Reviewer 1:

Comment 1: This paper addresses the Emotional Voice Conversion problem, which aims to transform the discrete emotional state of a speech utterance from a source emotion to a target emotion while preserving the linguistic content. The authors introduce a self-supervised learning (SSL) framework and a novel direction latent vector modeling approach to tackle this challenge. The effectiveness of their proposed method has been demonstrated in both English and Hindi. Strength:

1. Regulating emotional intensity is an important and challenging task.
2. The proposed methods are reasonable and demonstrate relatively good performance.

Response 1: We thank the reviewer for the positive feedback on our work.

Comment 2: This paper needs improvement of writing. The innovation of the proposed method is not clearly conveyed. The author claims that the SSL-based framework and the direction latent vector modelling approach are innovations, but it is unclear how these methods differ from previous work or whether the authors have simply adapted existing techniques from other areas.

Response 2: Thanks. We will improvise writing by adding further details related to the innovation of the proposed method with respect to the SOTAs in the literature in the camera ready version. To the best of the authors‘ knowledge this is the first attempt that achieves a high-quality emotional intensity regularization in the diffusion-based EVC framework. In particular, we propose a novel approach that utilizes direction latent vector modelling in the learnt SSL-based emotional embedding space for achieving fine control over intensity while transitioning from one emotional state to another emotional state. Earlier approaches (namely, Zhou 2022c,a; Matsumoto, 2020; Schnell 2021; Choi 2021; Um 2020) fail to capture high-level complex abstract representations, which results in artifacts in the converted output and leads to significant degradation in the quality. Also, emotional manipulations often involve inter-dependencies across different types of features, hence these methods fail in achieving emotional intensity regularization. On the other hand, our approach utilizes the SSL-based audio feature representations, which are obtained after finetuning the SSL-based framework for a downstream emotion recognition task. Also some approaches like (choi 2021 and Um 2020) manipulate learned emotion representations via scaling or interpolations. However, these methods do not work well since emotional embedding space does not align well with the assumption of linear interpolation.

Comment 3: There is room for enhancing the readability of the abstract and introduction. Some concepts are introduced too early without sufficient explanation.

Response 3: Thanks. We will revise the abstract and introduction with sufficient explanation for the introduced concepts and will also provide necessary reference in the camera-ready version.

Comment 4: More importantly, the compared methods are somewhat outdated and insufficient.

Response 4: Thanks. Emotion intensity regularization is an underexplored research problem. There are mainly two approaches (EmoVox 2022 and Mixed Emotion 2022) in the literature for emotion intensity control. In addition, we have compared our results with the SOTA EVC methods such as, Diff-EVC (i.e., Ablation w/o DVM - ICASSP 2024), CycleGAN (2021), StayGAN-EVC (2018), Seq-seq EVC (INTERSPEECH 2021). Also, added two additional

(Zhong 2022 and Gallil 2023) baselines with these rebuttals. These methods convert emotion but do not support emotion intensity regularizations.

Submitted in Rebuttal Box with 2500 limits: (2423 in limit)

We thank the reviewer for the positive feedback on our work.

>Clarity about innovation w.r.t. SOTA

We will improve the writing by adding further details related to the innovation of the proposed method with respect to the SOTAs in the literature in the camera-ready version. To the best of the authors‘ knowledge, this is the first attempt to achieve a high-quality emotional intensity regularization in the diffusion-based EVC framework. In particular, we propose a novel approach that utilizes direction latent vector modeling in the learned SSL-based emotional embedding space for achieving fine control over intensity while transitioning from one emotional state to another emotional state. Earlier approaches (namely, Zhou 2022c,a; Matsumoto, 2020; Schnell 2021; Choi 2021; Um 2020) fail to capture high-level complex abstract representations, which results in artifacts in the converted output and leads to significant degradation in the quality. Also, emotional manipulations often involve inter-dependencies across different types of features, hence these methods fail in achieving emotional intensity regularization. On the other hand, our approach utilizes the SSL-based audio feature representations, which are obtained after finetuning the SSL-based framework for a downstream emotion recognition task. Also, some approaches like (choi 2021 and Um 2020) manipulate learned emotion representations via scaling or interpolations. However, these methods do not work well since emotional embedding space does not align well with the assumption of linear interpolation.

