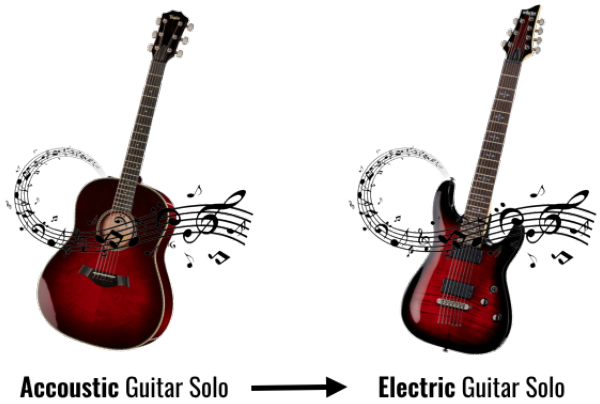
National Technical University of Athens
School of Electrical and Computer Engineering
Division of

EditGen: Harnessing Cross Attention Control for Instruction-Based Autoregressive Audio Editing

Diploma Thesis

of

VASSILEIOS SIOROS



Supervisor: Alexandros Potamianos
Associate Professor

Athens, May 2024

National Technical University of Athens
School of Electrical and Computer Engineering
Division of

EditGen: Harnessing Cross Attention Control for Instruction-Based Autoregressive Audio Editing

Diploma Thesis

of

VASSILEIOS SIOROS

Supervisor: Alexandros Potamianos
Associate Professor

Approved by the examination committee on 19th July 2024.

(Signature)	*(Signature)*	*(Signature)*
.....	.....	.....
Alexandros Potamianos	Athanasios Rontogiannis	Constantinos Tzafestas
Associate Professor	Associate Professor	Associate Professor

Athens, May 2024

National Technical University of Athens
School of Electrical and Computer Engineering
Division of

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism.

(Signature)

.....
Vassileios Sioros
19th July 2024

Abstract

In this study, we investigate leveraging cross-attention control for efficient audio editing using auto-regressive models. Inspired by image editing methodologies, we develop a Prompt-to-Prompt-like approach that guides edits through cross and self-attention mechanisms. Integrating a diffusion-based strategy, influenced by Auffusion, we extend the model's functionality to support prompt-guided refinement editing. Additionally, we introduce an alternative approach by incorporating MUSICGEN, a pre-trained frozen auto-regressive model, and propose three editing mechanisms, based on Replacement, Reweighting, and Refinement of the attention scores. We employ commonly-used music-specific evaluation metrics and a human study, to gauge time-varying controllability, adherence to global text cues, and overall audio realism. The automatic and human evaluations indicate that the proposed combination of prompt-to-prompt guidance with autoregressive generation models significantly outperforms the diffusion-based baseline in terms of melody, dynamics, and tempo of the generated audio.

Keywords

Audio Content Editing, Prompt-Guided Audio Manipulation, AI-Driven Audio Synthesis, Fine-Grained Audio Control, Text-Based Audio Editing, Cross-Modal Model Adaptation

Περίληψη

Στην παρούσα μελέτη, ερευνούμε την αξιοποίηση του ελέγχου διασταυρούμενης προσοχής για αποτελεσματική επεξεργασία ήχου χρησιμοποιώντας αυτοπαλινδρομικά μοντέλα. Εμπνευσμένοι από μεθοδολογίες επεξεργασίας εικόνας, αναπτύσσουμε μια προσέγγιση τύπου Prompt-to-Prompt που καθοδηγεί τις επεμβάσεις μέσω μηχανισμών διασταυρούμενης και αυτοπροσοχής. Ενσωματώνοντας μια στρατηγική διάχυσης, επηρεασμένη από το Auffusion, επεκτείνουμε τη λειτουργικότητα του μοντέλου για να υποστηρίξει την επεξεργασία βελτίωσης καθοδηγούμενη από προτροπές. Επιπλέον, εισάγουμε μια εναλλακτική προσέγγιση ενσωματώνοντας το MUSICGEN, ένα προ-εκπαιδευμένο παγωμένο αυτοπαλινδρομικό μοντέλο, και προτείνουμε τρεις μηχανισμούς επεξεργασίας, βασισμένους στην Αντικατάσταση, την Ανακατανομή βαρών και τη Βελτίωση των σκορ προσοχής. Χρησιμοποιούμε ευρέως χρησιμοποιούμενες μετρικές αξιολόγησης ειδικές για τη μουσική και μια μελέτη με ανθρώπους, για να αξιολογήσουμε την ελεγχόμενη μεταβλητότητα στον χρόνο, την τήρηση των γενικών κειμενικών οδηγιών και τον συνολικό ρεαλισμό του ήχου. Οι αυτόματες και οι ανθρώπινες αξιολογήσεις υποδεικνύουν ότι ο προτεινόμενος συνδυασμός καθοδήγησης τύπου προμπτ-το-προμπτ με αυτοπαλινδρομικά μοντέλα δημιουργίας υπερέχει σημαντικά σε σχέση με την βασική στρατηγική διάχυσης όσον αφορά τη μελωδία, τη δυναμική και τον ρυθμό του παραγόμενου ήχου.

Λέξεις Κλειδιά

Επεξεργασία Περιεχομένου Ήχου, Καθοδηγούμενη από Προτροπές Επεξεργασία Ήχου, Σύνθεση Ήχου με Τεχνητή Νοημοσύνη, Λεπτομερής Έλεγχος Ήχου, Επεξεργασία Ήχου Βασισμένη σε Κείμενο, Διασταυρούμενη Προσαρμογή Μοντέλων

Table of Contents

List of Figures

List of Tables

Εκτεταμένη Ελληνική Περίληψη

Η επίτευξη ικανοποιητικών αποτελεσμάτων σε εργασίες επεξεργασίας ήχου απαιτούσε συνήθως μεγάλα σύνολα δεδομένων η επισήμανση των οποίων είναι διαδικασία που συχνά είναι χρονοβόρα και δαπανηρή. Επιπλέον, η δημιουργία αποτελεσματικών μοντέλων προσαρμοσμένων στην επεξεργασία ήχου απαιτεί σημαντική εξειδίκευση και πειραματισμό. Η αξιοποίηση υπαρχουσών μοντέλων, αν και δυνατή, παραμένει μια δαπανηρή προσπάθεια.

Σε αυτή την εργασία, χρησιμοποιούμε μοντέλα καθοδηγούμενα από εντολές στον τομέα του ήχου και, συγκεκριμένα, προσαρμόζουμε την τεχνική Prompt-to-Prompt [1], που ήταν προηγουμένως επιτυχής στην επεξεργασία εικόνων, για την επεξεργασία ήχου. Αυτή η προσέγγιση επιτρέπει λεπτομερή επεξεργασία ήχου χωρίς να χρειάζεται επαναπροσαρμογή ενός μοντέλου ή πρόσθετων δεδομένων. Το Prompt-to-Prompt αξιοποιεί χάρτες διασταυρούμενης προσοχής για να ελέγχει τη διαδικασία δημιουργίας, επιτρέποντας στους χρήστες να επηρεάζουν τον τρόπο με τον οποίο τα παραγόμενα ηχητικά δεδομένα αλληλεπιδρούν με τις κειμενικές εντολές. Αυτή η ευελιξία επιτρέπει εργασίες όπως αλλαγές τιμών λέξεων, γενική επεξεργασία ήχου και χειρισμό σημασιολογικών εφέ χωρίς τροποποίηση του υποκείμενου μοντέλου.

Εφαρμόζουμε το Prompt-to-Prompt για την επεξεργασία μουσικής και το προσαρμόζουμε σε ένα αυτοπαλινδρομικό μοντέλο που ακολουθεί εντολές. Αυτή είναι η πρώτη επιτυχής εφαρμογή του Prompt-to-Prompt στο πλαίσιο επεξεργασίας ήχου με αυτοπαλινδρομικό μοντέλο. Η εκτενής αξιολόγησή μας αποδεικνύει τις δυνατότητες αυτών των τεχνικών για διαισθητική επεξεργασία ήχου βασισμένη σε κείμενο. Οι κύριες συμβολές μας είναι:

1. Εξετάζουμε τη χρήση ενός προεκπαιδευμένου αυτοπαλινδρομικού μοντέλου Transformer, αρχικά σχεδιασμένου για τη δημιουργία δειγμάτων υψηλής ποιότητας μουσικής από μια δεδομένη κειμενική εντολή, για την επεξεργασία ήχου.

2. Σχεδιάζουμε και υλοποιούμε τρεις διακριτούς μηχανισμούς επεξεργασίας ήχου, εμπνευσμένους από αυτούς που παρουσιάστηκαν στο [1].

3. Αξιολογούμε την προσέγγιση μας με βάση το αυτοπαλινδρομικό μοντέλο Transformer σε σύγκριση με υπάρχουσες μεθόδους διάχυσης, χρησιμοποιώντας αυτόματες μετρικές σχετικές με τη μουσική και ανατροφοδότηση από χρήστες.

Δημιουργία ήχου με βάση τη διάχυση: Τα μοντέλα που βασίζονται στη διάχυση έχουν εξερευνηθεί ευρέως στον τομέα του ήχου και της μουσικής. Οι Yang και λοιποί [2] χρησιμοποιούν ένα μοντέλο VQ-VAE εκπαιδευμένο σε mel-φασματογραφήματα για να τα μετατρέψουν σε διακριτούς κωδικούς. Αυτοί οι κωδικοί στη συνέχεια τροφοδοτούνται σε

ένα μοντέλο διάχυσης για τη δημιουργία ηχητικών σημάτων. Το Make-An-Audio [3] χρησιμοποιεί έναν αυτόματο κωδικοποιητή φασματογραφήματος, ενσωματώνει το CLAP [4] και εισάγει μια προσέγγιση ψευδο-βελτίωσης εντολών, ευθυγραμμίζοντας τις φυσικές γλώσσες με τα ηχητικά δεδομένα, επιτρέποντας την αξιοποίηση τεράστιων δεδομένων διαθέσιμων προς εκπαίδευση χωρίς επιτήρηση. Το AudioLDM [5] χρησιμοποιεί ένα λανθάνον μοντέλο διάχυσης (LDM) και αντιμετωπίζει την έλλειψη επισημασμένων δεδομένων εκπαιδεύοντας χρησιμοποιώντας αποκλειστικά ηχητικά δεδομένα. Το AudioLDM 2 [6] εισάγει την έννοια της Γλώσσας του Ήχου (LOA), χρησιμοποιώντας το AudioMAE [7] προ-εκπαιδευμένο σε ποικίλο ηχητικό περιεχόμενο. Το AUDIT [8] συνδυάζει λανθάνοντα μοντέλα διάχυσης για τη δημιουργία επεξεργασμένων ηχητικών τμημάτων χρησιμοποιώντας τόσο ηχητικά όσο και κειμενικά στοιχεία. Όσον αφορά τη μουσική, το TANGO [9], εμπνευσμένο από τα λανθάνοντα μοντέλα διάχυσης και το AudioLDM, χρησιμοποιεί ένα LLM αντί για embeddings που βασίζονται στο ΎΛΑΠ. Το Auffusion [10] χρησιμοποιεί ένα προ-εκπαιδευμένο Λανθάνον Μοντέλο Διάχυσης και το HiFi-GAN [11] vocoder. Επιπλέον, οι συγγραφείς εισάγουν έναν μηχανισμό διασταυρούμενης προσοχής που ενισχύει την ευθυγράμμιση και την ευελιξία. Το MusicLDM [12] προσαρμόζει τις αρχιτεκτονικές Stable Diffusion και AudioLDM στον τομέα της μουσικής. Για να αντιμετωπίσουν την πρόκληση των περιορισμένων δεδομένων εκπαίδευσης, προτείνονται νέες στρατηγικές mixup: το beat-synchronous audio mixup (BAM) και το beat-synchronous latent mixup (BLM). Το InstructME [13] αξιοποιεί ένα προσαρμοσμένο λανθάνον μοντέλο διάχυσης, διευκολύνοντας εργασίες όπως η προσθήκη, η αφαίρεση και η αναμίξη μουσικών στοιχείων διατηρώντας την αρμονική ακεραιότητα μέσω πινάκων προόδου συγχορδιών.

