MAXORA.AI Microsoft

NEURAL INFORMATION PROCESSING SYSTEMS

Meta-DiffuB: A Contextualized Sequence-to-Sequence Text Diffusion Model with Meta-Exploration

Neurips 2024


Yun-Yen Chuang[1,2]


Hung-Min Hsu[3]


Kevin Lin[4]


Chen-Sheng Gu[1,2]


Ling Zhen Li[1,2]


Ray-I Chang[2]


Hung-Yi Lee[2]

Introduction

Diffusion Models:

- A generative framework successful in image, audio, video, and text generation

Introduction

Diffusion Models:

- A generative framework successful in image, audio, video, and text generation
- Works by adding (Foward process) and removing (Reverse process) noise to generate data

Introduction

Diffusion Models:

- A generative framework successful in image, audio, video, and text generation
- Works by adding (Foward process) and removing (Reverse process) noise to generate data
- Seq2Seq Text Generation (S2S-Diffusion models): Generate target sentence from an input sentence

Introduction

Diffusion Models:

- A generative framework successful in image, audio, video, and text generation
- Works by adding (Foward process) and removing (Reverse process) noise to generate data
- Seq2Seq Text Generation (S2S-Diffusion models): Generate target sentence from an input sentence
- **Current Limitations** (Non-contextualized noise scheduling):
    - Fixed noise scheduling in existing models (e.g., DiffuSeq, Dinoiser, SeqDiffuSeq, LD4LG, RDM)
    - Fails to adapt to sentence semantic

Introduction

Diffusion Models:

- A generative framework successful in image, audio, video, and text generation
- Works by adding (Foward process) and removing (Reverse process) noise to generate data
- Seq2Seq Text Generation (S2S-Diffusion models): Generate target sentence from an input sentence
- **Current Limitations** (Non-contextualized noise scheduling):
  - Fixed noise scheduling in existing models (e.g., DiffuSeq, Dinoiser, SeqDiffuSeq, LD4LG, RDM)
  - Fails to adapt to sentence semantic

- **Key Contributions:**
  - Introduce Meta-DiffuB to dynamically schedule noise

Introduction

Diffusion Models:

- A generative framework successful in image, audio, video, and text generation
- Works by adding (Foward process) and removing (Reverse process) noise to generate data
- Seq2Seq Text Generation (S2S-Diffusion models): Generate target sentence from an input sentence
- **Current Limitations** (Non-contextualized noise scheduling):
  - Fixed noise scheduling in existing models (e.g., DiffuSeq, Dinoiser, SeqDiffuSeq, LD4LG, RDM)
  - Fails to adapt to sentence semantic

- **Key Contributions:**
  - Introduce Meta-DiffuB to dynamically schedule noise
  - Contextualized noise scheduling for better sentence generation

