8. Supplementary Material

8.1. Data Preprocessing

We harmonize the slice thickness of all images such that all the relevant anatomy fits in a 32 × 512 × 512 volume. The resampling of thick-sliced head CTs to thicker slices will create interpolation artifacts along the skull, if the new thickness is not a multiple of the original one. As such, we resample images with a slice-thickness of 2.5 mm to 5.0 mm. Images with a slice thickness above 4.0 mm are not resampled. As some images extend up to the patient's shoulders, we crop volumes with more than 32 slices. Since no masks are available for CQ500 and PC, we use an *nnUNet* model [15] pre-trained on INST to produce a pre-segmentation of all ICH.

8.2. Model Training

We use the official data split released with the INST and BHSD datasets. Due to our data curation process, our results are not directly comparable to other publications. Evaluating our models on the original data splits is also impossible, as the annotation does not exist for images which have not been selected. As some cases do not contain any relations, we only keep images with relations for relation detection. Split sizes can be found in (Tabs. 5 and 6). Additionally, given the number and length of experiments (7h for object detection and 1h for relation prediction), it is not feasible to find optimal hyperparameters for all setups. As such, we use the same hyperparameters for centralized and federated experiments. Exact splits and detailed configuration files will be made available along the configuration files.

Dataset	Training	Validation	Test
INSTANCE2022	81	12	27
Private Cohort	41	7	19
BHSD	51	10	39
CQ500	86	14	57

Table 5. Split sizes for object detection experiments.

Dataset	Training	Validation	Test
INSTANCE2022	56	9	24
Private Cohort	38	7	16
BHSD	35	6	29
CQ500	51	11	30

Table 6. Split sizes for relation prediction experiments.

8.3. Implementation Details

The methods are implemented in Python 3.10 and PyTorch 2.4 [22] and make use of the Voxel Scene Graph Generation framework [26] and of the open-source framework *TheODen*[2] as an overlay to enable federated training. The federated training is not simulated, as each of the 4 clients had an own computer with an NVIDIA RTX 4090 GPU. An additional fifth computer is used for model aggregation and does not require a GPU. For each federated training, we train for 16000 and 200 steps respectively for object detection and relation prediction. Each training round lasts 25 and 5 steps respectively and is followed by an aggregation round. The V-RAM requirement was optimized to use the GPUs' full 24 GB of memory and allows for a batch size of 1 and 13 respectively for object detection and relation prediction.

8.4. Detailed Results

The tables that follow provide results for each dataset separately.

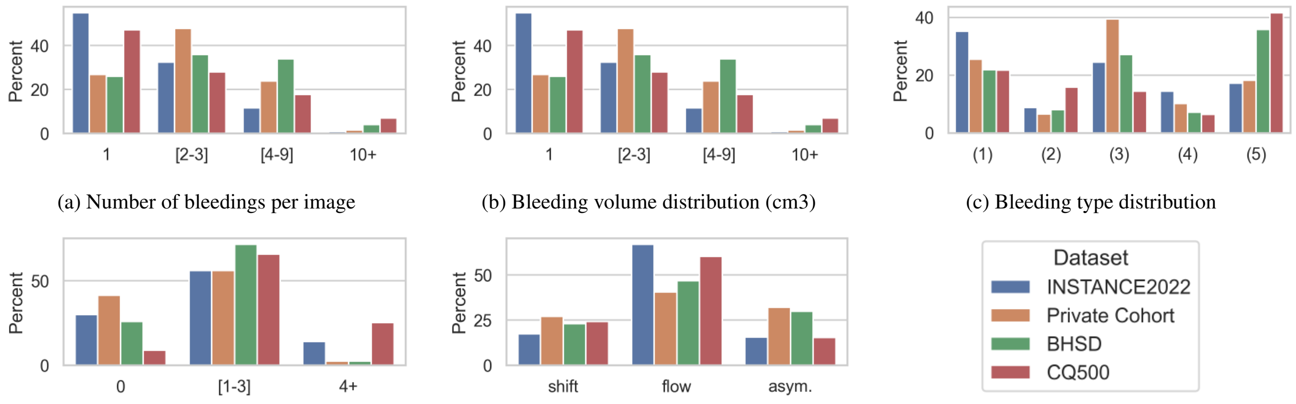
	Ventricle System Segmentation			
Train	INST	PC	BHSD	CQ500
INST	73.8±2.0	65.3±2.8	61.0±6.5	60.0±9.3
PC	56.2±7.1	80.4±2.3	28.9±9.5	16.9±6.1
BHSD	**78.5±1.0**	69.3±4.3	**78.3±1.1**	**78.1±1.0**
CQ500	77.2±1.0	71.0±3.3	76.0±2.1	**78.2±2.1**
FedAvg	78.2±0.4	**81.2±1.0**	72.8±2.9	74.6±3.5
FedSGD	77.3±0.6	79.6±0.9	72.0±2.9	73.6±2.6
all	77.0±0.6	79.1±0.9	72.9±3.3	74.9±2.1