Αυτοπαλινδρομική δημιουργία ήχου: Ως εναλλακτική στα μοντέλα διάχυσης για τη δημιουργία ήχου και μουσικής, υπάρχουν τα αυτοπαλινδρομικά μοντέλα. Το WaveNet [14] εισήγαγε μια αυτοπαλινδρομική μέθοδο ταξινόμησης για σύνθεση ομιλίας, υπερβαίνοντας τις παραδοσιακές προσεγγίσεις συνένωσης και παραμετρικής προσέγγισης, αν και πιο αργό. Το AudioGen [15] ξεπέρασε το Diffsound χρησιμοποιώντας αυτοπαλινδρομική μοντελοποίηση σε διακριτούς χώρους κυματομορφών. Το Jukebox [16] χρησιμοποιεί ένα πολυκλίμακο VQ-VAE [17] για να συμπιέσει τον ακατέργαστο ήχο σε διακριτούς κωδικούς, οι οποίοι στη συνέχεια μοντελοποιούνται χρησιμοποιώντας αυτοπαλινδρομικούς μετασχηματιστές. Το AudioLM [18] χρησιμοποιεί ετικέτες που δημιουργούνται από έναν νευρωνικό κωδικοποιητή SoundStream [19] [20, 21] ως στόχους για μια εργασία μοντελοποίησης ακολουθίας. Στο [22], οι Agostinelli και λοιποί εισάγουν το MusicLM, ακολουθώντας μια παρόμοια προσέγγιση με το AudioLM αλλά προσαρμοσμένο για εργασίες επεξεργασίας μουσικής. Το MUSICGEN, ένας αποκωδικοποιητής βασισμένος σε μετασχηματιστή που εισήχθη στο [23], χρησιμοποιεί το EnCodec [24]. Η προσαρμογή στο κείμενο ενσωματώνει τεχνικές όπως οι T5 encoder, FLAN-T5 και CLAP, ενώ η προσαρμογή στη μελωδία χρησιμοποιεί κυρίαρχα βινς χρόνου-συχνότητας για τον έλεγχο της μελωδικής δομής. Το μοντέλο εισάγει επίσης ένα πλαίσιο για μοτίβα εναλλαγής του βιβλίου κωδίκων, βελτιώνοντας την αποδοτικότητα και την ευελιξία.

Τεχνικές επεξεργασίας: Το Prompt-to-Prompt [1] αξιοποιεί εσωτερικά στρώματα διασταυρούμενης προσοχής για να ελέγξει ποια pixels δίνουν προσοχή σε ποια tokens, επιτρέποντας

εργασίες όπως αλλαγές τιμών token, παγκόσμια επεξεργασία εικόνας και ενίσχυση/αποδυνάμωση σημασιολογικών αποτελεσμάτων χωρίς επανεκπαίδευση του μοντέλου. Το Textual Inversion [25] χρησιμοποιεί μια σειρά θορυβωμένων λανθανόντων κωδίκων που αποκτώνται από την αρχική αναστροφή DDIM ως σημείο αναφοράς και βελτιστοποιεί την ενσωμάτωση του null-text. Με την εκπαίδευση ενός προ-εκπαιδευμένου μοντέλου text-to-image με μερικές εικόνες ενός αντικειμένου, το Dreambooth [26] συνδέει ένα μοναδικό αναγνωριστικό με το αντικείμενο. Το Dreambooth επιτρέπει τη σύνθεση φωτορεαλιστικών εικόνων του αντικειμένου στο πλαίσιο διάφορων σκηνών. Στο [27] οι συγγραφείς προτείνουν το Custom Diffusion, όπου βελτιστοποιούν μόνο μερικές παραμέτρους στον μηχανισμό προσαρμογής text-to-image για να αντιπροσωπεύσουν νέες έννοιες. Αυτή η προσέγγιση αποδίδει εξίσου καλά ή καλύτερα από υπάρχουσες μεθόδους διατηρώντας την υπολογιστική αποδοτικότητα. Το SVDiff [28], βελτιστοποιεί τις μοναδιαίες τιμές των μητρών βαρών, με αποτέλεσμα έναν συμπαγή χώρο παραμέτρων, μειώνοντας τον κίνδυνο υπερεκπαίδευσης και γλωσσικής εκτροπής [29]. Στο [30] χρησιμοποιούν το Textual Inversion [25] και το Dreambooth [26] για να εξατομικεύσουν τις εξόδους του AudioLDM για νεοεκπαιδευμένες μουσικές έννοιες με λίγα παραδείγματα. Στο [31], οι συγγραφείς διερευνούν δύο τεχνικές επεξεργασίας ήχου zero-shot χρησιμοποιώντας αναστροφή DDPM σε προ-εκπαιδευμένα δίκτυα διάχυσης. Η προσέγγισή τους, βασισμένη στο [32], περιλαμβάνει την εξαγωγή λανθανόντων διανυσμάτων θορύβου που αντιστοιχούν στο αρχικό σήμα και τη χρήση αυτών των διανυσμάτων σε μια διαδικασία δειγματοληψίας DDPM για να καθοδηγήσουν τη διάχυση προς την επιθυμητή επεξεργασία. Για την επεξεργασία βάσει κειμένου, προσαρμόζουν την κειμενική εντολή που δίνεται στο μοντέλο. Στο σενάριο χωρίς επίβλεψη, διαταράσσουν την έξοδο του μοντέλου κατά μήκος των κατευθύνσεων των κύριων συνιστωσών της κατανομής.

Βασιζόμενοι στα επιτεύγματα του Prompt-to-Prompt και στην ανώτερη ποιότητα ήχου του MUSICGEN, σκοπεύουμε να τα ενσωματώσουμε σε ένα αυτοπαλινδρομικό πλαίσιο. Αυτό θα συνδυάσει την επεξεργασιμότητα του Prompt-to-Prompt με την υψηλή ποιότητα ήχου των σύγχρονων αυτοπαλινδρομικών μοντέλων όπως το MUSICGEN.

Το μοντέλο MUSICGEN [23] χρησιμοποιεί το EnCodec [24], έναν συνελικτικό κωδικοποιητή που συμβάλλει στην κβαντοποίηση του χώρου λανθάνοντων μεταβλητών. Η είσοδος, μια τυχαία μεταβλητή αναφοράς ήχου X, κωδικοποιείται σε ένα συνεχές τανυστή με χαμηλότερο ρυθμό καρέ (f_r) σε σύγκριση με το ρυθμό δειγματοληψίας (f_s). Η συνεχής αναπαράσταση στη συνέχεια κβαντίζεται σε διακριτά σύμβολα (Q) χρησιμοποιώντας RVQ, με αποτέλεσμα K παράλληλες ακολουθίες (για κάθε χρονικό βήμα), καθεμία με T σύμβολα, όπου K είναι ο αριθμός των βιβλίων κώδικα, και M είναι το μέγεθος του βιβλίου κώδικα. Οι συγγραφείς εφαρμόζουν μια προσέγγιση αυτοπαλινδρόμησης που προβλέπει πολλαπλά βιβλία κώδικα ταυτόχρονα, επιταχύνοντας έτσι σημαντικά τόσο την εκπαίδευση όσο και την εξαγωγή. Πιο συγκεκριμένα, το MUSICGEN χρησιμοποιεί ένα μοτίβο διαπλοκής συμβόλων για να παράγει όλα τα βιβλία κώδικα σε μία και μόνο διέλευση αποκωδικοποιητή, εξαλείφοντας την ανάγκη για διαδοχικά πολλαπλά μοντέλα και καθιστώντας το μοντέλο πολύ αποδοτικό.

Το Prompt-to-Prompt [1] εκμεταλλεύεται το μηχανισμό διασταυρούμενης προσοχής. Ας θεωρήσουμε ένα ηχητικό δείγμα A που παράγεται χρησιμοποιώντας μια κειμενική εντολή P. Λάβετε υπόψη ένα ηχητικό δείγμα A που παράγεται χρησιμοποιώντας ένα κείμενο εντολής P.

Με την εισαγωγή των χαρτών προσοχής που αποκτώνται κατά τη διάρκεια της δημιουργίας του A σε μια νέα δημιουργία με τροποποιημένη εντολή P^*, μπορούμε να πραγματοποιήσουμε μια επεξεργασία που έχει ως αποτέλεσμα ένα νέο ηχητικό δείγμα A^* το οποίο διατηρεί τη δομή του αρχικού. Για να αντιμετωπίσουμε συγκεκριμένες λειτουργίες επεξεργασίας, χρησιμοποιούμε τρεις μηχανισμούς επεξεργασίας:

Αντικατάσταση: Ο χρήστης αντικαθιστά τους όρους της αρχικής εντολής με άλλους. Για παράδειγμα, αντικαθιστώντας μια ακουστική κιθάρα με μια ηλεκτρική κιθάρα. Εισάγουμε τους χάρτες προσοχής του αρχικού δείγματος στη διαδικασία δημιουργίας με την τροποποιημένη εντολή:

$$Edit(M_t, M_t^*, t) = \begin{cases} M_t^*, & \text{ιφ } t < \tau \\ M_t, & otherwise \end{cases} \tag{1}$$

όπου τ είναι μια παράμετρος χρονικής σήμανσης που καθορίζει μέχρι ποιο βήμα εφαρμόζεται η εισαγωγή.

Διόρθωση: Ο χρήστης προσθέτει νέους όρους στην εντολή. Σε αυτήν την περίπτωση, η εισαγωγή προσοχής εφαρμόζεται μόνο στους κοινούς όρους που μοιράζονται και οι δύο εντολές:

$$(Edit(M_t, M_t^*, t))_{i,j} = \begin{cases} (M_t^*)_{i,j}, & \text{ιφ } A(j) = \emptyset \\ (M_t)_{i,A(j)}, & otherwise \end{cases} \tag{2}$$

Αξίζει να αναφερθεί ότι ο δείκτης i αντιστοιχεί σε μια τιμή, ενώ ο δείκτης j αντιστοιχεί σε έναν όρο κειμένου. Και πάλι, μπορούμε να ορίσουμε μια χρονική σήμανση τ για να ελέγξουμε τον αριθμό των βημάτων στα οποία εφαρμόζεται η παραπάνω τεχνική.

