We kindly thank the reviewers for their feedback. All reviewers expressed their interest in more experimental evaluation. We report an additional experiment on sequence prediction and then address each reviewer's questions individually.

Sequence Prediction. We considered the *TIMIT Acoustic-Phonetic Continuous Speech Corpus* (`https://catalog.ldc.upenn.edu/LDC93S1`), containing recordings of ~ 6300 recorded sentences (630 English speakers, each reading ~ 10 sentences). We preprocessed the dataset following a standard practice in the speech community, taking 10ms frames (dropping the glottal stop 'q' labeled frames in the core test set) and mapping them as 40-dimensional vectors [D. Povey et al., *"The kaldi speech recognition toolkit"*, 2011]. The speech recognition problem is stated in terms of sequence prediction, i.e. each sentence, represented as a list of 40-dimensional vectors to be mapped in a same-length list of phonemes (131 phonemes per each 10ms frame). We considered:

1. A baseline performing independent classification of sequence elements using Kernel SVM with Gaussian Kernel. The regularization parameter C and the bandwidth of the kernel have been selected via hold-out cross validation in the logarithmic range $[10^{-3}, 10^3]$ and $[10^{-9}, 10^3]$.
2. Structural SVM for sequence tagging based on Hidden Markov Models (SVM-HMM) and exact Viterbi algorithm for inference (see below). Implementation from `https://www.cs.cornell.edu/people/tj/svm_light/svm_hmm.html` using Gaussian kernel. The regularization parameter C and the bandwidth of the kernel have been selected via hold-out cross validation in the logarithmic range $[10^{-3}, 10^3]$ and $[10^{-9}, 10^3]$.
3. The proposed *localized structured prediction algorithm*. In particular we defined parts on input as subsequences of 11 frames (110ms) corresponding to vectors of 440 elements, while the corresponding parts of the output are the couple of phonemes associated to the central and the next frame of the input.

Inference. For both SVM-HMM and our method, the inference is solved with the classic Viterbi algorithm over a table of dimension $L|x|$ where L is the number of labels per element ($L = 131$ phonemes) and $|x|$ is the lenght of the input sequence ($|x| \sim 200$ on average). When output parts are couples of phonemes this entiails $O(|x|L^2)$ computations.

Results. The Table reports the test performance (Hamming loss) of the three methods over 5 trials (dataset is reshuffled and randomly split in 3300 sentences for training and the rest for test). Leveraging the parts structure appears beneficial, with our method consistently outperforming both the baseline and the traditional structured prediction competitor.

Sequence prediction on TIMIT	Hamming
Independent SVM	0.31 ± 0.015
Struct SVM	0.29 ± 0.031
Localized Struct. Pred	$\mathbf{0.26 \pm 0.012}$

R1 * *Long Range Dependecies.* Our assumption does not limit long-range dependencies in the data as they do not imply that two parts to have identical (or very similar) appearance with respect to the similarity metric used. However, it is also true that while such dependencies could be leveraged to improve performance, our approach is not designed to capture them at training time (but only during inference). As future work we will explore the question of reformulating the part-based regression as a multi-task learning (MTL) problem, with each task corresponding to a part. Then, by leveraging ideas from the MTL literature we could impose (when known a-priori) or learn (when unknown) the multi-scale relations/dependencies between tasks/parts during the training phase. Algorithmically this extension would require a small modification of the current approach. We will add a comment on this when discussing future directions.
* *Inference.* The inference step in our setting is formulated as an optimization problem over the output space Z as most traditional structured prediction approaches (see Remark 2 in our paper). Hence, our inference step is as difficult as previous methods, from a computational viewpoint (while also providing strong theoretical guarantees on the prediction). For instance, in the experiments on TIMIT in this document, both our method and SVM-HMM use the viterbi algorithm for inference. Algorithm 3 offers an additional benefit of our formulation when first order optimization on Z is possible.

R2 * *Example 2.* We agree with R2 to improve exposition of Example 2. In particular we will clarify the difference between the loss function used to evaluate the error and the score function at inference time. While these are two separate concepts in the context of CRFs, they essentially coincide in our Localized Structured Prediction framework (score is a linear combination of losses). The conclusion of Example 2 highlights the connection between CRFs and our framework by observing that the score functions have essentially the same structure. Hence they lead to the same inference problem. *[Sutton, C and McCallum, A. "An introduction to conditional random fields." 2012].*
* *Notation* We agree with R2: will use $\Delta(z, y)$ in the main text and $\Delta(z, y|x)$ only in the appendix. This notation was originally introduced to account for settings where the parts might depen on the input (e.g. images with differnt sizes).

R3 **Attention Models.* Loosely speaking, an attention model could be formulated within our framwork by considering a parametrization of the possible part structures: in this sense, the attention would consist in the process of selecting iteratively the parts that are more relevant to the task and adapt them depending on individual inputs. This model however would require to perform an optimization over the parts parametrization, making both learning and inference significantly more challegning. Investigating this question could lead to interesting future work, we thank the reviewer and we will add a comment to the paper. **Code.* Our code is in python (+ pytorch) and will be made available in Github.