Table 7. **Patient Dice score** for the **ventricle system**, when training in a centralized setup using one or all datasets or using FedL.

	Midline Segmentation			
Train	INST	PC	BHSD	CQ500
INST	65.5±1.6	50.1±2.8	50.1±3.3	48.3±4.4
PC	55.1±5.9	72.1±1.0	28.1±8.2	22.1±7.6
BHSD	72.0±1.2	55.7±4.6	67.6±2.5	66.8±2.3
CQ500	72.5±0.6	58.4±2.3	67.7±0.9	67.2±2.5
FedAvg	**73.6±0.5**	**70.1±0.8**	**68.4±1.1**	**67.8±1.6**
FedSGD	70.4±1.5	64.9±1.1	62.1±2.7	60.6±2.6
all	71.7±0.4	64.5±0.9	64.8±1.9	62.4±2.0

Table 8. **Patient Dice score** for the **midline**, when training in a centralized setup using one or all datasets or using FedL.

[2] https://github.com/MECLabTUDA/TheODen



(a) Number of bleedings per image

(b) Bleeding volume distribution (cm3)

(c) Bleeding type distribution

(d) Number of relations per image

(e) Relation distribution

Figure 4. Distribution of bleedings and relations for each dataset. All datasets show in bleeding representation whether regarding their number, volume or type. The bleeding types refer to: 1) intraparenchymal, 2) epidural or subdural, 3) intraventricular, 4) basal subarachnoidal, and 5) non-basal subarachnoidal. "Basal" refers to the basal cistern, where the subarachnoidal bleeding can be more prominent.

	Bleeding Segmentation			
Train	INST	PC	BHSD	CQ500
INST	79.1±0.7	81.8±0.3	57.2±1.2	40.5±2.5
PC	62.0±4.7	**83.1±0.6**	29.4±5.0	19.0±3.6
BHSD	74.2±0.6	79.0±0.9	65.5±0.8	49.8±1.9
CQ500	70.0±2.0	73.4±1.0	61.3±1.1	52.7±1.3
FedAvg	**81.0±0.4**	82.6±0.5	**70.1±0.5**	**55.3±1.6**
FedSGD	79.6±0.7	81.4±0.6	68.2±0.9	53.4±1.4
all	78.3±0.8	80.0±0.7	67.3±0.8	53.5±1.2

Table 9. **Patient Dice score** for **bleeding**, when training in a centralized setup using one or all datasets or using FedL.