Ανακατανομή βάρους: Τέλος, ο χρήστης μπορεί να επιθυμεί να ενισχύσει ή να αποδυναμώσει την έκταση στην οποία κάθε όρος επηρεάζει το αποτέλεσμα. Για να επιτύχουμε αυτό, πολλαπλασιάζουμε τον χάρτη προσοχής του εκχωρημένου όρου j^* με μια παράμετρο c που κυμαίνεται από -2 έως 2, οδηγώντας σε ισχυρότερη ή ασθενέστερη επίδραση. Οι χάρτες προσοχής για τους άλλους όρους παραμένουν αμετάβλητοι:

$$(Edit(M_t, M_t^*, t))_{i,j} = \begin{cases} c \cdot (M_t)_{i,j}, & \text{ιφ } j = j* \\ (M_t)_{i,j}, & otherwise \end{cases} \tag{3}$$

Στην αρχική υλοποίηση του Prompt-to-Prompt, όπου χρησιμοποιούνται μοντέλα διάχυσης, οι συγγραφείς περιόριζουν τον αριθμό των βημάτων τ εφαρμογής του Prompt-to-Prompt, έτσι ώστε η παραγόμενη εικόνα να προσαρμόζεται στην γεωμετρία που επιτάσσει η νέα εντολή. Για να ενσωματόσουμε το Prompt-to-Prompt με τα αυτοπαλινδρομικά χαρακτηριστικά του MUSICGEN, εφαρμόζουμε τη διαδικασία σε όλα τα χρονικά βήματα. Δεδομένου ότι το MUSICGEN αντιμετωπίζει την παραγωγή ήχου ως μια εργασία ακολουθίας προς ακολουθία, η έννοια του χρόνου δεν αντιστοιχεί στην εφαρμογή μιας επαναληπτικής μεθόδου όπως στην περίπτωση των μοντέλων διάχυσης, αλλά στη δειγματοληψία νέων ηχητικών όρων. Έτσι, για να διασφαλίσουμε ότι η μέθοδός μας επηρεάζει ολόκληρο τον παραγόμενο ήχο, αυτή η προσαρμογή κρίθηκε απαραίτητη. Όπως αναφέρθηκε προηγουμένως, οι επεξεργασίες στο πλαίσιο του της αρχικής υλοποίησης του Prompt-to-Prompt, εφαρμόζονταν για έναν

καθορισμένο αριθμό επαναλήψεων της διαδικασίας διάχυσης. Αυτή η στρατηγική στόχευε να επιτύχει μια ισορροπία μεταξύ της δημιουργίας νέων δειγμάτων και της διατήρησης των βασικών χαρακτηριστικών του αρχικού. Αναγνωρίζοντας την αυτοπαλινδρομική φύση του MUSICGEN, εξετάζουμε μια εναλλακτική μέθοδο: τη μαλακή ανάμειξη. Αυτή η τεχνική συνδυάζει τους χάρτες χαρακτηριστικών που παράγονται με τους injected, χρησιμοποιώντας έναν σταθμισμένο μέσο όρο για την παραγωγή του χάρτη χαρακτηριστικών εξόδου. Ας θεωρήσουμε ότι X_t είναι ο χάρτης χαρακτηριστικών που παράγεται στο χρονικό βήμα t, και Y_t είναι ο injected χάρτης χαρακτηριστικών. Η τεχνική μαλακής ανάμειξης συνδυάζει αυτούς τους χάρτες χαρακτηριστικών χρησιμοποιώντας έναν σταθμισμένο μέσο όρο για να παράγει τον χάρτη χαρακτηριστικών εξόδου Z_t:

$$Z_t = aX_t + (1 - a)Y_t \tag{4}$$

όπου a είναι η παράμετρος ανάμειξης, με τιμές μεταξύ 0 και 1, και εκφράζεται ως:

$$a = \frac{i}{N}$$

όπου i αντιπροσωπεύει τον δείκτη του τρέχοντος στρώματος προσοχής που εξετάζεται και N αντιπροσωπεύει τον συνολικό αριθμό των στρωμάτων προσοχής στη στοίβα αποκωδικοποιητή. Αυτή η διατύπωση διασφαλίζει ότι ο παράγοντας ανάμειξης προσαρμόζεται δυναμικά με βάση τη θέση εντός της στοίβας αποκωδικοποιητή και αναπαράγει την αρχική προσέγγιση διάχυσης του Prompt-to-Prompt, όπου οι επεξεργασίες εφαρμόζονται για έναν καθορισμένο αριθμό επαναλήψεων της διαδικασίας διάχυσης.

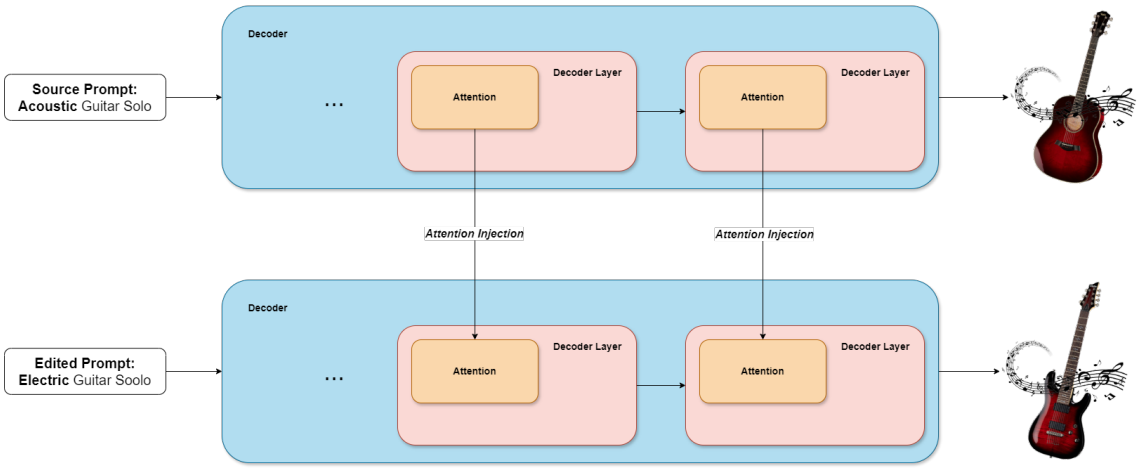


Figure 1. *Employing prompt-to-prompt throughout the decoder stack: The attention maps corresponding to the source prompt are injected into the forward process using the edited prompt at every decoder layer.*

Για την αξιολόγηση της μεθόδου μας, αρχικά δημιουργούμε ένα σύνολο δεδομένων που περιέχει ζεύγη εντολών για κάθε μηχανισμό επεξεργασίας: Αντικατάσταση, Διόρθωση και Ανακατανομή Βαρών. Κάθε ζεύγος αποτελείται από πρωτότυπες και επεξεργασμένες εντολές κειμένου. Η δημιουργία του συνόλου δεδομένων μας αποτελείται από δύο βήματα: (1)

δημιουργία ενός μικρού συνόλου ζευγών εντολών με το χέρι, και στη συνέχεια (2) χρήση του ChatGPT 3.5 για τη δημιουργία επιπλέον ζευγών. Αυτή η υβριδική προσέγγιση εξασφαλίζει ένα μείγμα χειροποίητων εντολών και δυναμικά παραγόμενων, παρέχοντας ένα ποικίλο εύρος προς αξιολόγηση. Εξετάζουμε διαφορετικούς άξονες επεξεργασίας ήχου για να οργανώσουμε αποτελεσματικά το σύνολο δεδομένων μας. Κάθε άξονας αντιπροσωπεύει μια διακριτή πτυχή του ηχητικού περιεχομένου:

Αλλαγή Οργάνου: Αντικατάσταση ενός οργάνου ή ηχητικής πηγής με άλλο, βελτιώνοντας τον ήχο με την προσθήκη λεπτομερών στοιχείων ή ανακατανέμοντας την έμφαση σε διάφορα ηχητικά στοιχεία. Αυτός ο άξονας επιτρέπει την εξερεύνηση διαφορετικών ηχοχρωμάτων, υφών και ηχητικών χαρακτηριστικών μέσα στη σύνθεση του ήχου.

Αλλαγή Διάθεσης/Τόνου: Η αλλαγή διάθεσης/τόνου περιλαμβάνει την τροποποίηση, αλλαγή ή βελτίωση της συναισθηματικής απήχησης και των τονικών αποχρώσεων της μουσικής. Αυτός ο άξονας περιλαμβάνει τροποποιήσεις που προκαλούν διαφορετικές συναισθηματικές αντιδράσεις ή αλλάζουν τη συνολική τονική χροιά του ηχητικού υλικού.

Αλλαγή Είδους: Η αλλαγή είδους περιλαμβάνει τη μετάβαση μεταξύ διαφορετικών μουσικών στυλ ή ειδών. Αυτός ο άξονας διευκολύνει την εξερεύνηση διαφορετικών στυλιστικών συμβάσεων, ρυθμικών μοτίβων και οργανικών διατάξεων σε διάφορα μουσικά είδη. Οι αλλαγές είδους προσφέρουν ευκαιρίες για δημιουργικό πειραματισμό και συγχώνευση ειδών.

Μελωδική Μεταμόρφωση: Η μελωδική μεταμόρφωση περιλαμβάνει την αλλαγή του μελωδικού περιεχομένου της μουσικής. Αυτός ο άξονας περιλαμβάνει τροποποιήσεις μελωδικών περιγραμμάτων, διαστημάτων, μοτίβων και θεμάτων.

Αρμονική Τροποποίηση: Αυτός ο άξονας περιλαμβάνει αλλαγές στις αρμονικές ακολουθίες, τον αρμονικό ρυθμό, την αρμονική πυκνότητα και την αρμονική ένταση, επιτρέποντας την αρμονική εμπλούτιση και την εξερεύνηση τονικών σχέσεων. Με την εξερεύνηση αρμονικών τροποποιήσεων, μπορούμε να διερευνήσουμε πώς οι αλλαγές στις αρμονικές ακολουθίες, τις αρμονικές φωνές και τις αρμονικές υφές επηρεάζουν τον αρμονικό χαρακτήρα και τη συναισθηματική απήχηση της μουσικής.

Παραλλαγή Μορφής/Δομής: Η παραλλαγή μορφής/δομής περιλαμβάνει παραλλαγές στη συνολική μορφή ή δομή της μουσικής. Αυτός ο άξονας περιλαμβάνει αλλαγές στη διαίρεση των τμημάτων, τις επαναλήψεις, τις μεταβάσεις και τις διαδικασίες ανάπτυξης, επιτρέποντας τον δομικό πειραματισμό και την αφηγηματική εξερεύνηση.