>Need a sufficient explanation for some concepts

We will revise the abstract and introduction with sufficient explanation for the introduced concepts and provide necessary references in the camera-ready version.

>Compared methods are outdated

Emotion intensity regularization is an underexplored research problem. There are mainly two approaches (EmoVox 2022 and Mixed Emotion 2022) in the literature for emotion intensity control. In addition, we have compared our results with the SOTA EVC methods such as, Diff-EVC (i.e., Ablation w/o DVM - ICASSP 2024), CycleGAN (2021), StayGAN-EVC (2018), Seq-seq EVC (INTERSPEECH 2021). Also, added two additional (Zhong 2022 and Gallil 2023) baselines with these rebuttals. These methods convert emotion but do not support emotion intensity regularizations.

A detailed rebuttal is provided on the demo page.

Reviewer 2:

Comment 1: The main concern is evaluation setup and its correctness while the proposed model, regardless of its limited novelty, still provides good results and allows controllable emotion changing.

Response 1: Thanks. We followed standard subjective and objective evaluation criteria from the domain of EVC and Emotion intensity regularization literature [Diff-EVC-ICASSP 2024, EmoVox 2022 and MixedEmo 2022 and SeamlessExpressive 2023].

Comment 2: limited novelty by adding DVM module and combining existing models (average phoneme encoder, emotion embedding and diffusion decoder)

Response 2: First of all, we propose a novel approach that utilizes direction latent vector modelling in the learnt SSL-based emotional embedding space for achieving fine control over intensity while transitioning from one emotional state to another emotional state. Conventional methods simply manipulate learned emotion representations via scaling or interpolations. They often fail to obtain continuous emotional intensity regularizations and produce poor quality. Moreover, emotional manipulations often involve inter-dependencies across different types of features. Hence, we proposed to learn better emotional representations via an SSL-based approach. In particular, DVM module leverages the GMM and the PCA for estimating direction vectors. This allows smooth traversal across emotional states with desired intensity levels. Furthermore, we utilize a diffusion-based decoder for high quality synthesis. To the best of the authors' knowledge, this is the first approach to achieve emotional intensity regulation within a diffusion-based EVC framework.

Comment 3: The authors claim that the proposed diffusion model is best suited for emotion intensity regularization tasks, but I do not see any evidence for that in the paper. The proposed DVM techniques should be tested with other, possibly non-diffusional models.

Response 3: Thanks for this excellent suggestion. We also obtained similar improvement in the performance (~1%) with DVM-based approach in the case of non-diffusion-based architecture, namely VITS. We will incorporate this in the camera ready version.

Comment 4: The incremental improvement of the DVM-based method against non-DVM doesn't strongly support the claims in the introduction of its effectiveness.

Response 4: Thank you for the comment. It is apparent from Fig. 4 that the emotion similarity score increases as the emotion intensity scale rises which demonstrates that our EmoReg with DVM can achieve fine control over emotion intensity. In contrast, non-DVM baselines (EmoVox and MixEmo) and the ablation study do not show increase in the similarity scores with an increase in intensity scale. Hence, non-DVM approaches fail to achieve fine control over emotion intensity. We also conducted now intensity scale wise subjective scores, that confirms the similar trend. Please refer Fig. 4, Table 1 and Fig. 5 in the camera ready version and refer additional samples on the demo page.