		INST Dataset					
		Predicate Classification			Scene Graph Generation		
Method	Model	R@8↑	mR@8↑	mAP@8↑	R@8↑	mR@8↑	mAP@8↑
INST dataset	*MOTIF*	62.9±5.0	65.1±5.7	57.2±6.8	47.1±6.1	48.7±6.3	38.0±5.9
INST dataset	*IMP*	61.5±3.2	65.0±2.8	53.9±3.4	58.1±4.6	60.1±3.9	24.7±5.4
Avg. unseen	*MOTIF*	63.6±16.3	65.0±15.5	62.8±8.8	46.8±16.8	46.4±15.6	38.7±15.8
Avg. unseen	*IMP*	61.7±12.0	63.8±11.2	55.8±11.4	52.5±14.6	54.0±14.5	25.9±7.0
FedAvg	*Fed-MOTIF*	**76.9±4.6**	**77.8±4.9**	**70.5±2.0**	60.7±6.0	58.3±6.9	**45.3±8.2**
FedAvg	*Fed-IMP*	73.1±3.1	73.8±3.3	65.5±5.3	**65.1±3.4**	**65.8±3.6**	29.6±1.1
FedSGD	*Fed-MOTIF*	43.3±9.6	42.7±10.5	69.2±8.1	46.7±6.5	43.7±3.7	57.1±6.0
FedSGD	*Fed-IMP*	54.5±1.8	55.3±2.2	50.6±8.0	46.0±4.6	44.2±4.2	31.0±2.7
All seen	*MOTIF*	74.3±1.5	76.1±2.4	60.2±4.4	57.6±6.1	57.6±5.9	37.1±2.2
All seen	*IMP*	70.7±1.9	73.8±1.0	61.2±2.3	60.3±5.7	62.2±5.1	27.9±3.0

Table 10. Results for the **Predicate Classification** (left) and **Scene Graph Generation** (right) tasks for Centralized and Federated Learning on the **INST** dataset). All configurations are run 5 times using random seeds.

		PC Dataset					
		Predicate Classification			Scene Graph Generation		
Method	Model	R@8↑	mR@8↑	mAP@8↑	R@8↑	mR@8↑	mAP@8↑
PC dataset	*MOTIF*	76.6±6.5	80.6±4.7	78.3±2.5	58.3±4.3	64.0±3.7	51.1±8.9
PC dataset	*IMP*	63.8±7.3	66.1±9.0	73.8±5.0	57.7±8.2	59.1±4.7	21.0±7.4
Avg. unseen	*MOTIF*	67.9±14.3	70.7±16.2	74.8±4.7	53.1±11.7	60.6±11.9	40.8±9.9
Avg. unseen	*IMP*	68.2±9.6	74.0±9.5	67.7±9.5	55.2±10.5	63.7±11.1	26.1±7.7
FedAvg	*Fed-MOTIF*	72.8±5.6	78.6±5.9	**76.7±5.2**	64.2±1.6	71.7±1.0	**46.9±5.2**
FedAvg	*Fed-IMP*	71.9±4.2	80.3±4.1	73.3±8.1	65.4±3.0	75.1±3.9	34.2±4.1
FedSGD	*Fed-MOTIF*	39.1±9.8	38.6±13.4	72.1±3.8	39.7±5.4	45.5±5.1	45.5±11.3
FedSGD	*Fed-IMP*	60.6±3.6	64.7±5.2	66.5±2.7	46.7±8.8	53.2±9.6	32.7±6.9
All seen	*MOTIF*	78.5±1.1	83.7±1.6	72.6±3.7	**70.0±5.7**	**78.2±3.5**	37.7±6.4
All seen	*IMP*	**81.0±5.7**	**86.9±4.3**	67.6±4.2	69.8±3.6	77.1±2.0	28.0±4.8

Table 11. Results for the **Predicate Classification** (left) and **Scene Graph Generation** (right) tasks for Centralized and Federated Learning on the **PC** dataset). All configurations are run 5 times using random seeds.