Χρησιμοποιώντας το MUSICGEN και την τεχνική Prompt-to-Prompt (όπως ορίζεται για τα αυτοπαλινδρομικά μοντέλα), δημιουργήσαμε 22 δείγματα ανά κατηγορία επεξεργασίας (Αντικατάσταση, Βελτίωση και Ανακατανομή Βαρών) με 5 τυχαίους σπόρους ανά ζεύγος εντολών. Αυτή η διαδικασία επαναλήφθηκε για το Auffusion, με αποτέλεσμα να παραχθούν συνολικά 660 δείγματα και από τα δύο μοντέλα.

Χρησιμοποιούμε πολλούς κοινούς δείκτες αξιολόγησης για την εκτίμηση των μουσικών χαρακτηριστικών των παραγόμενων δειγμάτων:

Ακρίβεια Μελωδίας: Αξιολογεί την ευθυγράμμιση των κλάσεων συχνοτήτων (C, C#, ..., B· σύνολο 12) σε βάση καρέ-καρέ μεταξύ του αρχικού ήχου και αυτού που προκύπτει από την εφαρμογή του Prompt-to-Prompt [33].

Συσχέτιση Δυναμικής: Αναφέρεται στη συσχέτιση Pearson μεταξύ των αρχικών τιμών δυναμικής σε βάση καρέ-καρέ και των τιμών που προκύπτουν από την εφαρμογή του

Configuration	T2A Similarity	A2A Similarity
Hard-blending	0.836 ∓ 0.087	0.400 ∓ 0.152
Soft-blending	0.849 ∓ 0.094	0.414 ∓ 0.157

Table 1. *Text-to-Audio & Audio-to-Audio Cosine Similarity with regards to blending strategy.*

Prompt-to-Prompt [33].

Rythm F1 Score: Μετρά την ευθυγράμμιση των εκτιμώμενων χρονικών σημείων των κτυπημάτων (beats) και των κύριων κτυπημάτων (downbeats) [34, 35]μεταξύ του αρχικού ήχου και του παραγόμενου ήχου από την εφαρμογή του Προμπτ-το-Προμπτ. Οι χρονικές στιγμές των κτυπημάτων/κύριων κτυπημάτων λαμβάνονται με την εφαρμογή ενός φίλτρου HMM [36] στις πιθανότητες που εκτιμώνται για κάθε καρέ. Η ευθυγράμμιση θεωρείται επιτυχής αν η διαφορά των χρονικών σημείων είναι μικρότερη από 70 χιλιοστά του δευτερολέπτου [35].

CLAP Score: [37, 38]Αξιολογεί την προσήλωση στην κειμενική εντολή υπολογίζοντας την συνημιτονοειδή ομοιότητα μεταξύ των embeddings κειμένου και του ήχου που εξάγονται από το μοντέλο CLAP. Το CLAP είναι ένα μοντέλο διπλού κωδικοποιητή με ξεχωριστούς κωδικοποιητές για τις εισόδους κειμένου και ήχου. Αυτοί οι κωδικοποιητές μαθαίνουν να ενσωματώνουν κείμενο και ήχο σε ένα κοινό χώρο [37, 38].

Διεξάγουμε ένα αρχικό πείραμα όπου συστηματικά μεταβάλλουμε τον αντίκτυπο *-ïσχύς εντολής¨-* των εγχυόμενων χαρτών προσοχής στην παραγωγή ήχου, αυξάνοντας σταδιακά την επίδραση των λεκτικών υποδείξεων στην επεξεργασία. Υπολογίζουμε τη μέση συνημίτονη ομοιότητα μεταξύ των αρχικών και επεξεργασμένων εντολών τόσο στο πλαίσιο Ήχος-προς-Ήχος όσο και Κείμενο-προς-Ήχο. Η γραφική παράσταση 2 δείχνει ότι ο επεξεργασμένος ήχος διατηρεί τις ιδιότητες του αρχικού, όπως υποδεικνύεται από την υψηλή μέση συνημίτονη ομοιότητα Ήχου-προς-Ήχο και τη μέση συνημίτονη ομοιότητα Κειμένου-προς-Ήχο με την αρχική εντολή, που παραμένει σταθερή ανεξαρτήτως της *ïσχύς εντολής¨*. Επιπλέον, η μέση συνημίτονη ομοιότητα Κειμένου-προς-Ήχο με την επεξεργασμένη εντολή, η οποία παραμένει σταθερή, αναδεικνύει ότι ο επεξεργασμένος ήχος παραμένει ευθυγραμμισμένος με την επεξεργασμένη εντολή ανεξαρτήτως της *ïσχύς εντολής¨*.

Επιπλέον, αξιολογούμε την αποτελεσματικότητα της μαλακής ανάμειξης χαρακτηριστικών διασταυρούμενης προσοχής ποσοτικά, υπολογίζοντας μετρικές συνημίτονης ομοιότητας Ήχου-προς-Ήχο και Κειμένου-προς-Ήχο μεταξύ των παραγόμενων δειγμάτων ήχου και της επεξεργασμένης εντολής. Οι τιμές συνημίτονης ομοιότητας Ήχου-προς-Ήχο και Κειμένου-προς-Ήχο υπολογίζονται ως μέσοι όροι σε ολόκληρο το σύνολο δεδομένων, διασφαλίζοντας μια συνολική και αντικειμενική αξιολόγηση. Όπως υποδεικνύεται από τον πίνακα 1, η χρήση της μαλακής ανάμειξης οδηγεί σε υψηλότερες μέσες τιμές συνημίτονης ομοιότητας Ήχου-προς-Ήχο και Κειμένου-προς-Ήχο. Σημειωτέον, η τυπική απόκλιση παραμένει σχετικά σταθερή, υπογραμμίζοντας την αξιοπιστία και συνέπεια της μεθόδου μαλακής ανάμειξης.

Τέλος, επιδιώκουμε να ερευνήσουμε την αποτελεσματικότητα του Prompt-to-Prompt όσον αφορά την επεξεργασία ήχου εξετάζοντας τόσο μοντέλα διάχυσης όσο και αυτοπαλινδρομικά μοντέλα. Αυτή η εξερεύνηση στοχεύει να προσφέρει γνώσεις για τα πλεονεκτήματα

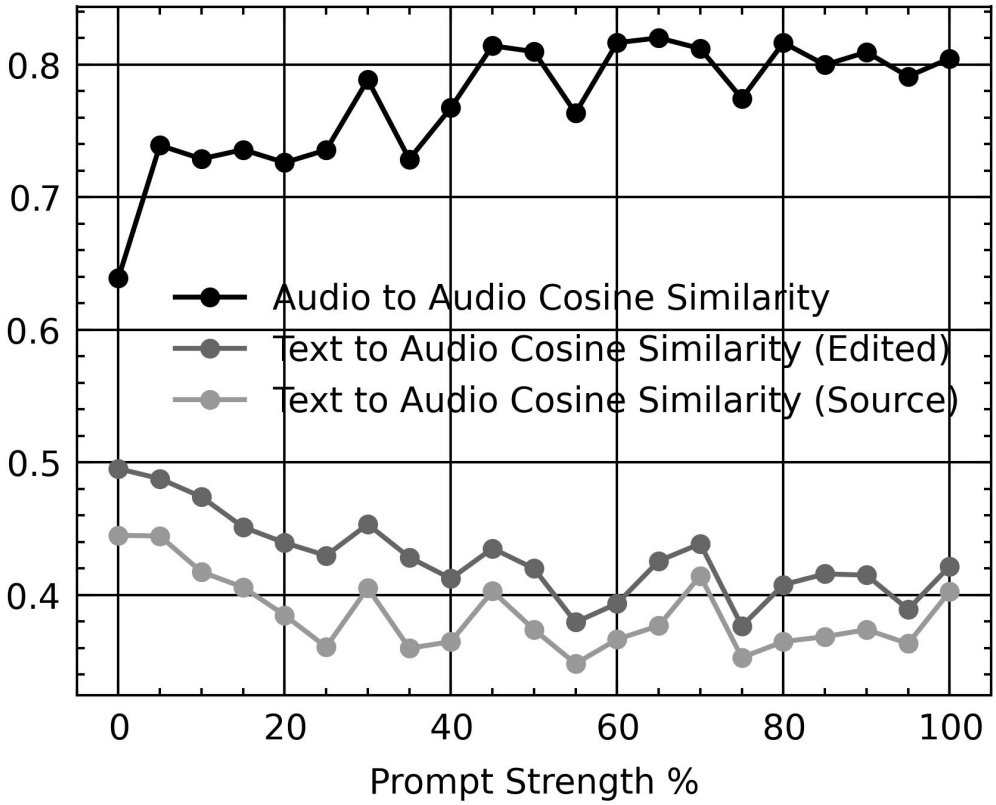


Figure 2. *Evaluation of audio and textual alignment with regards to "prompt strength".*

και τους περιορισμούς αυτών των μοντέλων σχετικά με τη επεξεργασία ήχου. Η βασική μας στρατηγική χρησιμοποιεί το Auffusion [10], μια προσέγγιση βασισμένη στη διάχυση. Αυτή η μέθοδος ενσωματώνεται με τη μεθοδολογία Prompt-to-Prompt, παρέχοντας μια βάση για την εξερεύνηση καθοδηγούμενης από εντολή επεξεργασίας ήχου. Επιπλέον, εισάγουμε μια εναλλακτική ενσωματώνοντας ένα προ-εκπαιδευμένο αυτοπαλινδρομικό μοντέλο, ονόματι ΜΥΣΙΓΕΝ [23].

Οι μετρικές υπολογίζονται ως μέσοι όροι σε ολόκληρο το σύνολο δεδομένων και παρουσιάζονται στη γραφική παράσταση 3 ενώ αναλυτικά αποτελέσματα παρουσιάζονται στον πίνακα 2. Όπως φαίνεται, το MUSICGEN υπερέχει του Auffusion σε όλες τις μετρικές αξιολόγησης. Διαπρέπει στην ακρίβεια μελωδίας, παρουσιάζει ανώτερη ομοιότητα τόσο με τον αρχικό ήχο όσο και με την εντολή στόχο, και ξεπερνά σημαντικά το Auffusion στη συσχέτιση δυναμικής και το Rythm F1 Score. Η μεθοδολογία μας, χρησιμοποιώντας την τεχνική Prompt-to-Prompt στο πλαίσιο ενός αυτοπαλινδρομικού μοντέλου με στόχο την επεξεργασίας ήχου, υπερέχει του Auffusion. Είναι σημαντικό να αναφερθεί πως πρόκειται γαι την πρώτη επιτυχή χρήση του Prompt-to-Prompt στο πλαίσιο επεξεργασίας ήχου με αυτοπαλινδρομικό μοντέλο.