Introduction

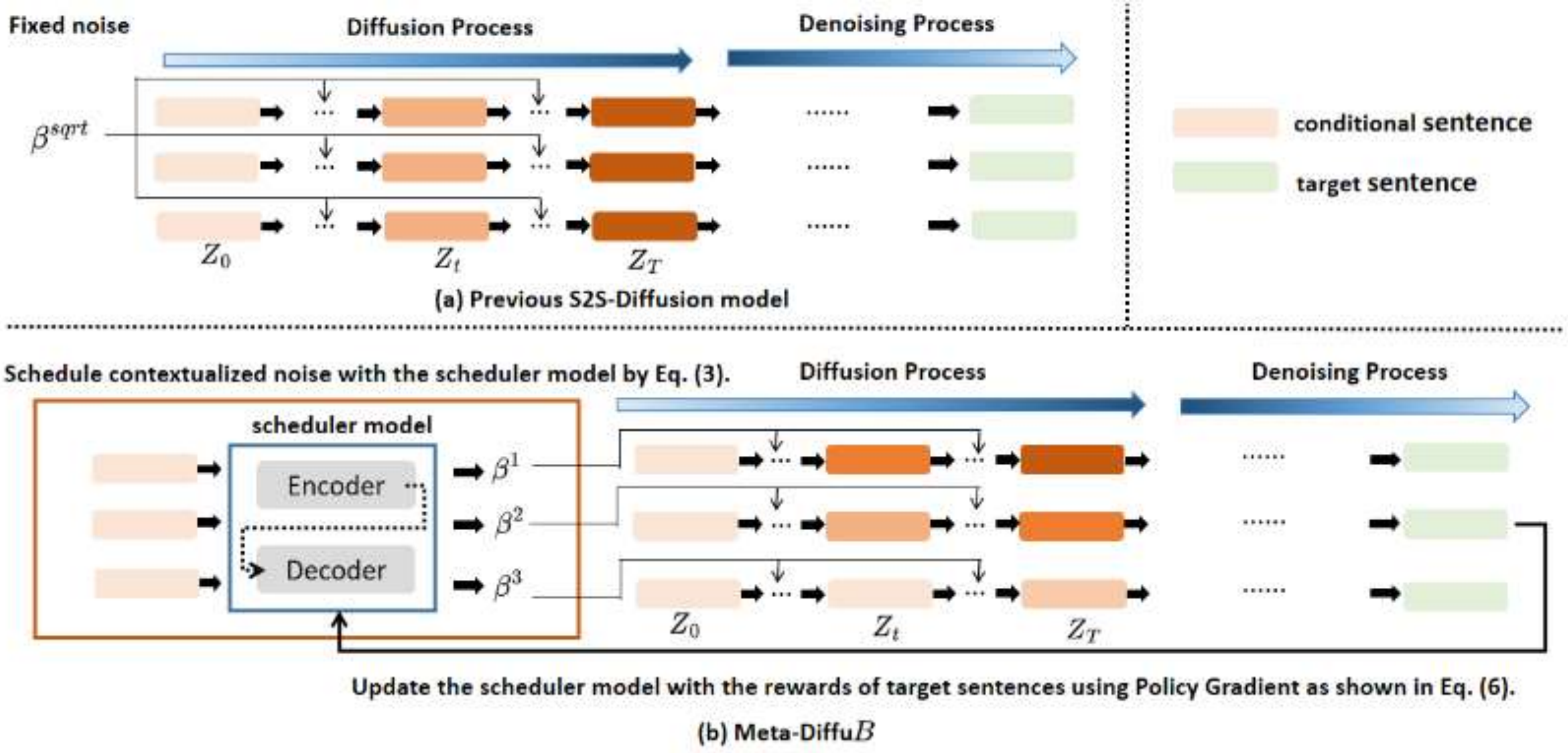
Diffusion Models:

- A generative framework successful in image, audio, video, and text generation
- Works by adding (Foward process) and removing (Reverse process) noise to generate data
- Seq2Seq Text Generation (S2S-Diffusion models): Generate target sentence from an input sentence
- **Current Limitations** (Non-contextualized noise scheduling):
  - Fixed noise scheduling in existing models (e.g., DiffuSeq, Dinoiser, SeqDiffuSeq, LD4LG, RDM)
  - Fails to adapt to sentence semantic

- **Key Contributions:**
  - Introduce Meta-DiffuB to dynamically schedule noise
  - Contextualized noise scheduling for better sentence generation
  - Plug-and-play scheduler model: Enhances existing models without fine-tuning

Meta-DiffuB: Framework

Scheduler-Exploiter Framework:

