Joint Contrastive Decoding Framework for Multilingual Machine Translation

Haozhe Liu
ETH Zurich
haozliu@ethz.ch

Tu Nguyen
ETH Zurich
tunguyen1@ethz.ch

Jules Schneuwly
ETH Zurich
jschneuwly@ethz.ch

Abstract

Neural machine translation often produces fluent but unfaithful outputs (hallucinations) and can drift into the wrong language, especially for low-resource pairs. Prior work tackles these issues separately via source, language, or model contrastive decoding. We introduce Joint Contrastive Decoding (Joint CD), which combines all three signals—perturbed sources, off-target language prompts, and a weaker student model—into a single inference step. By dynamically weighting these penalties based on model uncertainty, Joint CD selects tokens that are both reliable and faithful. On the FLORES-101 benchmark with M2M-small, our method improves chrF2 and spBLEU, reduces hallucinations, and cuts off-target outputs without extra computational cost.

1 Introduction

Despite their fluency, neural translators can hallucinate content or switch languages unexpectedly. *Source-contrastive decoding* penalizes tokens likely under a perturbed source, curbing hallucinations. *Language-contrastive decoding* penalizes tokens favored when the language prompt is swapped, preventing drift. *Model-contrastive decoding* filters tokens where a strong model outperforms a weaker variant.

These methods address different errors but don't fully complement each other when used alone. *Joint CD* unifies all three: at each step, it prunes to high-confidence tokens, applies penalties from multiple noisy inputs, and balances them via an uncertainty-driven weight. This yields translations that are both accurate and on-target. We demonstrate consistent gains on FLORES-101 with M2M-small, boosting quality metrics and reducing error modes, all at comparable inference speed.

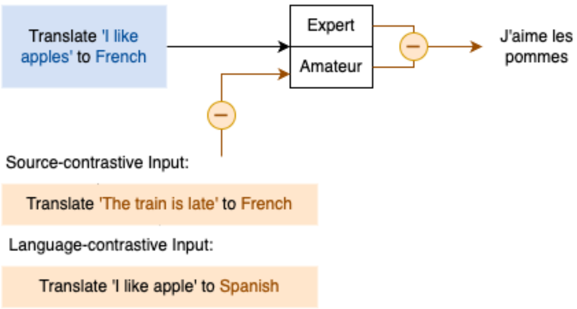

Translate 'I like apples' to French
Expert
Amateur
−
J'aime les pommes
−
Source-contrastive Input:
Translate 'The train is late' to French
Language-contrastive Input:
Translate 'I like apple' to Spanish

Figure 1: Joint Contrastive Decoding Framework

2 Methodology

In this work, we develop a unified contrastive decoding framework that mitigates hallucinations and off-target translations by combining ideas (decoding strategies) from both *source-contrastive* and *language-contrastive decoding*, alongside *amateur vs. expert model ensembling*. The structure of our framework is illustrated in Figure 1.

2.1 Joint Contrastive Decoding

The main idea of our project is to manipulate the generation by contrasting candidate tokens across multiple noisy variants of the input. The noisy variants include:

- **Source-contrastive inputs**: obtained by selecting a random input
- **Language-contrastive inputs**: We use the exact same strategy as in the paper (Sennrich et al., 2024). Let l_y be the target language. We replace its indicator token with contrastive variants $l_{y'} \in L_C$ for languages we wish to suppress. Based on the predominant off-target languages in multilingual MT, the set of contrastive languages L_c consists of English and the respective source language.

- **Model-contrastive inputs**: generated by using a student model, which intuitively might produce a lower quality translation compared to the main model. The choice of the student model can be varied, for example, using the output from an intermediate decoder layer of the main model, or using a completely different (smaller) model.

The source and language contrastive inputs are combined with the model contrastive inputs using a scoring function that blends two components: the **expert** and the **amateur**. The token-level score is defined as:

$$\log \text{score}_{\text{EXP}}(i) - \gamma \log \text{score}_{\text{AMA}}(i) \qquad (1)$$

where $\text{score}_{\text{EXP}}(i)$ is the log-probability from the expert model and $\text{score}_{\text{AMA}}(i)$ is computed from contrastive variants. The weight γ is determined dynamically:

$$\gamma = 1 - \max_i p_x(i)^\beta \qquad (2)$$

This formulation ensures that contrastive influence increases when the expert is uncertain about the prediction.