Για να εκτιμήσουμε πόσο καλά η μέθοδός μας διατηρεί τη φυσικότητα και τη συνοχή με το αρχικό ηχητικό περιεχόμενο, καλέσαμε 24 αξιολογητές και διεξήγαμε μια μελέτη. Η μελέτη ξεκίνησε με τους συμμετέχοντες να αξιολογούν ζεύγη ηχητικών κλιπ διάρκειας 10 δευτερολέπτων. Καλούνταν να επιλέξουν το κλιπ σε κάθε ζεύγος που εμφάνιζε τον υψηλότερο βαθμό φυσικότητας (ομοιότητα με γνωστά μουσικά όργανα αντί για θόρυβο κλπ.). Αυτά τα

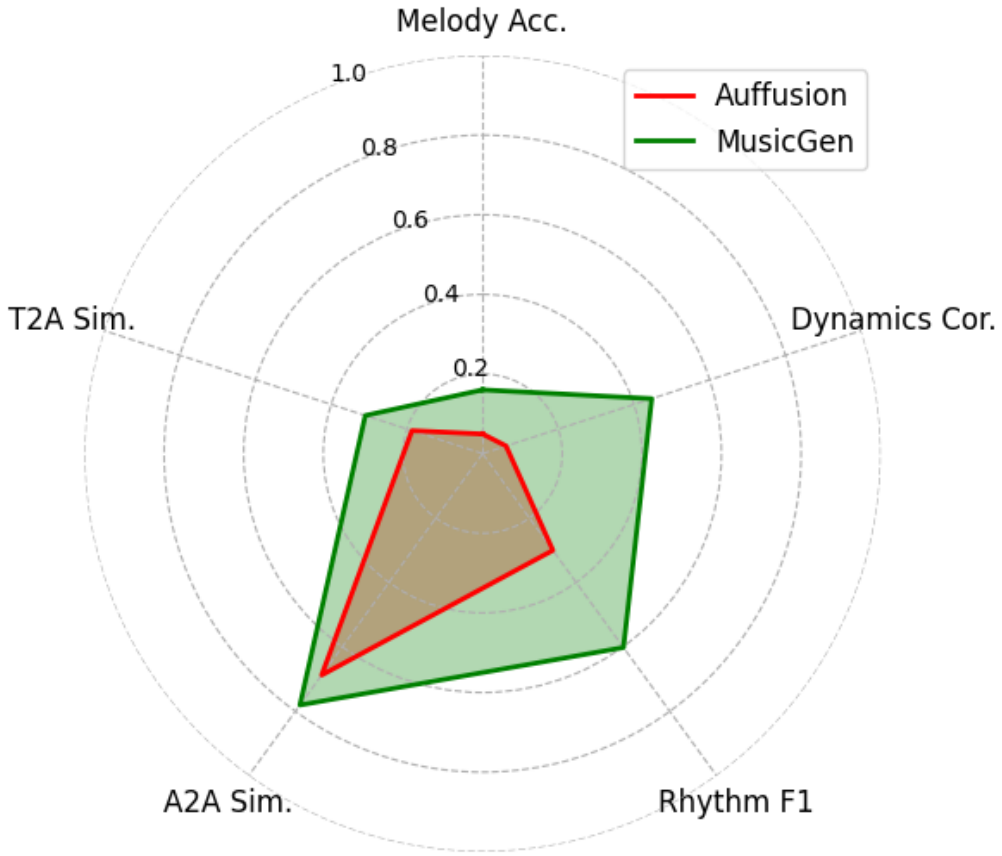


Figure 3. *Average evaluation metrics across editing mechanisms for both Auffusion and MUSICGEN.*

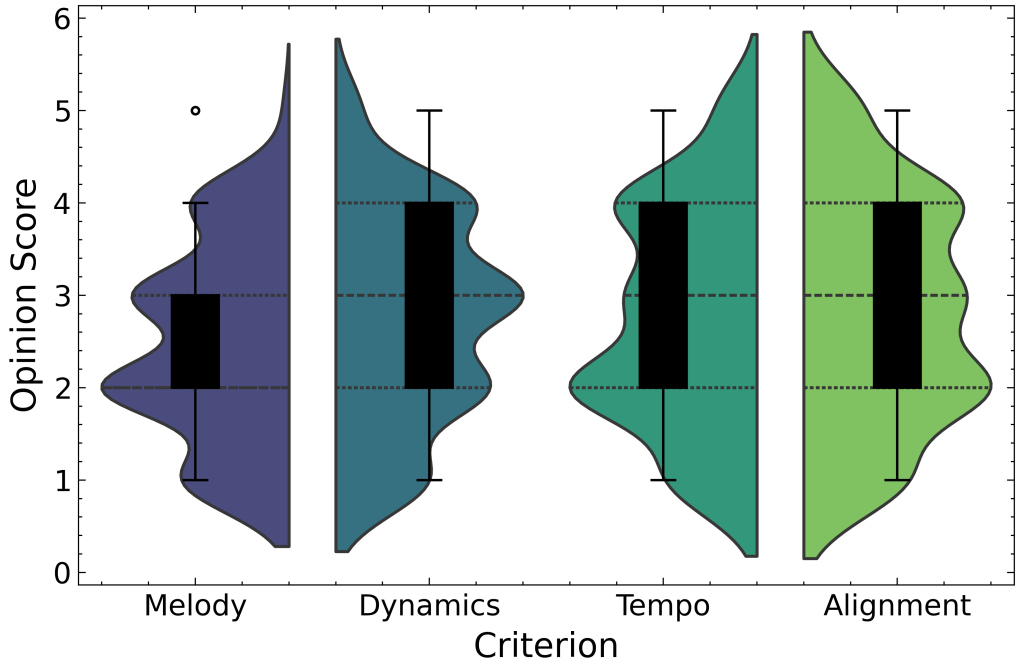


Figure 4. *Distribution of MOS (Mean opinion score) in the case of Auffusion per evaluation axis.*

Edit	Model	Alignment	Dynamics	Melody	Tempo
Refine	Auffusion	2.76	2.57	2.41	2.81
Refine	MUSICGEN	**3.43**	**3.08**	**3.22**	**3.62**
Replace	Auffusion	2.91	2.80	2.44	2.74
Replace	MUSICGEN	**3.05**	**3.07**	**3.16**	**3.56**
Reweight	Auffusion	2.78	2.88	2.58	2.78
Reweight	MUSICGEN	**3.54**	**3.24**	**3.38**	**3.86**

Table 2. *Comparison of audio editing capabilities of MUSICGEN and Auffusion based on MOS (Mean opinion score) of faithfulness.*

Edit	Model	Naturalness
Refine	Auffusion	35.14%
Refine	MUSICGEN	**64.86%**
Replace	Auffusion	40.00%
Replace	MUSICGEN	**60.00%**
Reweight	Auffusion	17.95%
Reweight	MUSICGEN	**82.05%**

Table 3. *Comparison of audio editing capabilities of MUSICGEN an**d Auffusion based** on MOS (Mean opinion score) of naturalness.*

κλιπ, προέρχονταν από την ίδια λεκτική εντολή αλλά είχαν παραχθεί από διαφορετικά μοντέλα – Auffusion και MUSICGEN. Στη συνέχεια, οι συμμετέχοντες αξιολόγησαν την πιστότητα 16 τυχαίων ζευγών κειμένου-ήχου στο σύνολο δεδομένων μας. Κάθε ζεύγος περιελάμβανε τόσο την αρχική λεκτική εντολή και τον ήχο όσο και μια επεξεργασμένη εκδοχή. Οι συμμετέχοντες αξιολόγησαν το κατά πόσο ο επεξεργασμένος ήχος παραμένει πιστός στην αρχική έκδοση, λαμβάνοντας υπόψη βασικά στοιχεία όπως η μελωδία, ο ρυθμός και η δυναμική, καθώς και την ευθυγράμμιση του με την επεξεργασμένη λεκτική εντολή, βαθμολογώντας αυτά τα χαρακτηριστικά σε μια κλίμακα Likert [39] από το 1 έως το 5. Βάσει των αποτελεσμάτων του πίνακα 3, η μέθοδός μας παράγει ηχητικά κλιπ που θεωρήθηκαν πιο φυσικά από τους συμμετέχοντες στη μελέτη μας. Οι Μέσοι Όροι Γνώμης (MOS) για κάθε κριτήριο απεικονίζονται στις γραφικες παραστάσεις 4 και 5. Για να εκτιμήσουμε τη στατιστική σημαντικότητα των αποτελεσμάτων μας, χρησιμοποιούμε ένα μη συζευγμένο t-test για τις κατανομές βαθμολογιών γνώμης. Ο πίνακας 4 συνοψίζει τα p-values που προκύπτουν, δείχνοντας ότι η βελτίωση που επιτεύχθηκε με τη χρήση της προτεινόμενης τεχνικής επεξεργασίας μουσικής σε συνδυασμό με το μοντέλο MUSICGEN ξεπερνά σημαντικά τη βάση.

Συνοψίζοντας, εξερευνήσαμε τη χρήση δύο μοντέλων για την επεξεργασία ήχου: Auffusion και MUSICGEN. Ξεκινήσαμε με το Auffusion, εκμεταλλευόμενοι τις υπάρχουσες δυνατότητές του για επεξεργασία βάσει οδηγιών. Επιπλέον, εισαγάγαμε μια εναλλακτική προσέγγιση ενσωματώνοντας το MUSICGEN, ένα προεκπαιδευμένο αυτοπαλινδρομικό μοντέλο γνωστό για τις προηγμένες δυνατότητές του. Για να ενσωματόσουμε το συγκεκριμένο μοντέλο με την τεχνική Prompt-to-Prompt, εφαρμόσαμε τη διαδικασία εισαγωγής χαρτών προσοχής σε όλα τα χρονικά βήματα. Επιπλέον, εισήγαμε την τεχνική της μαλακής ανάμειξης,

Melody	Dynamics	Tempo	Alignment
2.66×10^{-12}	2.32×10^{-4}	2.85×10^{-15}	3.76×10^{-06}

Table 4. *p-values, obtained using an unpaired t-test, indicating the statistical significance for the opinion scores obtained from the human study.*

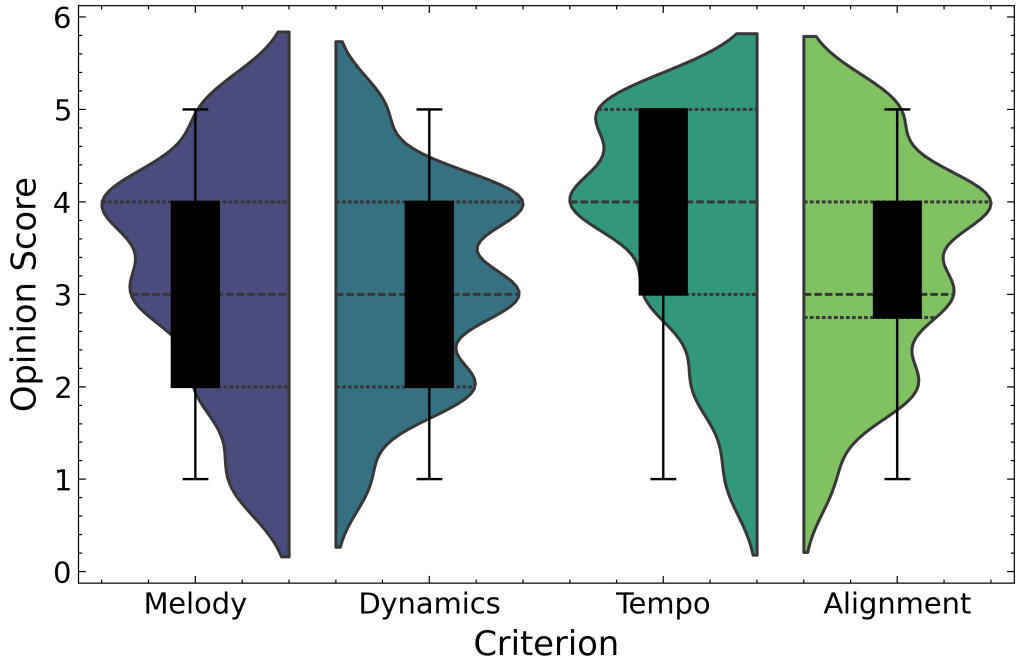


Figure 5. *Distribution of MOS (Mean opinion score) in the case of MUSICGEN per evaluation axis.*