Comment 5: The baselines are missing latest emotional vc models, such as:

- [1] Disentanglement of Emotional Style and Speaker Identity for Expressive Voice Conversion (https://arxiv.org/pdf/2110.10326.pdf)
- [2] Speaking Style Conversion With Discrete Self-Supervised Units (https://arxiv.org/pdf/2212.09730.pdf)

Response 5: Thanks. We have added results of the two recommended baselines in the camera ready version. The proposed approach obtains 9.5% and 9% of relative improvements compared to these baselines in the emotional similarity and AutoPCP-based evaluations, respectively.

Comment 6: The evaluation is not clear:

- it should be specified for each Figure and Table what dataset the authors used (ESD or Hindi)
- Is it done on unseen / seen speakers during training, is it trained and evaluated on the full dataset without split?

- Are the conversions done between the emotions of the same speaker in the evaluation?
- Are you using Chinese data in ESD for training / evaluation?

Response 6: Thank you for your feedback. We will clarify our evaluation setup in the final version. Dataset details have been added to the captions of each figure and table. We used a 90:10 train-test split for both ESD-English and Hindi data, focusing on seen speakers, and did not use Chinese data. In this work, emotional conversions utilize reference emotional samples from different speakers; in future work, we plan to analyse conversions within the same speaker.

Comment 7: In the emotion intensity experiments (Fig. 4) how the evaluation is done at scale=0, 0.1, do you compare the converted emotion to the source emotion or to the target emotion? Also the score looks very high > 91% and it looks strange to me how is it possible?

Response 7: Here, we computed the cosine distance between the emotional embeddings of the converted sample (at scale 0) and target emotional embeddings [VECL-TTS INTERSPEECH24]. Emotional embeddings usually contain information related to the content, gender, and other speech-related information. Hence, when the content, gender and acoustic environment and other elements of the converted and target emotional embeddings are same, it usually impacts the score. Therefore, while absolute scores may seem high (even at low intensity scales), the relative score changes with varying intensity levels should make more sense. In addition, we have now also calculated subjective evaluation scores for emotional similarity at different intensity levels, which also confirms the similar trend.

Comment 8: Can you provide the technical details on the models architectures, number of parameters?

Response8: Thanks. Yes, we will included details about the number of parameters in the experimental setup section of camera ready version. Number of parameters in the Encoder is 8.5mn and decoder is 118mn.

Submitted in Rebuttal Box with 2500 limits:

Thanks for the positive comments.

>Concerns on evaluation

We followed SOTA subjective and objective evaluations from the literature [DiffEVC24, EmoVox22].

>Limited novelty

We propose a novel approach that utilizes direction latent vector modeling in the learned SSL-based emotional embedding space for achieving fine control over intensity while transitioning from one emotional state to another. Conventional methods manipulate learned emotion representations via interpolation. They often fail to obtain emotion intensity regularization and produce poor quality. Hence, we proposed to learn better emotional representations via the SSL approach. In particular, the DVM module leverages the GMM and the PCA for estimating direction vectors. This allows smooth traversal across emotional states with desired intensity levels. We utilize a diffusion-based decoder for high-quality synthesis. To the best of the authors' knowledge, this is the first approach to achieve emotion intensity regulation in the diffusion-based EVC framework.

> Apply DVM in a non-diffusion model

We obtained a 1% improvement with DVM in non-diffusion-based architecture, namely VITS. We will incorporate this in the camera-ready version.

> Incremental improvements for non-DVMapproaches

It is apparent from Fig. 4 that the emotion similarity score increases as the emotion intensity scale rises which indicates that our EmoReg with DVM can achieve fine emotion intensity regularization. In contrast, non-DVM baselines (EmoVox and MixEmo) do not show an increase in the similarity scores with an increase in intensity scale. Hence, they fail to achieve fine emotion intensity regularization. We also conducted intensity scale-wise subjective scores, that confirm the similar trend. Please refer to Fig 4, Table 1, and Fig 5 in the manuscript.