2.2 Vocabulary Pruning

To make contrastive scoring efficient, we restrict computation to a filtered vocabulary subset $\mathcal{V}_{\text{thresh}}$, defined as:

$$\mathcal{V}_{\text{thresh}} = \\ \{i \in \mathcal{V} : \log(p_x(i)) \geq \log(\alpha) + \max_j \log(p_x(j))\}$$

This pruning avoids unnecessary scoring of implausible candidates.

2.3 Contrastive Scoring Function

For each candidate token i, we compute a contrastive score using:

$$\text{score}_x(i) = p_x(i) - \lambda p_z(i \mid X') - \sum_{l_f \in \mathcal{L}_c} \theta p_x(i \mid l_f) \qquad (3)$$

where:

- $p_x(i)$ is the expert model's token distribution,
- $p_z(i \mid X')$ is the distribution under a perturbed (source-contrastive) input,
- $p_x(i \mid l_f)$ are distributions from language-contrastive variants,
- λ and θ are weighting coefficients for source and language contrastive penalties.

This final score encourages selection of tokens that are both high-confidence and robust across contrastive conditions.

3 Experiments

3.1 Models and Dataset

We use the M2M (Fan et al., 2020) models and evaluate the model performance on the FLORES-101 (Goyal et al., 2022) dataset according to the setup of (Sennrich et al., 2024) and (Waldendorf et al., 2024). The M2M models are multilingual translation transformer (Vaswani et al., 2017) with the encoder-decoder architecture trained on 7.5N sentences, which can still generate hallucination in low and medium resource languages, and thus can serve as a baseline model for various decoding algorithms evaluation. We evaluate the model on the Small (330M) M2M model, which requires less GPU memory than medium-sized M2M models and can produce more hallucinations. The language pairs we evaluate are referred to Table 6 in (Sennrich et al., 2024). For evaluation, we use the FLORES-101 dev split. All experiments are run using Hugging Face transformers (Wolf et al., 2020).

3.2 Metrics

Non-hallucination Filtering For dataset preprocessing, we use a combination of chrf2 (Popović, 2015), spBLEU (Goyal et al., 2022), target language identifier (Burchell et al., 2023), and TNG (top n-gram count) (Guerreiro et al., 2023) to filter out non-hallucinations in the baseline translations. chrf2 and spBLEU are reasonable quality metrics, which count the n-gram on a character level and subword level, respectively. Target language identifier uses OpenLID model for language identification and off-target language translation detection. TNG decides whether the top repeating n-gram in generation is more common than that in the gold translation by at least t and can detect oscillation hallucinations. The threshold of chrf2 and spBLEU

Experiments	chrf2				spBLEU			
	HLMT	X-branch	high-res	all	HLMT	X-branch	high-res	all
Direct	44.09	30.81	48.22	37.80	17.91	8.38	17.18	13.16
Source-language	44.97	33.52	48.88	39.53	18.40	9.60	18.78	14.06
Decoder-only	43.62	31.40	48.15	37.84	17.06	7.97	17.70	12.60
Attention Scaling	43.78	31.51	48.20	37.97	17.19	8.36	17.48	12.83
Joint Decoder-only	44.81	32.15	48.82	38.75	18.27	8.24	18.42	13.29
Joint Attention Scaling	44.57	31.78	48.80	38.50	18.23	9.07	18.30	13.70

Table 1: Automatic Evaluation Results

	EN	SRC	LOW
Direct	74	47	47
Source-language	29	17	17
Decoder-only	44	53	53
Attention Scaling	73	59	59
Joint Decoder-only	7	38	38
Joint Attention Scaling	26	42	42

Table 2: Number of off-target outputs, in English (EN), the source language (SRC), or the low-resource source language (LOW) from af, ast, hr, ps, ur, zu.

are 45.6 and 18.7 respectively. We use $n = 4$ and $t = 2$.

Evaluation Metrics: We report chrf2 and spBLEU scores for our automatice evaluation results. We split the language pairs into groups of non-English-centric directions (**HLMT**), low resource languages (**X-branch**), and high-resource languages (**high-res**) following (Sennrich et al., 2024) and evaluate the effect of Joint CD algorithms on each group. We also report the results for the entire set of language pairs.

3.3 Hyperparameters

We use beam search with a beam size of 5. We choose 2 contrastive source sentences and set English and the source language as contrastive languages. We tune the following hyperparameters: λ, θ, β and α on the af → zu data. Hyperparameters are selected based on chrf2 and spBLEU on the baseline model hallucinations. We set the attention scaling parameter to be 0.25. The complete set of hyperparameters is given in Table 5.