η οποία ενώνει τους χάρτες χαρακτηριστικών που παράγονται με τους εισαγόμενους, χρησιμοποιώντας έναν ζυγισμένο μέσο όρο για την έξοδο. Αυτή η νέα μέθοδος στοχεύει στη βελτίωση της ποιότητας των παραγώμενων δειγμάτων αναπαράγοντας την αρχική προσέγγιση βασισμένη στη διάχυση, όπου η τεχνική Prompt-to-Prompt εφαρμόζεται για ένα προκαθορισμένο πλήθος επαναλήψεων. Στην αξιολόγησή μας, το MUSICGEN υπερτερεί του Auffusion βάσει όλων των μετρικών αξιολόγησης και σε όλες τις κατηγορίες επεξεργασίας. Συγκεκριμένα, εμφάνισε υψηλότερη ακρίβεια στην αναπαραγωγή μελωδιών, καλύτερη ομοιότητα τόσο με τον αρχικό ήχο όσο και με την νέα κειμενική εντολή, και ξεπέρασε σημαντικά το Auffusion στη δυναμική συσχέτιση και το Rythm F1 Score. Η μελέτη αυτή σηματοδοτεί την πρώτη επιτυχημένη εφαρμογή της τεχνικής Prompt-to-Prompt στο πλαίσιο επεξεργασίας ήχου με αυτοπαλινδρομικά μοντέλα.

Μελλοντικά, θα θέλαμε να χρησιμοποιήσουμε τη μέθοδό μας για τη δημιουργία ενός συνόλου δεδομένων επεξεργασίας ήχου που αποτελείται από ζεύγη κειμενικών εντολών και δεδομένων ήχου, παρέχοντας πολύτιμους πόρους για περαιτέρω έρευνα. Επιπλέον, στοχεύουμε να πραγματοποιήσουμε εκτενείς μελέτες χρηστών, επικεντρωμένες ιδιαίτερα στην ποικιλομορφία της εθνικής μουσικής και στα κοινωνικά πλαίσια, προκειμένου να αξιολογήσουμε την αποτελεσματικότητα και την ενσωμάτωση των προτεινόμενων τεχνικών επεξεργασίας ήχου σε διαφορετικές κουλτούρες. Ο μηχανισμός προσοχής του MUSICGEN, συγκεν-

τρώνει όλη την πληροφορία σε μία μόνο τιμή προσοχής. Η επέκταση αυτού του μηχανισμού θα επέτρεπε την περεταίρω διερεύνηση της αλληλεπίδρασης μεταξύ των διακριτικών tokens κειμένου και ήχου και θα μπορούσε να οδηγήσει σε σημαντικές βελτιώσεις στην ποιότητα επεξεργασίας ήχου.

Chapter 1

Introduction

Obtaining satisfactory outcomes in audio manipulation tasks has typically required large datasets with intricate annotations, a process that is often labor-intensive and costly. Moreover, crafting effective model architectures tailored to the nuances of audio processing requires substantial expertise and experimentation. Fine-tuning existing model architectures, while a possibility, remains a costly endeavor.

In this work, we employ prompt-guided models in the audio domain, and, specifically, we adapt the Prompt-to-Prompt [1] technique, previously successful in image manipulation, to audio editing. This approach allows for fine-grained, semantically meaningful audio editing without the need for model retraining or additional data. Prompt-to-Prompt leverages cross-attention maps to control the generation process, enabling users to influence how generated data units interact with tokens. This versatility enables tasks such as token value changes, global audio editing, and the manipulation of semantic effects without modifying the underlying model.

We apply Prompt-to-Prompt for the task of music editing and adapt it for an auto-regressive prompt-based model. This marks the first successful use of prompt-to-prompt in the auto-regressive model audio editing context. Our extensive evaluation demonstrates these techniques' potential for intuitive text-based audio editing. Our key contributions are:

1. We explore the utilization of a pre-trained frozen auto-regressive transformer model, initially designed for generating high-quality music samples from a given text prompt, for audio editing.

2. We design and implement three distinct audio editing mechanisms, inspired by those presented in [1].

3. We evaluate our auto-regressive transformer model-based approach against existing diffusion-based methods using automatic music-related metrics and feedback from users.

This work was submitted to ISMIR (International Society for Music Information Retrieval) 2024 conference. This document is organized as follows; *Part I* introduces the reader to basic concepts and ideas that are related to the problem at hand rather than

completely reviewing the domain. *Part II* provides a more formal description of the problem and elaborates on our technical approach. *Part III* presents the outcomes of our diverse experiments and offers conclusions on our work, highlighting the advantages and limitations of our framework, while also suggesting avenues for future research.

Part I

Background Knowledge

Chapter 2

Transformers

This chapter provides an introduction to fundamental concepts in deep learning, including artificial neurons, which serve as the foundational units of artificial neural networks (ANNs). The chapter also explores Feed Forward neural networks and the Single layer perceptron and finally the Transformer architecture, unraveling its complex attention mechanism.

Deep learning (also known as Deep Structured Learning) is part of a broader family of machine learning methods based on artificial neural networks with representation learning.

Representation learning or Feature Learning is a set of techniques that allows a system to automatically discover the representations needed for feature detection from raw data. This replaces manual feature engineering, which is the process of using domain knowledge to extract features (characteristics, properties, attributes) from raw data, and allows a machine to both learn the features and use them to perform a specific task. By the term feature, we refer to an individual measurable property or characteristic of a phenomenon. Features are usually numeric, but structural features such as strings and graphs can also be used. The concept of "feature" is related to that of explanatory variables used in statistical techniques such as linear regression.

Artificial neural networks (ANNs) were inspired by information processing and distributed communication nodes in biological systems. ANNs, though are quite different from biological brains. ANNs are comprised of an input layer, one or more hidden layers, and an output layer. Each node or artificial neuron has inputs and produces a single output that can be sent to multiple other neurons. The inputs can be the feature values of a sample of external data, such as images or documents, or they can be the outputs of other neurons. The outputs of the final output neurons of the neural net accomplish the task, such as recognizing an object in an image.

An ANN wherein connections between the nodes do not form cycles or loops is referred to as **Feed-Forward Neural Network**. The Feed-Forward Neural Network was the first and simplest type of artificial neural network devised. In this network, the information moves in only one direction, forward from the input nodes, through the hidden nodes (if any), and to the output nodes.

2.1 An Artificial Neuron

Artificial neurons are elementary units in an artificial neural network. The artificial neuron receives one or more inputs and sums them to produce an output. Each input is separately weighted, and the sum is passed through a non-linear function known as an activation function.

Other than the neuron's weights, another term is added to the total sum before being passed through the activation function. This term is the so-called **bias**. Bias allows you to shift the activation function, analogously to a constant in the context of a linear function, whereby the line is effectively transposed by the constant value.

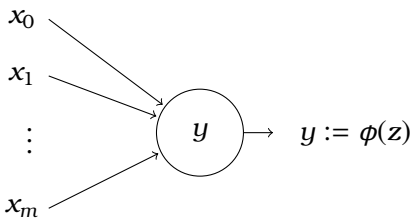


Figure 2.1. *An artificial neuron. This visualization was produced using code adapted from David Stutz's work [40].*

2.2 Single-Layer Perceptron Network (SLP)

The simplest kind of neural network is a **Single-Layer Perceptron Network**, which consists of a single layer of output nodes; the inputs are fed directly to the outputs via a series of weights. The sum of the products of the weights and the inputs is calculated in each node and passes through a, commonly non-linear, function. Single-layer perceptrons are only capable of learning linearly separable patterns.

For a given artificial neuron k, let there be $m + 1$ inputs with signals x_0 through x_m and weights $w_{k,0}$ through $w_{k,m}$. To achieve a bias-inclusive representation, the x_0 input is assigned the value +1 and corresponds to the neuron's bias, with $w_{k,0} = b_k$. Then the output of neuron k is given by the following equation:

$$y_k = \phi\left(\sum_{j=0}^{m} w_{k,j} x_j\right) \tag{2.1}$$

where ϕ stands for the activation function of choice. This operation is demonstrated by *Figure 2.1*, where k is left out as we are demonstrating the case of a single neuron.

2.3 Transformer

The Transformer architecture, introduced in [41], was originally employed as a sequence-to-sequence [42] model designed for machine translation. The Transformer has been successfully adopted in computer vision [43, 44, 45], audio processing [46, 47, 48], and other scientific fields [49, 50]. The transformer involves tokenizing text to numerical rep-

resentations, contextualizing tokens within a window of context, and employing parallel multi-head attention. This process emphasizes important tokens while diminishing the relevance of less significant ones, enabling the Transformer to capture extensive dependencies and contextual details within sequences effectively.

2.3.1 Self-Attention

The self-attention mechanism computes a set of attention scores, often referred to as attention weights, for each element in the input sequence. Given an input sequence of embeddings $X = \{x_1, x_2, \ldots, x_n\}$, the self-attention mechanism computes a set of attention scores $a_{i,j}$ for each pair of words i and j:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{2.2}$$

Here, Q, K, and V are the query, key, and value matrices obtained by linear projections of the input embeddings. The division by $\sqrt{d_k}$ helps stabilize the gradients during training, where d_k is the dimension of the key vectors. The output of the self-attention mechanism is a weighted sum of the values V, where the weights are determined by the attention scores.

2.3.2 Multi-Head Attention

To capture different aspects of the relationships between words, the self-attention mechanism is often extended to multiple heads. The outputs of these heads are concatenated and linearly projected as follows:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O \tag{2.3}$$

where $\text{head}_i = \text{Attention}(QW_{Qi}, KW_{Ki}, VW_{Vi})$ and W_O is the output projection matrix.

2.3.3 Positional Encoding

Since the model processes tokens in parallel rather than sequentially, it requires a mechanism to differentiate between the positions of tokens. In [41], Vaswani et al. introduced the concept of positional encoding. This involves adding positional encodings to the input embeddings of the tokens, allowing the model to discern the order of tokens in a sequence. The positional encoding is usually implemented using sine and cosine functions:

$$\text{PE}(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d}}\right) \tag{2.4}$$

$$\text{PE}(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d}}\right) \tag{2.5}$$

Here, $\text{PE}(pos, 2i)$ and $\text{PE}(pos, 2i+1)$ represent the positional encoding for the pos-th position and the $2i$-th and $2i+1$-th dimensions, respectively. The term d is the dimension

of the positional encoding. This positional encoding is then added to the input embeddings of the tokens. The resulting embeddings contain both the semantic information of the tokens and positional information, enabling the model to consider the order of tokens during processing.