		BHSD Dataset					
		Predicate Classification			**Scene Graph Generation**		
Method	Model	R@8↑	mR@8↑	mAP@8↑	R@8↑	mR@8↑	mAP@8↑
BHSD dataset	*MOTIF*	67.9±2.5	58.1±4.2	36.6±2.3	38.3±5.9	37.3±8.4	17.7±3.5
BHSD dataset	*IMP*	61.7±6.3	52.2±6.0	34.9±5.9	38.8±3.7	35.2±4.9	6.9±2.1
Avg. unseen	*MOTIF*	51.1±17.7	41.0±16.3	41.8±10.7	31.6±14.9	25.9±14.0	15.7±9.7
Avg. unseen	*IMP*	51.7±15.8	40.5±12.9	39.1±9.5	31.6±15.3	30.1±15.5	9.7±6.8
FedAvg	*Fed-MOTIF*	68.0±5.9	54.4±5.9	47.8±5.7	**47.1±4.0**	39.8±2.6	**20.9±5.3**
FedAvg	*Fed-IMP*	69.8±6.9	53.9±4.6	48.1±5.0	46.5±3.0	43.4±5.2	14.8±4.1
FedSGD	*Fed-MOTIF*	32.4±10.2	20.9±7.4	**50.8±17.9**	29.0±5.6	19.7±5.1	17.4±2.4
FedSGD	*Fed-IMP*	37.4±7.2	29.8±3.0	34.5±7.4	26.7±6.6	29.2±8.9	10.9±3.1
All seen	*MOTIF*	**72.4±4.1**	**59.9±5.8**	42.4±5.3	**47.2±3.8**	39.8±6.1	17.1±3.7
All seen	*IMP*	66.2±5.1	54.4±5.3	47.2±5.7	49.8±3.3	**48.1±3.7**	12.4±0.9

Table 12. Results for the **Predicate Classification** (left) and **Scene Graph Generation** (right) tasks for Centralized and Federated Learning on the **BHSD** dataset). All configurations are run 5 times using random seeds.

		CQ500 Dataset					
		Predicate Classification			**Scene Graph Generation**		
Method	Model	R@8↑	mR@8↑	mAP@8↑	R@8↑	mR@8↑	mAP@8↑
CQ500 dataset	*MOTIF*	51.5±7.7	55.0±7.3	59.0±5.3	34.8±3.4	38.6±3.9	31.0±5.4
CQ500 dataset	*IMP*	53.0±7.5	53.9±7.8	52.3±7.1	38.5±2.7	37.9±5.0	20.0±7.1
Avg. unseen	*MOTIF*	42.0±14.3	45.5±15.9	51.0±9.2	30.1±15.5	32.6±16.7	24.2±12.6
Avg. unseen	*IMP*	44.7±12.3	47.1±14.0	41.8±10.6	34.4±18.0	35.3±18.2	16.7±8.6
FedAvg	*Fed-MOTIF*	54.4±3.2	59.3±3.8	**60.3±4.0**	49.7±7.8	54.6±7.5	**36.9±4.6**
FedAvg	*Fed-IMP*	**60.5±1.1**	**66.0±2.1**	51.4±4.5	**59.4±3.9**	**61.9±2.4**	25.8±2.4
FedSGD	*Fed-MOTIF*	22.1±8.8	21.7±8.6	54.0±11.4	23.8±2.8	25.3±2.7	34.8±7.4
FedSGD	*Fed-IMP*	32.9±4.9	34.3±4.9	46.5±8.1	29.5±7.4	28.8±6.6	21.4±5.6
All seen	*MOTIF*	59.1±1.9	62.8±2.0	54.3±7.5	46.7±4.7	49.1±4.3	27.7±5.9
All seen	*IMP*	53.2±1.5	55.6±1.8	44.7±4.4	52.2±3.7	52.1±4.1	18.5±2.4

Table 13. Results for the **Predicate Classification** (left) and **Scene Graph Generation** (right) tasks for Centralized and Federated Learning on the **CQ500** dataset). All configurations are run 5 times using random seeds.