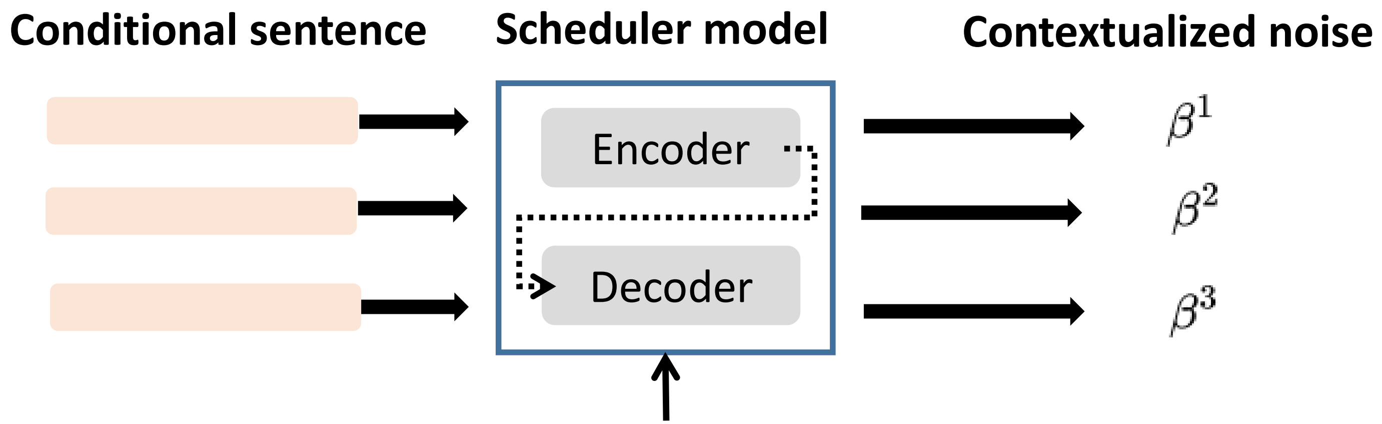


Methodology

- **Noise Scheduling (Scheduler Model):**
  - Generates Meta-Instructions for noise adjustment

$$\iota^x = B_\psi(\mathbf{w}^x)$$
$$\beta^x = skipping(\iota^x, \beta^{sqrt}).$$

  - Customizes noise schedule based on sentence complexity
  - Scheduler adapts noise levels based on exploiter's learning by Policy Gradient

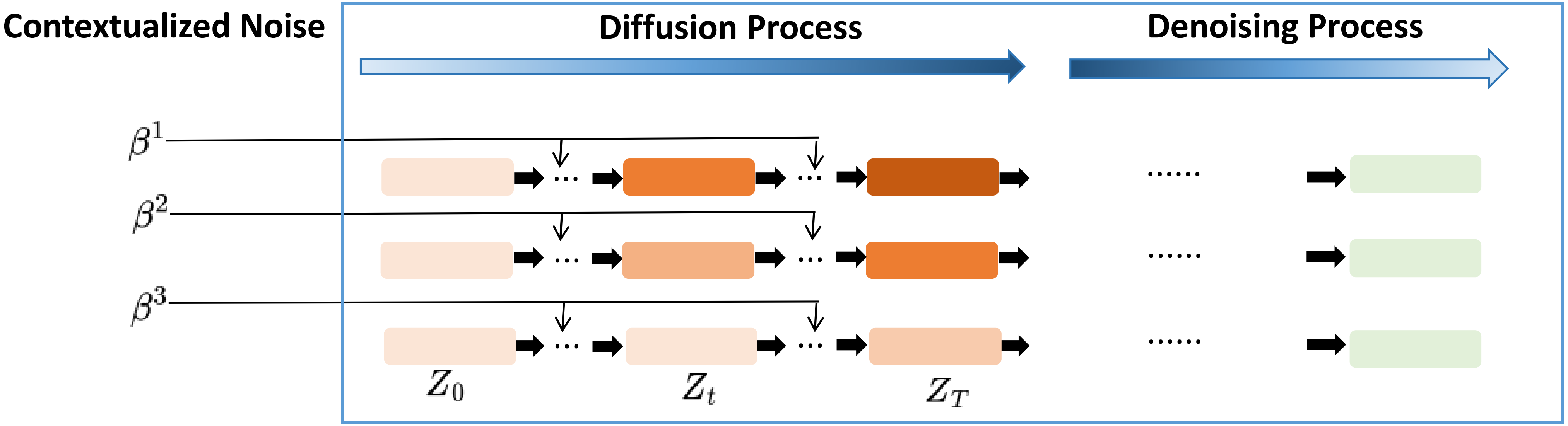


Update with the rewards of target sentences using Policy Gradient as shown in Eq. (6).