2.3.4 The Encoder & Decoder Stacks

The complete Transformer model is composed of an encoder and a decoder. The encoder processes the input sequence, and the decoder generates the output sequence.

The encoder consists of a stack of $N = 6$ identical layers, each featuring a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. Residual connections and layer normalization surround each of these sub-layers, maintaining stability during training. The dimensions of the model outputs are set to $d_model = 512$ to facilitate these residual connections. On the other hand, the decoder, also comprising $N = 6$ layers, extends the encoder's design by introducing a third sub-layer for multi-head attention over the encoder's output. Residual connections and layer normalization persist, with a modification in the self-attention sub-layer to prevent positions from attending to subsequent positions. This alteration, coupled with output embeddings offset by one position, ensures that predictions at position i depend solely on known outputs at positions less than i.

The Transformer model leverages multi-head attention in distinct ways within the encoder and decoder stacks. In "encoder-decoder attention" layers, queries originate from the previous decoder layer, while memory keys and values come from the encoder output, allowing each decoder position to attend over all encoder positions. Encoder self-attention layers enable each position in the encoder to attend to all positions in the preceding encoder layer. Similarly, self-attention layers in the decoder facilitate each position's attention to all positions in the decoder, preserving an auto-regressive property. To prevent leftward information flow in the decoder, masking is applied to illegal connections, implemented within scaled dot-product attention by setting corresponding values in the softmax input to ∞.

This chapter explored artificial neurons and Single-Layer Perceptron Networks. These foundational concepts pave the way for grasping the Transformer architecture, a groundbreaking advancement in deep learning. The Transformer utilizes the attention mechanism to process sequential data efficiently, capturing complex dependencies within sequences.

Chapter 3

Diffusion Models

In this chapter, we delve into the mechanics of the diffusion process. Furthermore, we explore advancements in diffusion modeling, such as classifier-free guidance [51] and latent diffusion [52].

Sohl-Dickstein et al. introduced the diffusion probabilistic model [53], or diffusion model for short, which has dominated the task of image synthesis [54, 55, 56, 57] and has shown promise in various domains, ranging from computer vision [58, 59, 60] to natural language processing [61, 62] as well as other domains [63, 64, 65, 66, 67]. A diffusion model is a Markov chain with parameters trained via variational inference. It generates samples that match the data within a set timeframe. The chain's transitions are trained to reverse a diffusion process, gradually adding noise to the data opposite to the sampling direction until the signal is lost. When the diffusion involves Gaussian noise, setting the sampling chain transitions to conditional Gaussians simplifies the neural network parameterization. In the following sections, we describe the forward and reverse diffusion trajectories in more detail, based on the work of Ho et al. [55].

3.1 Forward Trajectory

Given a data distribution $q(x_0)$, the forward process consists of gradually converting $q(x_0)$ into a well-behaved distribution $\pi(y)$ by repeated application of a Markov diffusion kernel. The approximate posterior $q(x_{1:\tau} \mid x_0)$, called the forward trajectory, is fixed to a Markov chain that gradually adds Gaussian noise to the data:

$$q(x_{1:\tau} \mid x_0) = \prod_{t=1}^{\tau} q(x_t \mid x_{t-1}) \tag{3.1}$$

$$q(x_t \mid x_{t-1}) == \mathcal{N}(x_t;\ \sqrt{1-\beta_t}x_{t-1}, b_t\mathbf{I}) \tag{3.2}$$

where $\beta_1, \cdots \beta_\tau$ are the forward process variances. These can be learned by or held constant as hyperparameters. The forward process admits sampling x_t at an arbitrary timestep t in closed form as such:

$$q(x_t \mid x_0) = \mathcal{N}(x_t;\ \sqrt{\hat{a}_t}x_0, (1-\hat{a}_t)\mathbf{I}) \tag{3.3}$$

$$\hat{a}_t = \prod_{s=1}^{t} a_s \tag{3.4}$$

$$a_t = 1 - b_t \tag{3.5}$$

3.2 Reverse Trajectory

The reverse trajectory $p_\partial(x_{0:T})$ is defined as a Markov chain with learned Gaussian transitions starting at $p_\partial(x_\tau) = \mathcal{N}(x_t; 0, \mathbf{I})$:

$$p_\partial(x_{0:T}) = p_\partial(x) \prod_{t=1}^{\tau} p_\partial(x_{t-1} \mid x_t) \tag{3.6}$$

$$p_\partial(x_{x-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\partial(x_t, t), \sigma_\partial(x_t, t)) \tag{3.7}$$

Training a DPM amounts to maximizing its log-likelihood. Given a data point x_0 a DPM assigns it a probability equal to:

$$p_\partial(x_0) = \int p_\partial(x_0 : T) dx_{1:\tau} \tag{3.8}$$

$$= \int p_\partial(x_{0:\tau}) \frac{q(x_{1:\tau} \mid x_0)}{q(x_{1:\tau} \mid x_0)} dx_{1:\tau} \tag{3.9}$$

$$= \int q(x_{1:\tau} \mid x_0) \frac{p_\partial(x_{0:\tau})}{q(x_{1:\tau} \mid x_0)} dx_{1:\tau} \tag{3.10}$$

$$= \int q(x_{1:\tau} \mid x_0) dx_{1:\tau} \tag{3.11}$$

$$= p_\partial(x_\tau) \prod_{t=1}^{\tau} \frac{p_\partial(x_{t-1} \mid x_t)}{q(x_t \mid x_{t-1})} \tag{3.12}$$

Instead of maximizing the log-likelihood of the model Sohl-Dickstein et al. [53] used the negative log-likelihood L as a cost/score function and optimized the model using a conventional gradient descent approach. The negative log-likelihood L is defined as follows:

$$L = \mathrm{E}_q\left[-\log \frac{p_\partial(x_{0:\tau})}{q(x_{1:\tau} \mid x_0)}\right] \tag{3.13}$$

$$= \mathrm{E}_q\left[-\log p_\partial(x_\tau) - \sum_{t\geq 1} \log \frac{p_\partial(x_{t-1} \mid x_t)}{q(x_t \mid x_{t-1})}\right] \tag{3.14}$$

$$= \mathrm{E}_q\left[-\log p_\partial(x_\tau) - \sum_{t>1} \log \frac{p_\partial(x_{t-1} \mid x_t)}{q(x_t \mid x_{t-1})} - \log \frac{p_\partial(x_0 \mid x_1)}{q(x_1 \mid x_0)}\right] \tag{3.15}$$

$$= \mathrm{E}_q\left[-\log p_\partial(x_\tau) - \sum_{t>1} \log \frac{p_\partial(x_{t-1} \mid x_t)}{q(x_{t-1} \mid x_t, x_0)} \cdot \frac{q(x_{t-1} \mid x_0)}{q(x_t \mid x_0)} - \log \frac{p_\partial(x_0 \mid x_1)}{q(x_1 \mid x_0)}\right] \tag{3.16}$$

$$= \mathrm{E}_q\left[-\log \frac{p_\partial(x_\tau)}{q(x_\tau \mid x_0)} - \sum_{t>1} \log \frac{p_\partial(x_{t-1} \mid x_t)}{q(x_{t-1} \mid x_t, x_0)} - \log p_\partial(x_0 \mid x_1)\right] \tag{3.17}$$

L has a lower bound provided by Jensen's inequality:

$$L = \mathrm{E}_q\left[-\log \frac{p_\partial(x_{0:\tau})}{q(x_{1:\tau} \mid x_0)}\right] \tag{3.18}$$

$$\geq \mathrm{E}_q\left[-\log p_\partial(x_0)\right] \tag{3.19}$$

$$= K \tag{3.20}$$

Training consists of finding the reverse Markov transitions which maximize this lower bound K on the log-likelihood. Given that we can sample the forward process at an arbitrary timestep t, we optimize random terms of L with stochastic gradient descent, yielding much more efficient training.

L can be rewritten in terms of KL divergence as such:

$$\mathrm{E}_q\left[D_{KL}(q(x_\tau \mid x_0) \parallel p_\partial(x_\tau)) + \sum_{t>1} D_{KL}(q(x_{t-1} \mid x_t, x_0) \parallel p_\partial(x_{t-1} \mid x_t)) - \log p_\partial(x_0 \mid x_1)\right] \tag{3.21}$$

The forward process posteriors are tractable when conditioned on x_0:

$$q(x_{t-1} \mid x_t, x_0) = \mathcal{N}(x_{t-1}; \hat{\mu}_t(x_t, x_0), \hat{\beta}_t \mathbf{I}) \tag{3.22}$$

$$\hat{\mu}_t(x_t, x_0) = \frac{\sqrt{\hat{a}_{t-1}}\hat{\beta}_t}{1-\hat{a}_t} x_0 + \frac{\sqrt{a_t}(1-\hat{a}_{t-1})}{1-\hat{a}_t} x_t \tag{3.23}$$

$$\hat{\beta}_t = \frac{1-\hat{a}_{t-1}}{1-\hat{a}_t}\beta_t \tag{3.24}$$

All KL divergences are comparisons between Gaussians, so they can be calculated in a Rao-Blackwellized fashion with closed-form expressions instead of high-variance Monte Carlo estimates.

In their work [55], the authors opt to set the forward process variances β_t as constants rather than learnable parameters. Consequently, the approximate posterior q

lacks any learnable parameters, leading to $E_q[D_{KL}(q(x_T \mid x_0) \parallel p_\theta(x_\tau))]$ remaining constant throughout training, and thus negligible. Furthermore, for the reverse process, the authors set $\Sigma_\theta = \sigma_t^2 \mathbf{I}$ as untrained time-dependent constants. Similar outcomes can be achieved by setting $\sigma_t^2 \beta_t$ and $\sigma_t^2 \hat{\beta}_t$. Moreover, considering the analysis of L_t with $p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 \mathbf{I})$, we have

$$E_q[D_{KL}(q(x_{t-1} \mid x_t, x_0) \parallel p_\theta(x_{t-1} \mid x_t))] = E_q\left[\frac{1}{2\sigma_t^2} \|\hat{\mu}_\theta(x_t, x_0) - \mu_\theta(x_t, t)\|^2 + \mathbf{C}\right] \tag{3.25}$$

where $\mathbf{C}$ is a constant independent of θ. The most direct parameterization of μ_θ would involve a model that predicts $\hat{\mu}_t$.

3.3 Classifier-free guidance

To enhance the quality of samples generated by diffusion models, previous work [68] introduced classifier guidance, which involved leveraging an additional trained classifier. This technique enabled the generation of low-temperature samples, a capability that was previously unknown in diffusion models as simple approaches such as scaling model score vectors or reducing Gaussian noise during diffusion sampling proved ineffective. Classifier guidance mixed a diffusion model's score estimate with the input gradient of a classifier, allowing for a trade-off between Inception Score (IS) and Fréchet Inception Distance (FID). Building upon this, [51] introduced classifier-free guidance, eliminating the need for an additional classifier. Instead of sampling based on the gradient of an image classifier, classifier-free guidance combines the score estimates of a conditional diffusion model and a jointly trained unconditional diffusion model. By adjusting the mixing weight, they achieve a similar FID/IS trade-off as with classifier guidance. Their results demonstrate that pure generative diffusion models can produce highly realistic samples, showcasing the effectiveness of classifier-free guidance in synthesizing high-fidelity images.