4 Results

First, we present results comparing our different CD strategies by combining all language pairs except the af → zu data. We compare Joint CD models against previous state-of-the-art CD algorithms by reporting the chrf2 and spBLEU scores. Next, we use different hallucination criteria to report the frequencies of detected hallucinations when using different CD algorithms and briefly compare the translations on one low-resource language pair. The code to replicate all results can be found in https://github.com/liuhaozhe6788/ContraDecode

Joint CD outperforms direct and model-contrastive in automatic evaluation. We report results using 5 CD algorithms, including input-contrastive decoding (Source-language), decoder-only amateur model contrastive decoding (Decoder-only), attention scaling amateur model contrastive decoding (Attention Scaling), joint input-contrastive and decoder-only amateur model contrastive decoding (Joint Decoder-only), and joint input-contrastive and attention scaling amateur model contrastive decoding (Joint Attention Scaling) in Table 1. Among all 58 language pairs, two Joint CD algorithms have lower chrf2 and spBLEU than the input-contrastive decoding method but still improve the direct method. By combining any of the model-contrastive methods with input-contrastive decoding, chrf2 improves by 0.91 (Decoder-only) and 0.53 (Attention scaling) compared with the model-contrastive method. Measurement with spBLEU shows improvement of 0.69 on Decoder-only and 0.87 on Attention scaling.

Joint CD improves off-target language rates for English hallucination, but falls short in low-resource off-target language suppression. In Table 2, Joint Decoder-only and Joint Attention Scaling models follow the input-contrastive decoding model. Joint CD algorithms significantly reduce the number of off-target outputs compared with each model-contrastive method. In translations where English is the off-target language, Joint CD algorithms outperform model-contrastive ones. In the low-resource language setting, the poor per-

	HLMT	X-branch	high-res	all
Direct	9.099%	14.433%	7.316%	12.286%
Source-language	8.833%	14.290%	6.648%	12.086%
Decoder-only	9.499%	14.678%	7.204%	12.581%
Attention Scaling	9.381%	14.644%	7.300%	12.519%
Joint Decoder-only	8.831%	14.328%	6.838%	12.112%
Joint Attention Scaling	8.879%	14.366%	6.838%	12.153%

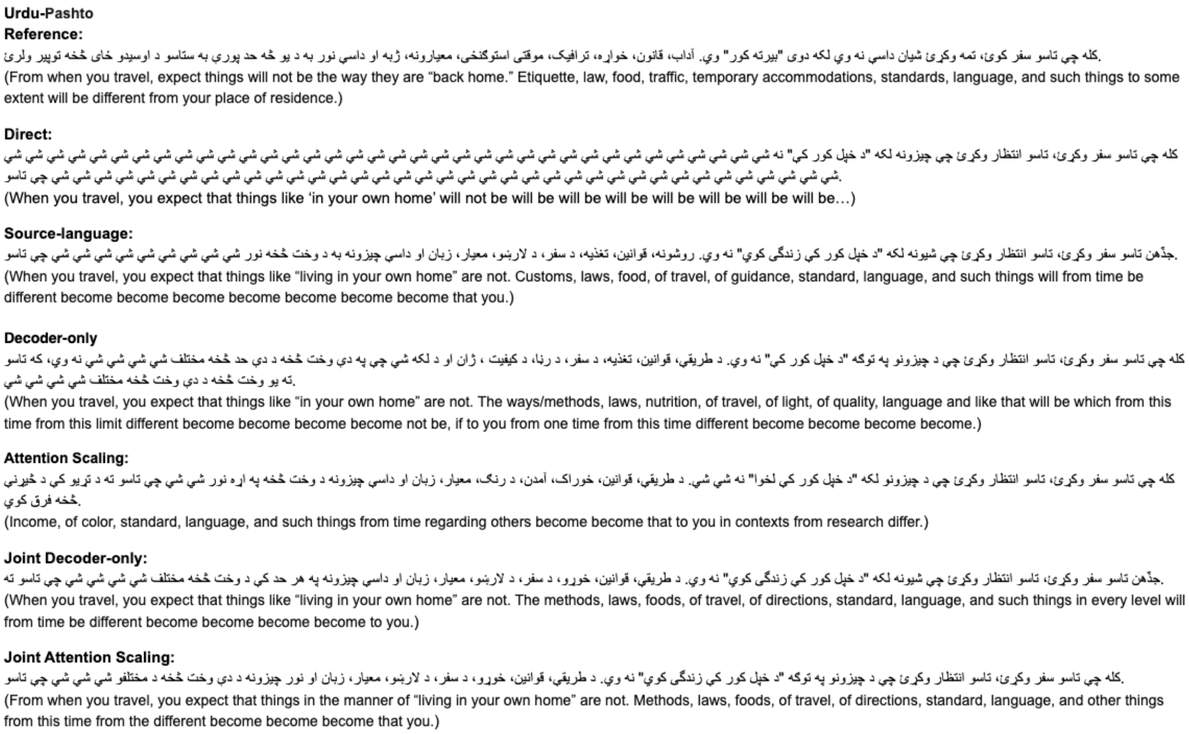
Table 3: Proportion of translations with chrf2 < 45.6.