		INST Dataset					
		Predicate Classification			**Scene Graph Generation**		
Method	Model	R@8↑	mR@8↑	mAP@8↑	R@8↑	mR@8↑	mAP@8↑
INST dataset	*MOTIF*	62.9±5.0	65.1±5.7	57.2±6.8	59.7±1.4	61.5±2.4	55.5±6.2
INST dataset	*IMP*	61.5±3.2	65.0±2.8	53.9±3.4	57.2±6.5	60.6±5.3	46.7±6.6
Avg. unseen	*MOTIF*	63.6±16.3	65.0±15.5	62.8±8.8	63.0±11.5	62.1±9.8	50.8±15.3
Avg. unseen	*IMP*	60.9±11.6	63.0±10.7	54.3±11.2	57.7±9.5	60.3±8.3	46.2±11.2
FedAvg	*Fed-MOTIF*	**76.9±4.6**	**77.8±4.9**	**70.5±2.0**	**75.1±1.2**	**71.8±1.5**	60.0±5.8
FedAvg	*Fed-IMP*	73.1±3.1	73.8±3.3	65.5±5.3	70.6±3.1	71.3±2.6	56.8±2.8
FedSGD	*Fed-MOTIF*	43.3±9.6	42.7±10.5	69.2±8.1	59.9±1.9	57.4±2.1	**63.3±7.0**
FedSGD	*Fed-IMP*	54.5±1.8	55.3±2.2	50.6±8.0	54.9±5.0	58.0±3.9	48.0±5.8
All seen	*MOTIF*	74.3±1.5	76.1±2.4	60.2±4.4	71.3±2.6	**71.5±2.7**	54.5±2.9
All seen	*IMP*	66.4±3.6	69.2±2.9	52.0±2.6	61.3±5.0	64.2±3.9	46.8±5.5

Table 14. Results for the **Predicate Classification** (left) and **Scene Graph Generation** (right) tasks for Centralized and Federated Learning on the **INST** dataset). All configurations are run 5 times using random seeds.

		PC Dataset					
		Predicate Classification			**Scene Graph Generation**		
Method	Model	R@8↑	mR@8↑	mAP@8↑	R@8↑	mR@8↑	mAP@8↑
PC dataset	*MOTIF*	76.6±6.5	80.6±4.7	78.3±2.5	61.0±3.1	66.9±2.4	66.6±5.2
PC dataset	*IMP*	63.8±7.3	66.1±9.0	73.8±5.0	58.7±5.2	59.7±4.3	48.1±10.1
Avg. unseen	*MOTIF*	67.9±14.3	70.7±16.2	74.8±4.7	60.2±8.7	69.9±8.0	63.1±8.5
Avg. unseen	*IMP*	65.5±8.5	70.4±8.2	65.5±10.8	59.6±6.2	66.1±4.3	55.1±10.5
FedAvg	*Fed-MOTIF*	72.8±5.6	78.6±5.9	**76.7±5.2**	63.5±6.4	72.6±5.8	**66.0±3.8**
FedAvg	*Fed-IMP*	71.9±4.2	80.3±4.1	73.3±8.1	58.6±2.0	68.8±1.1	73.2±2.1
FedSGD	*Fed-MOTIF*	39.1±9.8	38.6±13.4	72.1±3.8	52.4±5.6	59.9±6.1	65.2±6.2
FedSGD	*Fed-IMP*	60.6±3.6	64.7±5.2	66.5±2.7	58.7±7.3	66.9±7.2	48.2±5.0
All seen	*MOTIF*	**78.5±1.1**	**83.7±1.6**	72.6±3.7	**66.2±1.9**	**77.1±2.8**	**65.8±2.8**
All seen	*IMP*	64.7±10.4	65.4±7.8	54.7±6.0	58.0±7.4	63.3±3.9	53.5±5.5

Table 15. Results for the **Predicate Classification** (left) and **Scene Graph Generation** (right) tasks for Centralized and Federated Learning on the **PC** dataset). All configurations are run 5 times using random seeds.