NEURAL INFORMATION PROCESSING SYSTEMS

Methodology

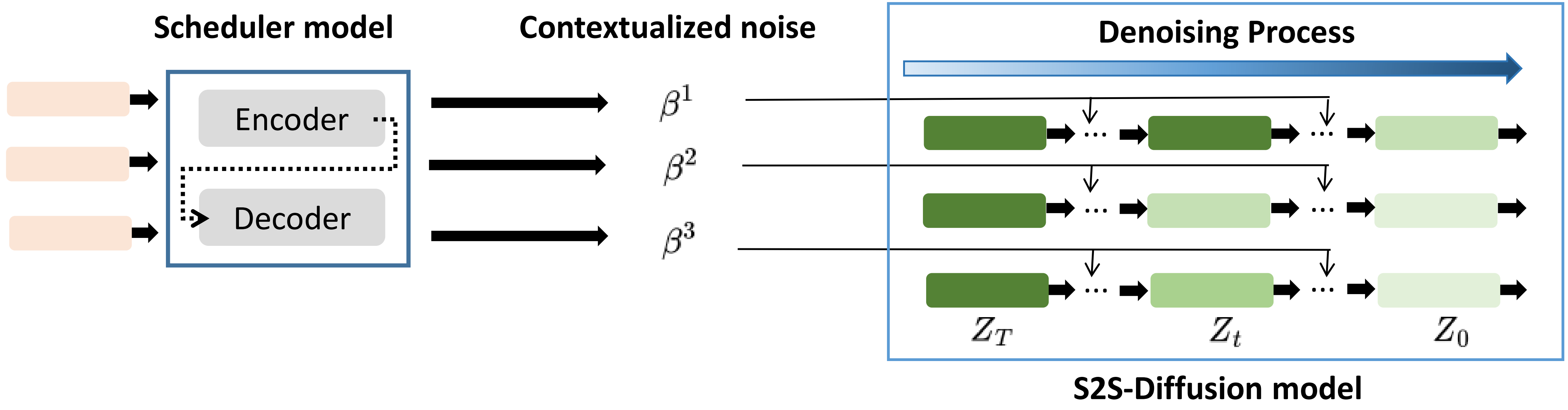
- **Training the Exploiter (S2S-Diffusion model):**
  - S2S-Diffusion model uses scheduled noise for generation



S2S-Diffusion model

Methodology

- **Contextualized Inference:**
  - Dynamic noise scheduling during inference for better text generation
  - Adapts to sentence difficulty, improving quality and diversity



Experiments & Results

NEURAL INFORMATION PROCESSING SYSTEMS

Benchmark dataset experiment

Tasks	Methods	BLEU (↑)	ROUGH-L (↑)	BERTScore (↑)	Dist-1 (↑)	Self-BLEU (↓)	M-R (↓)
QQP	⋆ GPT2-base	0.1980	0.5212	0.8246	0.9798	0.5480	5.20
	⋆ GPT2-large	0.2059	0.5415	0.8363	0.9819	0.7325	3.80
	⋆ LevT	0.2268	0.5795	0.8344	0.9790	0.9995	4.80
	⋆ DiffuSeq	0.2413	0.5880	0.8365	0.9807	0.2732	2.60
	⋆ SeqDiffuSeq	0.2434	-	0.8400	0.9807	-	2.33
	† Dinoiser	0.1949	0.5316	0.8036	0.9723	0.8643	6.20
	Meta-Diffu$B$	**0.2632**	**0.5933**	**0.8519**	**0.9902**	**0.2595**	**1.00**
WA	⋆ GPT2-base	0.3083	0.5461	0.8021	0.9439	0.5444	3.40
	⋆ GPT2-large	0.2693	0.5111	0.7882	0.9464	0.6042	4.00
	⋆ LevT	0.2052	0.4402	0.7254	**0.9715**	0.9907	5.00
	⋆ DiffuSeq	0.3622	0.5849	0.8126	0.9264	0.4642	3.00
	⋆ SeqDiffuSeq	0.3712	-	0.8214	0.9077	-	3.33
	† Dinoiser	0.2388	0.4821	0.6787	0.8421	0.9132	6.20
	Meta-Diffu$B$	**0.3877**	**0.6047**	**0.8233**	0.9355	**0.3888**	**1.60**
QT	⋆ GPT2-base	0.0741	0.2714	0.6052	0.9602	**0.1403**	3.80
	⋆ GPT2-large	0.1110	0.3215	**0.6346**	**0.9670**	0.2910	2.60
	⋆ LevT	0.0930	0.2893	0.5491	0.8914	0.9830	5.40
	⋆ DiffuSeq	0.1731	0.3665	0.6123	0.9056	0.2789	3.20
	⋆ SeqDiffuSeq	0.1746	-	0.6174	0.9248	-	3.33
	† Dinoiser	0.0477	0.1872	0.4690	0.8191	0.5273	6.40
	Meta-Diffu$B$	**0.1820**	**0.3870**	0.6286	0.9323	0.2527	**1.80**
CC	⋆ GPT2-base	0.0108	0.1508	0.5279	0.9194	0.0182	4.00
	⋆ GPT2-large	0.0125	0.1002	0.5293	0.9244	0.0213	4.00
	⋆ LevT	0.0158	0.0550	0.4760	**0.9726**	0.7103	3.80
	⋆ DiffuSeq	0.0139	0.1056	0.5131	0.9467	0.0144	3.40
	⋆ SeqDiffuSeq	0.0112	-	0.4425	0.9608	-	2.80
	† Dinoiser	0.0096	0.1166	0.3545	0.2485	0.9994	6.00
	Meta-Diffu$B$	**0.0220**	**0.1528**	**0.5316**	0.9670	**0.0133**	**1.20**