	HLMT	X-branch	high-res	all
Direct	0.035%	2.038%	0.016%	1.246%
Source-language	0.005%	0.464%	0.016%	0.283%
Decoder-only	0.018%	1.108%	0.000%	0.677%
Attention Scaling	0.040%	0.829%	0.016%	0.517%
Joint Decoder-only	0.008%	1.067%	0.000%	0.648%
Joint Attention Scaling	0.012%	0.84%	0.000%	0.513%

Table 4: Proportion of translations with oscillatory hallucinations according to TNG.

Experiment	Source(λ)	Language(θ)	Alpha(α)	Amateur(β)	Scaling
Source-language	0.7	0.1	-	-	-
Decoder-only	-	-	0	0.05	-
Attention Scaling	-	-	0	0.05	0.25
Joint Decoder-only	0.1	0.1	0.01	0.05	-
Joint Attention Scaling	0.1	0.1	0.01	0.05	0.25

Table 5: Hyperparameters Setting



Urdu-Pashto
Reference:
کله چي تاسو سفر کوئ، تمه وکړئ شیان داسي نه وي لکه دوی "بیرته کور" وي. آداب، قانون، خواړه، ترافیک، موقتي استوګنځی، معیارونه، ژبه او داسي نور به د یو څه حد پوري به ستاسو د اوسیدو ځای څخه توپیر ولری.
(From when you travel, expect things will not be the way they are “back home.” Etiquette, law, food, traffic, temporary accommodations, standards, language, and such things to some extent will be different from your place of residence.)

Direct:
کله چي تاسو سفر وکړئ، تاسو انتظار وکړئ چي چیزونه لکه "د خپل کور کي" نه شي
شي چي تاسو.
(When you travel, you expect that things like ‘in your own home’ will not be will be will be will be will be will be will be will be…)

Source-language:
جڈهن تاسو سفر وکړئ، تاسو انتظار وکړئ چي شیونه لکه "د خپل کور کي زندګی کوي" نه وي. روشونه، قوانین، تغذیه، د سفر، د لاربنو، معیار، زبان او داسي چیزونه به د وخت څخه نور شي شي شي شي شي شي شي شي شي چي تاسو.
(When you travel, you expect that things like “living in your own home” are not. Customs, laws, food, of travel, of guidance, standard, language, and such things will from time be different become become become become become become become that you.)

Decoder-only
کله چي تاسو سفر وکړئ، تاسو انتظار وکړئ چي د چیزونو په توګه "د خپل کور کي" نه وي. د طریقي، قوانین، تغذیه، د سفر، د رڼا، د کیفیت ، ژان او د لکه شي چي په دي وخت څخه د دي حد څخه مختلف شي شي شي شي نه وي، که تاسو
ته یو وخت څخه د دي وخت څخه مختلف شي شي شي شي.
(When you travel, you expect that things like “in your own home” are not. The ways/methods, laws, nutrition, of travel, of light, of quality, language and like that will be which from this time from this limit different become become become become not be, if to you from one time from this time different become become become become.)

Attention Scaling:
کله چي تاسو سفر وکړئ، تاسو انتظار وکړئ چي د چیزونو لکه "د خپل کور کي لخوا" نه شي شي. د طریقي، قوانین، خوراک، آمدن، د رنګ، معیار، زبان او داسي چیزونه د وخت څخه په اړه نور شي شي چي تاسو ته د تریو کي د څېړني
څخه فرق کوي.
(Income, of color, standard, language, and such things from time regarding others become become that to you in contexts from research differ.)

Joint Decoder-only:
جڈهن تاسو سفر وکړئ، تاسو انتظار وکړئ چي شیونه لکه "د خپل کور کي زندګی کوي" نه وي. د طریقي، قوانین، خوړو، د سفر، د لاربنو، معیار، زبان او داسي چیزونه په هر حد کي د وخت څخه مختلف شي شي شي شي چي تاسو ته.
(When you travel, you expect that things like “living in your own home” are not. The methods, laws, foods, of travel, of directions, standard, language, and such things in every level will from time be different become become become become to you.)