		BHSD Dataset					
		Predicate Classification			**Scene Graph Generation**		
Method	Model	R@8↑	mR@8↑	mAP@8↑	R@8↑	mR@8↑	mAP@8↑
BHSD dataset	*MOTIF*	67.9±2.5	58.1±4.2	36.6±2.3	47.8±4.8	44.8±6.5	26.5±3.0
BHSD dataset	*IMP*	61.7±6.3	52.2±6.0	34.9±5.9	37.2±3.4	28.7±5.6	21.9±3.8
Avg. unseen	*MOTIF*	51.1±17.7	41.0±16.3	41.8±10.7	47.9±15.8	37.4±12.4	26.0±9.1
Avg. unseen	*IMP*	49.6±14.9	38.5±11.5	37.4±8.9	41.4±10.6	30.9±7.1	23.7±9.9
FedAvg	*Fed-MOTIF*	68.0±5.9	54.4±5.9	47.8±5.7	61.0±4.7	47.4±4.3	29.0±2.4
FedAvg	*Fed-IMP*	69.8±6.9	53.9±4.6	48.1±5.0	49.7±5.8	33.3±4.0	34.0±5.6
FedSGD	*Fed-MOTIF*	32.4±10.2	20.9±7.4	**50.8±17.9**	39.0±9.9	29.1±5.4	34.2±6.2
FedSGD	*Fed-IMP*	37.4±7.2	29.8±3.0	34.5±7.4	39.0±10.8	35.4±10.3	27.5±6.2
All seen	*MOTIF*	**72.4±4.1**	**59.9±5.8**	42.4±5.3	**64.1±4.8**	**48.9±5.4**	**29.5±3.6**
All seen	*IMP*	53.5±10.1	42.2±6.8	36.7±6.2	45.7±7.9	31.4±6.9	21.5±3.0

Table 16. Results for the **Predicate Classification** (left) and **Scene Graph Generation** (right) tasks for Centralized and Federated Learning on the **BHSD** dataset). All configurations are run 5 times using random seeds.

		CQ500 Dataset					
		Predicate Classification			**Scene Graph Generation**		
Method	Model	R@8↑	mR@8↑	mAP@8↑	R@8↑	mR@8↑	mAP@8↑
CQ500 dataset	*MOTIF*	51.5±7.7	55.0±7.3	59.0±5.3	48.2±8.6	51.9±8.2	44.2±5.4
CQ500 dataset	*IMP*	53.0±7.5	53.9±7.8	52.3±7.1	36.4±5.1	36.9±6.5	37.2±11.5
Avg. unseen	*MOTIF*	42.0±14.3	45.5±15.9	51.0±9.2	42.0±12.2	46.1±12.9	38.4±15.2
Avg. unseen	*IMP*	42.2±12.2	44.4±14.0	40.6±10.8	37.1±9.1	39.0±9.1	34.7±15.3
FedAvg	*Fed-MOTIF*	54.4±3.2	59.3±3.8	**60.3±4.0**	55.5±6.0	62.0±5.6	**53.9±1.8**
FedAvg	*Fed-IMP*	**60.5±1.1**	**66.0±2.1**	51.4±4.5	52.6±2.2	55.1±2.3	59.4±3.4
FedSGD	*Fed-MOTIF*	22.1±8.8	21.7±8.6	54.0±11.4	30.8±3.0	33.7±4.5	40.2±3.5
FedSGD	*Fed-IMP*	32.9±4.9	34.3±4.9	46.5±8.1	34.9±4.3	36.6±5.3	40.0±2.7
All seen	*MOTIF*	59.1±1.9	62.8±2.0	54.3±7.5	**59.5±3.0**	**62.9±2.3**	47.5±3.1
All seen	*IMP*	38.1±7.3	39.0±7.4	37.6±6.5	38.3±2.6	38.8±4.2	31.4±4.1

Table 17. Results for the **Predicate Classification** (left) and **Scene Graph Generation** (right) tasks for Centralized and Federated Learning on the **CQ500** dataset). All configurations are run 5 times using random seeds.