Experiments & Results

Benchmark dataset experiment

Tasks	Methods	BLEU (↑)	ROUGH-L (↑)	BERTScore (↑)	Dist-1 (↑)	Self-BLEU (↓)	M-R (↓)
QQP	⋆ GPT2-base	0.1980	0.5212	0.8246	0.9798	0.5480	5.20
	⋆ GPT2-large	0.2059	0.5415	0.8363	0.9819	0.7325	3.80
	⋆ LevT	0.2268	0.5795	0.8344	0.9790	0.9995	4.80
	⋆ DiffuSeq	0.2413	0.5880	0.8365	0.9807	0.2732	2.60
	⋆ SeqDiffuSeq	0.2434	-	0.8400	0.9807	-	2.33
	† Dinoiser	0.1949	0.5316	0.8036	0.9723	0.8643	6.20
	Meta-Diffu$B$	**0.2632**	**0.5933**	**0.8519**	**0.9902**	**0.2595**	**1.00**
WA	⋆ GPT2-base	0.3083	0.5461	0.8021	0.9439	0.5444	3.40
	⋆ GPT2-large	0.2693	0.5111	0.7882	0.9464	0.6042	4.00
	⋆ LevT	0.2052	0.4402	0.7254	**0.9715**	0.9907	5.00
	⋆ DiffuSeq	0.3622	0.5849	0.8126	0.9264	0.4642	3.00
	⋆ SeqDiffuSeq	0.3712	-	0.8214	0.9077	-	3.33
	† Dinoiser	0.2388	0.4821	0.6787	0.8421	0.9132	6.20
	Meta-Diffu$B$	**0.3877**	**0.6047**	**0.8233**	0.9355	**0.3888**	**1.60**
QT	⋆ GPT2-base	0.0741	0.2714	0.6052	0.9602	**0.1403**	3.80
	⋆ GPT2-large	0.1110	0.3215	**0.6346**	**0.9670**	0.2910	2.60
	⋆ LevT	0.0930	0.2893	0.5491	0.8914	0.9830	5.40
	⋆ DiffuSeq	0.1731	0.3665	0.6123	0.9056	0.2789	3.20
	⋆ SeqDiffuSeq	0.1746	-	0.6174	0.9248	-	3.33
	† Dinoiser	0.0477	0.1872	0.4690	0.8191	0.5273	6.40
	Meta-Diffu$B$	**0.1820**	**0.3870**	0.6286	0.9323	0.2527	**1.80**
CC	⋆ GPT2-base	0.0108	0.1508	0.5279	0.9194	0.0182	4.00
	⋆ GPT2-large	0.0125	0.1002	0.5293	0.9244	0.0213	4.00
	⋆ LevT	0.0158	0.0550	0.4760	**0.9726**	0.7103	3.80
	⋆ DiffuSeq	0.0139	0.1056	0.5131	0.9467	0.0144	3.40
	⋆ SeqDiffuSeq	0.0112	-	0.4425	0.9608	-	2.80
	† Dinoiser	0.0096	0.1166	0.3545	0.2485	0.9994	6.00
	Meta-Diffu$B$	**0.0220**	**0.1528**	**0.5316**	0.9670	**0.0133**	**1.20**

Model agnostic experiment

Tasks	Methods	BLEU (↑)	BERTScore (↑)	Dist-1 (↑)
QQP	⋆ DiffuSeq	0.2413	0.8365	0.9807
	Meta-Diffu$B$ (D_θ = DiffuSeq)	**0.2552**	**0.8821**	**0.9922**
	⋆ SeqDiffuSeq	0.2434	0.8400	0.9807
	Meta-Diffu$B$ (D_θ = SeqDiffuSeq)	**0.2632**	**0.8919**	**0.9902**
	† Dinoiser	0.1949	0.8036	0.9723
	Meta-Diffu$B$ (D_θ = Dinoiser)	**0.2271**	**0.8525**	**0.9752**
WA	⋆ DiffuSeq	0.3622	0.8126	0.9264
	Meta-Diffu$B$ (D_θ = DiffuSeq)	**0.3877**	**0.8233**	**0.9355**
	⋆ SeqDiffuSeq	0.3712	0.8214	0.9077
	Meta-Diffu$B$ (D_θ = SeqDiffuSeq)	**0.3957**	**0.8451**	**0.9412**
	† Dinoiser	0.2388	0.6787	0.8421
	Meta-Diffu$B$ (D_θ = Dinoiser)	**0.2471**	**0.7285**	**0.8694**

