but with BlueBERT instead of RadBERT; **full_desc** – full free-text description; **mixed** – mixture of prompts. **Exp1** and **Exp2** are ablation baselines.

Table 15: Median performance on the independent cohort (with 95% confidence intervals).

Model	Dice	PPV pixels	(mean) PPV clusters	(median) PPV clusters	IoU	Specificity	Sensitivity
MELD	0.358 (0.209–0.465)	0.261 (0.142–0.428)	0.464	1.000 (1.000–1.000)	0.218 (0.117–0.303)	–	57 / 82
Exp3: hemi	0.405 (0.286–0.560)	0.388 (0.201–0.605)	0.360	1.000 (0.750–1.000)	0.254 (0.150–0.347)	–	65 / 82
Exp3: lobe	0.422 (0.330–0.560)	0.438 (0.236–0.615)	0.319	1.000 (1.000–1.000)	0.268 (0.197–0.389)	–	69 / 82

Note. Colors denote rank within each column: best , second , third .
Exp labels. **Exp3: hemi** – hemisphere conditioning; **lobe_regions** – lobe- and region-level prompts; **hemi+lobe** – hemisphere + coarse lobe name; **hemi+lobe+BlueBERT** – same as previous but with BlueBERT instead of RadBERT; **full_desc** – full free-text description; **mixed** – mixture of prompts. **Exp1** and **Exp2** are ablation baselines.

TODO

General conclusions

- From the **Basic Experiments** chapter, we conclude that:

  - **Exp3_hemi+lobe_regions** shows the most balanced performance, achieving the **highest sensitivity**. Incorporating hemisphere and lobe-region information constrains the search space and improves localization.
  - **Exp3_hemi+lobe+PubMedBERT** achieves the best overall trade-off on both cohorts. The concise prompt without redundant context, combined with PubMedBERT's domain-specific pretraining, consistently improves all metrics, highlighting its influence and importance for the final results.
  - **Exp3_full_desc** underperforms compared to the hemi+lobe variants, suggesting that long atlas-style descriptions introduce noise and redundancy. In contrast, shorter, targeted prompts generalize better.

- From the **Mixed Text** chapter, we conclude that training the *model* on mixed-type descriptions yields better results than the **MELD** baseline, but remains slightly worse than models trained on a single, well-specified prompt type. Notably, replacing correct descriptions with incorrect ones does not collapse performance, which demonstrates robustness and stability of the proposed architecture.

- From the **Linking MELD to GNN** chapter, we conclude that adding an additional GNN block that aggregates information from neighboring nodes helps the model accumulate contextual information more effectively. This leads to improved performance and higher sensitivity. In future experiments, integrating this block into the final architecture could further enhance results.

- From the **Text Distribution** chapter, we conclude that conditioning on fine-grained region names as textual inputs tends to induce overfitting, because the underlying label distribution is heavily imbalanced and long-tailed. Frequent categories dominate the learning signal, while numerous rare classes lack sufficient variability for robust generalization.

- From the **Semantic Similarity Analysis of Frontal Lobe Regions** chapter, we conclude that the choice of the text encoder is crucial. Biomedical language models pretrained on domain-specific corpora (e.g., PubMedBERT, RadBERT, BioClinicalBERT) achieve higher semantic coherence between anatomically related brain regions, while general-purpose or less specialized models show inconsistent similarity patterns. This highlights that semantically aligned text embeddings can improve multimodal fusion and ultimately enhance the model's interpretability and performance.