Joint Attention Scaling:
کله چي تاسو سفر وکړئ، تاسو انتظار وکړئ چي د چیزونو په توګه "د خپل کور کي زندګی کوي" نه وي. د طریقي، قوانین، خوړو، د سفر، د لاربنو، معیار، زبان او نور چیزونه د دي وخت څخه د مختلفو شي شي شي چي تاسو.
(From when you travel, you expect that things in the manner of “living in your own home” are not. Methods, laws, foods, of travel, of directions, standard, language, and other things from this time from the different become become become that you.)

Figure 2: An example of different CD algorithms trying to fix the oscillation hallucination in the Direct model. All English versions of the Pashto generation are literal translations from ChatGPT.

formance of Decoder-only and Attention Scaling lies in the difficulty of accentuating what the expert model has learned using model-contrastive methods. Therefore, the Joint CD methods cannot outperform the input-contrastive method.

Joint CD is almost as good as input CD in defective hallucination reduction but fails to correct oscillation in model CD. The proportion of translations with chrf2 < 45.6 is shown in Table 3. We observe a significant decrease in the number of defective translations using Joint CD methods, with a reduction of -1.4% and -1.1% respectively, while the input-contrastive method yields a -1.6% decrease. Joint CD methods outperform their model-contrastive counterpart in reducing defective translations. For oscillation hallucinations, we notice the obvious advantages of all 5 CD methods. However, the improvement from model-contrastive to joint-contrastive methods is modest, which shows that joint-contrastive methods cannot effectively auto-correct oscillation hallucinations from model-contrastive decoding.

Repaired hallucinations still have errors, but Joint CD improves translation. In Figure 2, we provide examples of a sentence for a low-resource language pair. All CD methods try to fix the oscillation hallucination in the Direct method. Particularly, the Joint CD methods contain the least oscillation patterns and produce the most natural translations. However, beyond hallucination, they show syntactic issues such as verb adj/adv disorder, which cannot be solved by CD decoding.

5 Related Work

Hallucinations and Off-Target Translation. Neural machine translation models are known to produce *hallucinations*, fluent but unfaithful content, particularly when translating into low-resource languages or under domain shift (Guerreiro et al., 2023). Similarly, multilingual NMT systems can *drift* into an unintended language if language indicators are ambiguous (Burchell et al., 2023). Prior studies have proposed filtering and reranking strategies using quality-based metrics such as chrF (Popović, 2015), spBLEU, and n-gram repetition counts to detect and remove defective outputs.

Contrastive Decoding. Contrastive decoding techniques penalize tokens that the model would over-confidently generate under *perturbed* or *alternative* inputs. Sennrich et al. (2024) introduce *source-contrastive* and *language-contrastive* decoding, which respectively suppress hallucinations by contrasting against noisy source variants, and prevent off-target drift by contrasting against swapped language prompts. Independently, Waldendorf et al. (2024) propose *model-contrastive* decoding, wherein an "expert" model's predictions are compared to those of a weaker "amateur" variant, and tokens with low expert–amateur margin are penalized.

Decoding Enhancements. Beyond contrastive approaches, various decoding modifications have been explored to improve translation faithfulness and efficiency, including attention-scaling (Waldendorf et al., 2024), length and coverage penalties (Vaswani et al., 2017), and dynamic vocabulary pruning (Sennrich et al., 2024).

Multilingual NMT. Our work builds on large-scale multilingual translation backbones such as M2M (Fan et al., 2020), evaluated on benchmarks like FLORES-101 (Goyal et al., 2022). Despite their strong performance, these models remain vulnerable to hallucinations and off-target outputs, motivating the integration of contrastive techniques into a unified decoding framework.

6 Conclusions

This paper proposes a Joint CD framework in hallucination mitigation for multilingual NMT. Our method incorporates input CD and model CD during model inference. We show that Joint CD improves model CD in chrf2, spBLEU, and hallucination criteria. The results suggest more accurate and faithful translations in the Joint CD framework.

7 Limitations

The model CD methods in the experiments show poor performance in hallucination mitigation, which may result from minimal hyperparameter finetuning, amateur model type, or the selection of language pairs that favours the input CD method. Given sufficient resources, we would need to combine the two sets of language pairs in the experiments, which doesn't favour either of the two original CD methods. We would set M2M-100 418M and 1.2B versions as expert models and Small as the amateur models for model CD in the Joint CD framework.

References

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2024. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Jonas Waldendorf, Barry Haddow, and Alexandra Birch. 2024. Contrastive decoding reduces hallucinations in large multilingual machine translation models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2539, St. Julian's, Malta. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.