Experiments & Results

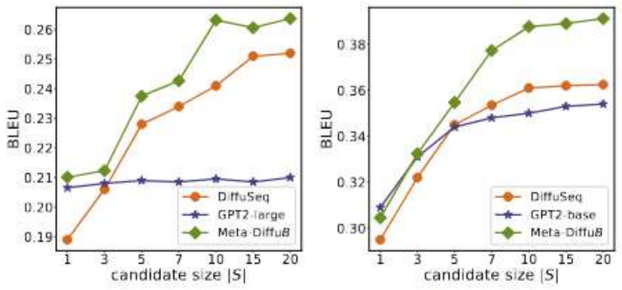
Benchmark dataset experiment

Tasks	Methods	BLEU (↑)	ROUGH-L (↑)	BERTScore (↑)	Dist-1 (↑)	Self-BLEU (↓)	M-R (↓)
QQP	⋆ GPT2-base	0.1980	0.5212	0.8246	0.9798	0.5480	5.20
	⋆ GPT2-large	0.2059	0.5415	0.8363	0.9819	0.7325	3.80
	⋆ LevT	0.2268	0.5795	0.8344	0.9790	0.9995	4.80
	⋆ DiffuSeq	0.2413	0.5880	0.8365	0.9807	0.2732	2.60
	⋆ SeqDiffuSeq	0.2434	-	0.8400	0.9807	-	2.33
	† Dinoiser	0.1949	0.5316	0.8036	0.9723	0.8643	6.20
	Meta-Diffu$B$	**0.2632**	**0.5933**	**0.8519**	**0.9902**	**0.2595**	**1.00**
WA	⋆ GPT2-base	0.3083	0.5461	0.8021	0.9439	0.5444	3.40
	⋆ GPT2-large	0.2693	0.5111	0.7882	0.9464	0.6042	4.00
	⋆ LevT	0.2052	0.4402	0.7254	**0.9715**	0.9907	5.00
	⋆ DiffuSeq	0.3622	0.5849	0.8126	0.9264	0.4642	3.00
	⋆ SeqDiffuSeq	0.3712	-	0.8214	0.9077	-	3.33
	† Dinoiser	0.2388	0.4821	0.6787	0.8421	0.9132	6.20
	Meta-Diffu$B$	**0.3877**	**0.6047**	**0.8233**	0.9355	**0.3888**	**1.60**
QT	⋆ GPT2-base	0.0741	0.2714	0.6052	0.9602	**0.1403**	3.80
	⋆ GPT2-large	0.1110	0.3215	**0.6346**	**0.9670**	0.2910	2.60
	⋆ LevT	0.0930	0.2893	0.5491	0.8914	0.9830	5.40
	⋆ DiffuSeq	0.1731	0.3665	0.6123	0.9056	0.2789	3.20
	⋆ SeqDiffuSeq	0.1746	-	0.6174	0.9248	-	3.33
	† Dinoiser	0.0477	0.1872	0.4690	0.8191	0.5273	6.40
	Meta-Diffu$B$	**0.1820**	**0.3870**	0.6286	0.9323	0.2527	**1.80**
CC	⋆ GPT2-base	0.0108	0.1508	0.5279	0.9194	0.0182	4.00
	⋆ GPT2-large	0.0125	0.1002	0.5293	0.9244	0.0213	4.00
	⋆ LevT	0.0158	0.0550	0.4760	**0.9726**	0.7103	3.80
	⋆ DiffuSeq	0.0139	0.1056	0.5131	0.9467	0.0144	3.40
	⋆ SeqDiffuSeq	0.0112	-	0.4425	0.9608	-	2.80
	† Dinoiser	0.0096	0.1166	0.3545	0.2485	0.9994	6.00
	Meta-Diffu$B$	**0.0220**	**0.1528**	**0.5316**	0.9670	**0.0133**	**1.20**