Bias	Method	Model	INST Dataset					
			Predicate Classification			Scene Graph Generation		
			R@8↑	mR@8↑	mAP@8↑	R@8↑	mR@8↑	mAP@8↑
All datasets	INST dataset	*MOTIF*	61.8±5.9	65.2±5.5	58.1±4.1	47.6±6.3	49.2±6.2	34.1±6.5
	INST dataset	*IMP*	61.9±5.2	66.2±5.3	49.6±6.3	53.1±4.6	54.4±5.4	22.4±3.2
	Avg. unseen	*MOTIF*	66.2±13.1	68.0±11.0	59.5±6.9	49.9±16.7	50.2±15.5	38.8±14.5
	Avg. unseen	*IMP*	60.4±13.1	62.7±11.8	51.8±11.0	54.2±16.3	55.7±16.1	20.0±10.1
Disabled	*FedAvg*	*Fed-MOTIF*	76.0±3.5	**76.4±3.3**	**64.8±0.8**	65.0±3.9	63.3±3.7	**49.6±4.7**
	FedAvg	*Fed-IMP*	73.8±5.5	74.0±4.3	62.7±6.3	**67.6±5.1**	67.4±5.2	20.8±1.4
	FedSGD	*Fed-MOTIF*	57.4±14.3	59.2±12.6	53.4±11.3	53.5±5.1	52.9±4.7	48.5±6.2
	FedSGD	*Fed-IMP*	55.0±6.3	55.9±5.6	55.8±7.3	50.1±5.8	54.3±4.9	10.1±1.5
	All seen	*MOTIF*	**74.7±4.3**	75.8±5.4	57.9±3.0	56.8±6.1	58.1±4.9	40.3±1.3
	All seen	*IMP*	72.5±5.1	75.3±3.5	35.6±4.7	**67.4±2.2**	**68.4±2.8**	17.1±1.6

Table 18. **Ablation study**: effect of frequency bias layers (results on the **INST** dataset). All configurations are run 5 times using random seeds.

Bias	Method	Model	PC Dataset					
			Predicate Classification			Scene Graph Generation		
			R@8↑	mR@8↑	mAP@8↑	R@8↑	mR@8↑	mAP@8↑
All datasets	PC dataset	*MOTIF*	72.4±9.1	76.9±7.0	77.6±5.4	56.8±3.6	62.5±5.0	53.2±7.1
	PC dataset	*IMP*	67.4±3.0	68.5±5.4	74.3±8.8	64.8±7.6	67.8±8.7	39.6±10.9
	Avg. unseen	*MOTIF*	72.0±7.3	76.1±7.3	72.0±7.4	56.7±8.0	63.8±9.2	38.9±10.5
	Avg. unseen	*IMP*	65.2±9.6	71.1±8.4	58.5±13.0	55.7±14.7	64.1±14.8	24.4±7.7
Disabled	*FedAvg*	*Fed-MOTIF*	75.6±1.3	80.7±2.5	**74.1±2.9**	64.1±2.9	71.6±3.1	**50.2±3.1**
	FedAvg	*Fed-IMP*	65.3±6.1	75.4±4.0	60.8±5.1	**73.8±2.3**	78.9±2.3	30.1±2.6
	FedSGD	*Fed-MOTIF*	61.4±6.1	65.7±6.5	68.2±7.3	56.6±7.1	60.7±5.5	44.1±6.3
	FedSGD	*Fed-IMP*	57.5±5.4	64.1±6.9	55.7±5.7	35.1±7.0	42.4±8.4	15.1±7.3
	All seen	*MOTIF*	**81.9±3.4**	**85.3±2.7**	70.5±10.7	65.8±1.3	75.6±0.6	34.5±6.2
	All seen	*IMP*	71.4±5.6	77.5±3.2	37.7±3.0	**73.9±2.5**	**81.9±2.6**	22.9±2.4

Table 19. **Ablation study**: effect of frequency bias layers (results on the **PC** dataset). All configurations are run 5 times using random seeds.