> Suggested baselines and additional details

We will add results from the two suggested baselines and additional evaluation details in the final version. The proposed approach obtains 9.5% and 9% relative improvements over these baselines in the emotion similarity and AutoPCP evaluations, respectively. We will add dataset details in each figure and table caption. We used a 90:10 train-test split for both ESD-English (excluding Chinese) and Hindi data, focusing on seen speakers. Here, we use reference emotional samples from different speakers. We will include the same speaker analysis in the future. The number of parameters in the Encoder is 8.5mn and the decoder is 118mn.

Reviewer 3:

Comment 1: The idea of using a directional vector in the latent space and combining it with the required intensity value is interesting and shown to perform well.

Response 1: Thanks for appreciating our idea and acknowledging better results. Our proposed approach is quite interesting in achieving emotional intensity regularization in the EVC task. In fact this is relatively new research area and not much work is done due to unavailability of a large emotional speech database with diverse emotional expressions and a wide range of emotion intensities.

Comment 2: The application of the method to Hindi is another point in this paper which differentiates it from other works in this domain.

Response 2: Thanks for highlighting our contributions. We tried to present that our proposed idea is not limited to one language and works well across different languages. In particular, we obtain 3% and 1% of improvements across different evaluations with DVM-based approaches for English and Hindi, respectively.

Comment 3: Although the idea is simple, it is shown to be effective for speech recognition quality. In the metric for emotion similarity the method is shown to be better, but this requires further evaluation

Response 3: Thanks. We will add two additional evaluation criteria to present the effectiveness of the proposed approach, namely, AutoPCP [Seamless23]-based objective evaluations for emotional similarity and scale-wise subjective evaluations for emotional controllability. We obtained 2 to 11% of improvements in both the evaluations with proposed DVM-based approach.

Comment 4: After going through the work, the question whether this requires parallel data is not clearly answered. If it requires parallel data, it is clearly limited in application and this has to brought out as a limitation. If not, what is the output from the DVM module during training that is concatenated. If paired data is used while training the model, a clear distinction should be made between which competing methods require paired data and those which do not

Response 4: No, our method does not require parallel data or paired data for both English and Hindi languages. The output of the DVM is the intensity of regularized emotional embedding that is conditioned to the diffusion-based decoder. Reference embeddings need not be necessarily from paired data and we have in fact taken it from the different speakers. We will clarify this in Section 4.1 of the camera ready version.

Comment 5: Only a single objective measure about emotional similarity is proposed. There are two suggestions [i] Mention the model using which the emotion similarity has been computed [ii] There are better metrics like AutoPCP [a] for judging emotional similarity. The emotion similarity scores in Table 1 are almost identical making it difficult to judge the superiority of the proposed method References

[a] Barrault, Loïc, et al. "Seamless: Multilingual Expressive and Streaming Speech Translation." arXiv preprint arXiv:2312.05187 (2023).

Response 5: Thanks. (i) We used parallel CNN-Transformer Emotion recognition model1* for evaluating emotion similarity scores. We have now incorporated additional details about the model in the updated manuscript. (ii) Thanks for suggesting an additional evaluation metric (AutoPCP). We will add AutoPCP evaluation scores in Table 1 of the camera ready version. We found that the proposed approach achieves 2%-11% relative improvements in the AutoPCP scores compared to the SOTA. The AutoPCP scores are mentioned in the below Table1. It is apparent from Table 1 that proposed method on an average achieves higher emotional similarity scores, which are statistically significant across all the emotions.

Table1: AutoPCP evaluation scores for on ESD dataset

Methods	**Neutral-Angry**	**Neutral-Sad**	**Neutral-Happy**	**Avg**
Emovox	3.3555	2.9537	2.8372	3.0488
Mixed Emotion	3.2550	2.7127	2.9570	2.9749
CycleGAN-EVC	3.2115	2.5084	2.8831	2.8677
StarGAN-EVC	3.2912	2.4995	2.7613	2.8506
Seq2Seq-EVC	3.1873	2.8387	**2.9844**	3.0035
StyleVC	3.1447	2.8557	2.7193	2.9112
DISSC	-	**3.04**	2.47	2.75
Proposed EmoReg w/o DVM	3.3016	2.8547	2.9115	3.0226
Proposed EmoReg with DVM	**3.3991**	2.9865	2.9382	**3.1079**

1*: https://github.com/Data-Science-kosta/Speech-Emotion-Classification-with-PyTorch

Submitted in Rebuttal Box with 2500 limits:

Thanks for appreciating our idea and acknowledging better results.