Model agnostic experiment

Tasks	Methods	BLEU (↑)	BERTScore (↑)	Dist-1 (↑)
QQP	⋆ DiffuSeq	0.2413	0.8365	0.9807
	Meta-Diffu$B$ (D_θ = DiffuSeq)	**0.2552**	**0.8821**	**0.9922**
	⋆ SeqDiffuSeq	0.2434	0.8400	0.9807
	Meta-Diffu$B$ (D_θ = SeqDiffuSeq)	**0.2632**	**0.8919**	**0.9902**
	† Dinoiser	0.1949	0.8036	0.9723
	Meta-Diffu$B$ (D_θ = Dinoiser)	**0.2271**	**0.8525**	**0.9752**
WA	⋆ DiffuSeq	0.3622	0.8126	0.9264
	Meta-Diffu$B$ (D_θ = DiffuSeq)	**0.3877**	**0.8233**	**0.9355**
	⋆ SeqDiffuSeq	0.3712	0.8214	0.9077
	Meta-Diffu$B$ (D_θ = SeqDiffuSeq)	**0.3957**	**0.8451**	**0.9412**
	† Dinoiser	0.2388	0.6787	0.8421
	Meta-Diffu$B$ (D_θ = Dinoiser)	**0.2471**	**0.7285**	**0.8694**

MBR decoding experiment

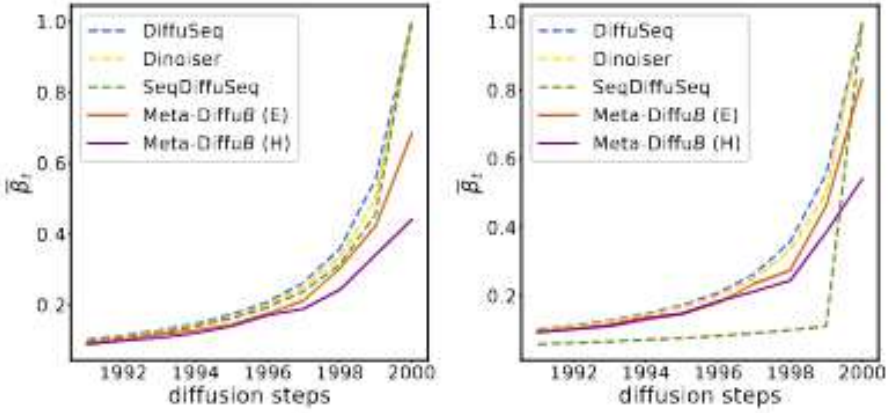


(a) QQP dataset (b) WA dataset

Figure 2: Increase in BLEU score with varying candidate sizes $|S|$ on the QQP and WA datasets.

Contextualized Noise Scheduling

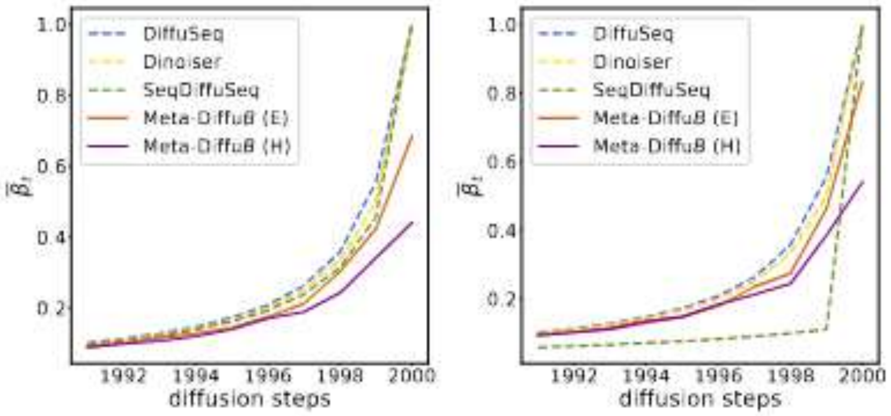
- **Noise adjustment based on sentence complexity:**
  - More noise for simpler sentences to enhance diversity
  - Less noise for complex sentences to improve accuracy
- **Significant improvement in generation performance:**
  - Noticeable enhancement in both quality and diversity for sentences of varying difficulty



(a) QQP dataset (b) WA dataset

Contextualized Noise Scheduling

- **Noise adjustment based on sentence complexity:**
  - More noise for simpler sentences to enhance diversity
  - Less noise for complex sentences to improve accuracy
- **Significant improvement in generation performance:**
  - Noticeable enhancement in both quality and diversity for sentences of varying difficulty