			BHSD Dataset					
			Predicate Classification			Scene Graph Generation		
Bias	Method	Model	R@8↑	mR@8↑	mAP@8↑	R@8↑	mR@8↑	mAP@8↑
All datasets	BHSD dataset	*MOTIF*	72.1±4.7	59.9±4.6	39.1±2.0	38.6±4.5	36.9±6.5	17.5±4.9
	BHSD dataset	*IMP*	64.5±6.2	54.6±5.6	37.5±4.7	42.6±3.3	41.4±6.8	14.1±2.9
	Avg. unseen	*MOTIF*	54.1±17.7	42.6±15.4	37.9±6.5	32.5±14.4	26.4±13.3	15.8±10.2
	Avg. unseen	*IMP*	49.7±17.8	40.0±14.3	35.7±7.1	34.1±16.9	35.6±18.0	8.3±4.2
Disabled	*FedAvg*	*Fed-MOTIF*	70.7±4.0	55.1±3.6	**44.4±5.4**	43.6±4.7	35.1±5.2	18.7±4.2
	FedAvg	*Fed-IMP*	64.8±7.6	49.7±5.1	43.4±2.7	51.8±4.1	49.7±3.3	11.9±2.6
	FedSGD	*Fed-MOTIF*	38.9±10.6	29.3±9.0	36.4±3.8	39.6±5.9	26.3±5.1	14.4±7.5
	FedSGD	*Fed-IMP*	34.0±8.3	33.4±5.9	35.2±4.9	30.0±5.8	43.7±6.9	4.0±1.8
	All seen	*MOTIF*	**74.6±5.3**	59.6±7.7	39.0±4.3	46.7±4.1	40.1±3.8	**23.0±4.0**
	All seen	*IMP*	69.9±5.8	**60.4±5.0**	31.8±4.2	**54.5±4.1**	**54.5±5.2**	9.7±1.1

Table 20. **Ablation study**: effect of frequency bias layers (results on the **BHSD** dataset). All configurations are run 5 times using random seeds.

			CQ500 Dataset					
			Predicate Classification			Scene Graph Generation		
Bias	Method	Model	R@8↑	mR@8↑	mAP@8↑	R@8↑	mR@8↑	mAP@8↑
All datasets	CQ500 dataset	*MOTIF*	52.9±7.7	56.9±6.9	59.3±5.1	39.0±3.7	42.2±4.0	29.8±5.7
	CQ500 dataset	*IMP*	59.9±4.2	61.9±5.7	51.9±9.8	41.1±3.6	42.2±3.1	21.3±6.8
	Avg. unseen	*MOTIF*	42.5±14.0	46.4±14.8	51.0±10.9	32.0±15.5	34.6±16.8	23.4±12.0
	Avg. unseen	*IMP*	43.5±13.3	46.8±13.7	39.6±10.2	39.8±19.8	41.0±20.1	12.6±7.8
Disabled	*FedAvg*	*Fed-MOTIF*	56.1±3.7	60.7±3.4	**61.5±3.0**	52.3±2.1	57.2±2.9	**37.2±2.3**
	FedAvg	*Fed-IMP*	57.5±5.5	60.1±5.9	50.9±2.7	**63.1±4.7**	**65.7±3.6**	22.5±4.4
	FedSGD	*Fed-MOTIF*	29.6±10.6	31.6±10.3	55.5±12.6	32.9±3.5	32.2±4.8	29.2±9.8
	FedSGD	*Fed-IMP*	31.4±4.2	35.4±3.9	44.0±4.7	46.0±7.6	46.1±7.3	8.3±2.6
	All seen	*MOTIF*	**59.0±3.6**	**62.5±4.4**	49.4±9.6	46.5±5.2	49.7±4.8	26.6±2.6
	All seen	*IMP*	57.7±5.2	61.4±6.5	30.5±5.6	57.1±4.3	56.2±4.6	13.3±1.6

Table 21. **Ablation study**: effect of frequency bias layers (results on the **CQ500** dataset). All configurations are run 5 times using random seeds.