>Appreciation of our idea and Hindi language-related contribution

Our proposed approach is quite interesting in achieving emotional intensity regularization in the EVC task. This is a relatively new research area and not much work is done due to the unavailability of a large emotional speech database with diverse emotional expressions and a wide range of emotion intensities. In addition, we aimed to present that our proposed idea is not limited to one language and works well across different languages. In particular, we obtain 3% and 1% improvements across different evaluations with DVM-based approaches for English and Hindi, respectively.

>Additional objective evaluation for emotional similarity

We will add two additional evaluation criteria to present the effectiveness of the proposed approach, namely, AutoPCP [Seamless23]-based objective evaluations for emotional similarity and scale-wise subjective evaluations for emotional controllability. We obtained 2 to 11% improvements in both the evaluations with the proposed DVM-based approach compared to the SOTAs. The AutoPCP scores will be added in Table 1 of camera camera-ready version.

>Clarification about the model for calculating the Emotion Similarity score

We used a parallel CNN-transformer-based model for evaluating emotion similarity scores. We have now incorporated additional details about the model in the camera-ready version. It is apparent from Table 1 that the proposed method on average achieves higher emotional similarity scores, which are statistically significant across all the emotions.

>Clarifications about parallel data

No, our method does not require parallel data or paired data for both English and Hindi languages. The output of the DVM is the intensity of regularized emotional embedding that is conditioned to the diffusion-based decoder. Reference embeddings need not be necessarily from paired data, and we have in fact taken it from the different speakers. We will clarify this in Section 4.1 of the camera-ready version.

A detailed rebuttal is provided on the demo page.

Reviewer 4:

Comment 1: The work excels in its innovative integration of SSL representations with an emotion-focused approach and the introduction of direction vector modeling. The framework has proven robust and effective in experiments, exhibiting high emotion similarity, better intelligibility, and improved MOS.

Response 1: Thanks for the positive comments.

Comment 2: While the intensity of emotion transfer is a major claim of the proposed framework, the paper provides limited evaluation of this aspect. The authors have some efforts to use an emotion classifier, the change is however subtle (from 0.94 to 0.96), which is not very convincing given the emotion classifier is unknown and the behaviour is not clear in 0.0x. The reviewer would recommend testing in subjective evaluation instead.

Response 2: We would like to thank the reviewer for the suggestion. Following your suggestion, we conducted a subjective evaluation for the proposed model at different intensity scales and will include these results in the final version. One possible reason for the subtle change in the emotional similarity score (e.g., from 0.94 to 0.96) is that the emotional embeddings we use also capture information on content, gender, and other speech characteristics, not just emotion. Hence, when the content, gender and acoustic environment and other elements of the converted and target emotional embeddings are same, they can impact the score, leading to smaller changes. Despite subtle changes, the emotion similarity score increases with an increase in emotion intensity scale which shows that the proposed EmoReg with DVM can achieve fine control over emotion intensity. This fine-grained control is also clearly perceptible in our demo samples. In contrast, non-DVM baselines (EmoVox and MixEmo) show no variation in similarity score with intensity changes, confirming their lack of control over emotion intensity.