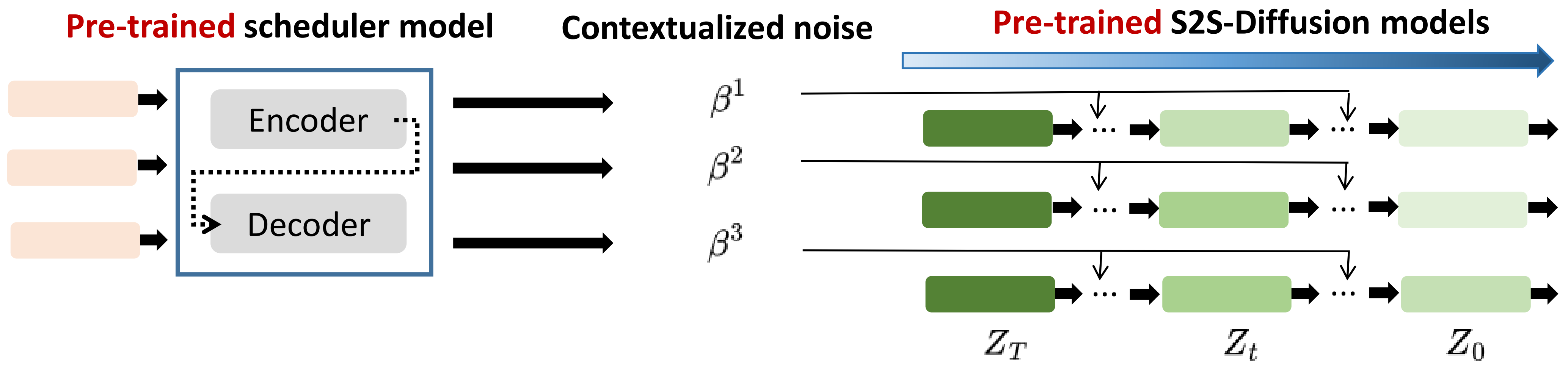
(a) QQP dataset (b) WA dataset

Methods	BLEU (↑)	Self-BLEU (↓)
DiffuSeq (E)	0.3721	0.4345
SeqDiffuSeq (E)	0.3752	0.4652
Dinoiser (E)	0.2892	0.8852
Meta-Diffu$B$ (E)	**0.3997**	**0.3688**
DiffuSeq (H)	0.3216	0.5085
SeqDiffuSeq (H)	0.3282	0.6251
Dinoiser (H)	0.2092	0.9528
Meta-Diffu$B$ (H)	**0.3724**	**0.4056**

Plug-and-Play experiment

- **Pre-trained scheduler integrated into other models:**
  - Applied Meta-DiffuB's pre-trained scheduler to DiffuSeq and other S2S-Diffusion models
  - No fine-tuning required
  - Improved performance without additional training

Scheduler	DiffuSeq	BLEU (↑)	ROUGH-L (↑)	BERTScore (↑)	Dist-1 (↑)	Self-BLEU (↓)
WA	QQP	**0.2594**	**0.5912**	**0.8459**	**0.9834**	**0.2653**
QT		**0.2603**	**0.5947**	**0.8503**	**0.9812**	**0.2649**
Null		0.2413	0.5880	0.8365	0.9807	0.2732



Conclusion & Applications

- **Conclusion & Applications**
  - Superior performance compared with previous S2S-Diffusion models in Seq2Seq tasks
  - Dynamic noise scheduling for better quality and diversity
  - Plug-and-play scheduler for easy integration

Conclusion & Applications

- **Conclusion & Applications**
  - Superior performance compared with previous S2S-Diffusion models in Seq2Seq tasks
  - Dynamic noise scheduling for better quality and diversity
  - Plug-and-play scheduler for easy integration
- **Applications:**
  - Language Translation
  - Text Summarization
  - Dialogue Systems

Conclusion & Applications

- **Conclusion & Applications**
  - Superior performance compared with previous S2S-Diffusion models in Seq2Seq tasks
  - Dynamic noise scheduling for better quality and diversity
  - Plug-and-play scheduler for easy integration
- **Applications:**
  - Language Translation
  - Text Summarization
  - Dialogue Systems
- **Future Impact:**
  - It is a valuable asset in the field of Seq2Seq Diffusion models.