Table: Scale wise Subjective Evaluation Scores

Methods	**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
Emovox	45.846	46.949	45.590	43.949	43.513	44.000
Mixed Emotion	34.56	42.06	47.39	51.70	51.26	54.87
Proposed EmoReg w/o DVM	68.205	66.949	68.385	71.923	71.179	73.000
Proposed EmoReg with DVM	**70.000**	**71.923**	**71.974**	**72.256**	**73.026**	**73.103**

Comment 3: There are missing details in the experimental setup and evaluation, which are crucial for reproduction, reasoning, and thorough analysis:

- Baseline information: The baseline models (EmoVox and Mixed Emotion) perform poorly in the presented results, especially compared to their original evaluations. It would be helpful to clarify whether the models used were from official checkpoints or re-implemented and trained by the authors. If they were re-implemented, differences in model parameters and setups should be discussed.
- Missing details on models used in objective evaluations:
    - Which ASR model is used for WER/CER evaluation?
    - What kind of speech emotion classifier is used for emotion similarity? If this classifier is related to the emotion2vec model, could this introduce a bias in favor of EmoReg, which has a dependency on emotion2vec?

Response 3: Thanks. We will incorporate all the missing details in Section 4 related to experimental setup and evaluation in the camera ready version. We have utilized their (EmoVox 2022 and MixedEmo 2022) pre-trained models for the English language. We found the quality of the generated samples matched with the samples shared by the original authors on their demo page (It was indeed poor!). For Hindi, we have finetuned models from official repos of those papers. Thanks. We used the Whisper-small-244mn model for WER/CER calculations. For a fair comparison, we used a completely different emotion classifier based on CRNN architecture and utilized the Mel spectrogram as a feature. Note, we did not the emotion2vec features to avoid possible bias in the similarity score for the propose model. We have clarified these in our manuscript..

Comment 4: The authors use a neural emotion as the base from which other emotion vectors are derived through subtraction. Since all test scenarios are based on neutral emotion, it would be interesting to explore whether the method can change emotions starting from non-neutral emotions.

Response 4: Thanks for this suggestion. We will briefly highlight this missing results in the paper. Due to space limitations, we have only shown results from neutral emotions to non-neutral emotions. However, we have observed that our method works equally well in non-neutral to neutral emotion conversion scenarios as well. We will add corresponding samples on the demo page.

Submitted in Rebuttal Box with 2500 limits:

Thanks for appreciating our idea and evaluation.

>Additional scale-wise subjective evaluation recommendation.

As per the suggestion, we conducted a subjective evaluation of the proposed model at different intensity scales and will include these results in the final version. One possible reason for the subtle change in the emotional similarity score (e.g., from 0.94 to 0.96) is that the emotional embeddings we use also capture information on content, gender, and other speech characteristics, not just emotion. Hence, when the content, gender, acoustic environment, and other elements of the converted and target emotional embeddings are the same, they can impact the score, leading to smaller changes. Despite subtle changes, the emotion similarity score increases with an increase in emotion intensity scale which shows that the proposed EmoReg with DVM can achieve fine control over emotion intensity. This fine-grained control is also clearly perceptible in our demo samples. In contrast, non-DVM baselines (EmoVox and MixEmo) show no variation in similarity scores with intensity changes, confirming their lack of control over emotion intensity.

>Missing details

We will incorporate all missing details in Section 4 of the camera-ready version. We have utilized their (EmoVox and MixedEmo 22) pre-trained models for the English language. We found the quality of the generated samples matched with the samples shared by the original authors on their demo page (It was indeed poor!). For Hindi, we have finetuned models from official repos of those papers. Thanks. We used the Whisper-small-244mn model for WER/CER calculations. For a fair comparison, we used a completely different emotion classifier based on CRNN architecture and utilized the Mel spectrogram as a feature. Note, that we did not the emotion2vec features to avoid possible bias in the similarity score for the proposed model. We have clarified these in our manuscript.

> Non-neutral to neutral emotion conversion

Thanks for this suggestion. We will briefly highlight these missing results in the paper. Due to space limitations, we have only shown results from neutral emotions to non-neutral emotions. However, we have observed that our method works equally well in non-neutral to neutral emotion conversion scenarios as well. We will add corresponding samples on the demo page.

A detailed rebuttal is provided on the demo page.