Both classifier and classifier-free guidance alter the original denoising objective:

$$\epsilon_\theta(z_\lambda, c) \approx -\sigma_\lambda \nabla_{z_\lambda} \log p(z_\lambda | c) \tag{3.26}$$

$$\min_\theta \mathbb{E} z_0, \epsilon \sim \mathcal{N}(0, I), t \sim \text{Uniform}(1, T) |\epsilon - \epsilon\theta(z_t, c)|_2^2 \tag{3.27}$$

Dhariwal et al. [68] introduced a method to achieve a truncation-like effect in diffusion models. This method involves incorporating classifier guidance into the diffusion score. Specifically, they modify the diffusion score to include the gradient of the log-likelihood of an auxiliary classifier model, denoted as $p_\theta(c|z_\lambda)$. This modification results in a new score:

$$\tilde{\epsilon}_\theta(z_\lambda, c) = \epsilon_\theta(z_\lambda, c) - \omega\sigma_\lambda \nabla_{z_\lambda} \log p_\theta(c|z_\lambda) \tag{3.28}$$

$$\approx -\sigma_\lambda \nabla_{z_\lambda}[\log p(z_\lambda|c) + \omega \log p_\theta(c|z_\lambda)] \tag{3.29}$$

where ω is a parameter that controls the strength of the classifier guidance. $\tilde{\epsilon}_\partial(z_\lambda, c)$ is then used in place of $\epsilon_\partial(z_\lambda, c)$ when sampling from the diffusion model. The key impact of this modification is to give higher importance to data that the classifier $p_\partial(c|z_\lambda)$ assigns a high likelihood to the correct label. In other words, data that can be effectively classified receives a higher score on the Inception score of perceptual quality. This design effectively rewards generative models for their ability to produce data that can be accurately classified, thereby enhancing the quality of generated content.

While classifier guidance successfully balances the trade-off between IS and FID, as anticipated in truncation or low-temperature sampling, it remains reliant on gradients derived from an image classifier. The integration of classifier guidance introduces complexities into the training pipeline of diffusion models, as it necessitates the training of an additional classifier. Moreover, this classifier must be trained on noisy data, making it generally impractical to employ a pre-trained classifier in this context. Instead of opting for the training of a distinct classifier model, classifier-free guidance [51] introduced by Ho et al. involves the training of two interconnected models: an unconditional denoising diffusion model $p_\partial(z)$, parameterized by a score estimator $\epsilon_\partial(z_\lambda)$, and a conditional model $p_\partial(z|c)$, parameterized by $\epsilon_\partial(z_\lambda, c)$. Both models share a common neural network for parameterization. In the case of the unconditional model, we can simply input a null token $\emptyset$ as the class identifier c when predicting the score, denoted as $\epsilon_\partial(z_\lambda) = \epsilon_\partial(z_\lambda, c = \emptyset)$. To facilitate the training process, we jointly train both the unconditional and conditional models. This is achieved by randomly assigning the unconditional class identifier $\emptyset$ to c with a specified probability, which is set as a hyperparameter. During sampling, we utilize a linear combination of the conditional and unconditional score estimates as follows:

$$\tilde{\epsilon}_\partial(z_\lambda, c) = (1 + \omega)\epsilon_\partial(z_\lambda, c) - \omega\epsilon_\partial(z_\lambda) \tag{3.30}$$

The results demonstrate that classifier-free guidance has the capacity to balance the trade-off between FID and IS, much like classifier guidance. In the absence of a classifier gradient, moving in the direction of $\epsilon_\partial(z_\lambda, c)$ cannot be construed as a gradient-based adversarial attack on an image classifier.

3.4 Latent Diffusion

Rombach et al. [52] suggested moving from the high-dimensional image space to a learned latent space with an autoencoding model. This makes diffusion models more computationally efficient by sampling in a lower-dimensional space. This involves a two-stage training process for perceptual image compression and latent diffusion. In the first stage, an autoencoder is trained to create a lower-dimensional, perceptually equivalent latent space. Rombach et al. employ a strategy that integrates perceptual loss and patch-based adversarial objectives, ensuring reconstructions adhere to the image manifold. In the second stage, termed "Latent Diffusion," a Diffusion Probabilistic Model (DPM) is trained on the learned lower-dimensional latent space from the autoencoder, instead of the high-dimensional pixel space. This not only makes training scalable but also facilitates efficient

image generation with a single network pass. The autoencoder, trained in the first stage, transforms the high-dimensional RGB image into a compressed two-dimensional representation, allowing the DPM to work in a more suitable, computationally efficient space. In [52], the model is extended to a conditional image generator by incorporating a cross-attention mechanism and introducing another encoder, which transforms text prompts into an intermediate representation for the UNet layers. The UNet [69] comprises a contracting path for feature extraction, a bottleneck to retain essential information, and an expansive path for precise localization. It incorporates temporal conditioning, giving rise to the concept of a time-conditional UNet. In a time-conditional UNet, the architecture is adapted to consider information across different time steps, making it applicable to tasks where the input data has a temporal dimension.

Diffusion models have proven their effectiveness in producing realistic samples across various domains. Classifier-free guidance improves sample quality without needing an extra classifier. Meanwhile, latent diffusion employs auto-encoding techniques to make computations more efficient without compromising sample quality.

Chapter 4

State of the Art

In this chapter, we delve into the state-of-the-art techniques in audio generation, focusing on diffusion-based and auto-regressive models, as well as editing techniques.

4.1 Diffusion-based audio generation

Diffusion-based models have been widely explored for generation tasks in the audio and music domains. Yang et al. [2] employ a Vector Quantised-Variational AutoEncoder (VQ-VAE) [17] model trained on mel-spectrograms to convert them into discrete codes. These codes are then fed into a diffusion model to generate audio signals. Make-An-Audio [3] employs a spectrogram autoencoder, integrates Contrastive Language-Audio Pretraining (CLAP) [4] and introduces a pseudo prompt enhancement approach, aligning natural languages with audio data, enabling utilization of vast unsupervised language-free data. AudioLDM [5] employs a Latent Diffusion Model (LDM) and addresses the limitations of paired data methods by training generative models exclusively with audio data. AudioLDM 2 [6] introduces the Language of Audio (LOA), employing AudioMAE [7] pre-trained on diverse audio content. AUDIT [8] combines latent diffusion models with human instructions to generate edited audio segments using both audio and text cues. With regards to music, TANGO [9], inspired by latent diffusion models (LDM) and AudioLDM, utilizes a Large Language Model (LLM) instead of CLAP-based embeddings. Auffusion[10] employs a pretrained Latent Diffusion Model (LDM) and HiFi-GAN [11] vocoder. Additionally, the authors introduce a cross-attention mechanism that enhances alignment and flexibility. MusicLDM [12], adapts Stable Diffusion and AudioLDM architectures to the music domain. To address the challenge of limited training data, novel mixup strategies are proposed: beat-synchronous audio mixup (BAM) and beat-synchronous latent mixup (BLM). InstructME [13] leverages a tailored latent diffusion model, facilitating tasks such as adding, removing, and remixing musical elements while preserving harmonic integrity via chord progression matrices.

4.2 Auto-regressive audio generation

As an alternative to diffusion-based models for audio and music generation, autoregressive models have shown promise in recent years. WaveNet [14] introduced an auto-

regressive classification method for speech synthesis, surpassing traditional concatenative and parametric approaches, albeit with slower inference. AudioGen [15], has outperformed Diffsound by employing auto-regressive modeling in discrete waveform spaces. Jukebox [16] uses a multi-scale VQ-VAE [17] to compress raw audio into discrete codes, which are then modeled using auto-regressive transformers. AudioLM [18] utilizes tokens generated by a SoundStream [19] neural codec [20, 21] as targets for a sequence modeling task. In [22], Agostinelli et al. introduce MusicLM, following a similar approach to AudioLM but tailored for music editing tasks. MUSICGEN, a transformer-based decoder introduced in [23], employs EnCodec [24] for audio tokenization. Text conditioning integrates techniques like T5 encoder, FLAN-T5, and CLAP, while melody conditioning utilizes dominant time-frequency bins to control melodic structure. The model also introduces a framework for codebook interleaving patterns, improving efficiency and flexibility.

4.3 Editing techniques

Prompt-to-Prompt [1] leverages internal cross-attention layers to control which pixels attend to which tokens, allowing tasks like token value changes, global image editing, and semantic effects amplification/attenuation without model retraining. Textual Inversion [25] utilizes a sequence of noised latent codes obtained from initial Denoising Diffusion Implicit Model (DDIM) [70] inversion as a pivot and optimizes the null-text embedding. By fine-tuning a pre-trained text-to-image model with a few images of a subject, Dreambooth [26] associates a unique identifier with the object. Leveraging the semantic prior embedded in the model along with a new autogenous class-specific prior preservation loss, Dreambooth enables the synthesis of photorealistic images of the subject contextualized in various scenes. In [27] the authors proposed Custom Diffusion, where they optimize only a few parameters in the text-to-image conditioning mechanism to represent new concepts. This approach performs on par with or better than existing methods while maintaining computational efficiency. SVDiff [28], fine-tunes the singular values of the weight matrices resulting in a compact parameter space, reducing the risk of overfitting and language-drifting [29]. [30] employs Textual Inversion [25] and Dreambooth [26] to personalize the outputs of AudioLDM for newly learned musical concepts in a few-shot manner. In [31], the authors investigate two zero-shot audio editing techniques utilizing Denoising Diffusion Probabilistic Models (DDPMs) [55] inversion on pre-trained diffusion networks. Their approach, based on [32], involves extracting latent noise vectors corresponding to the source signal and using these vectors in a DDPM sampling process to guide the diffusion toward the desired edit. For text-based editing, they adjust the text prompt given to the denoiser model. In the unsupervised scenario, they perturb the denoiser output along the directions of the top principal components (PCs) of the posterior.

Building upon the achievements of Prompt-to-Prompt and the superior audio quality of MUSICGEN, we aim to integrate them into an auto-regressive framework. This will blend Prompt-to-Prompt's editability with the high-quality sound of modern auto-regressive models like MUSICGEN.

Part II

Methodology

Chapter 5

Methodology

5.1 MUSICGEN

MUSICGEN [23] employs EnCodec [24], a convolutional auto-encoder utilizing Residual Vector Quantization (RVQ) for latent space quantization. The input, a reference audio random variable X, is encoded into a continuous tensor with a lower frame rate (f_r) compared to the sample rate (f_s). The continuous representation is then quantized into discrete tokens (Q) using RVQ, resulting in K parallel sequences (for each time step), each with T tokens, where K is the number of codebooks, and M is the codebook size. The authors employ an autoregressive decomposition approach that predicts multiple codebooks simultaneously, and thusly greatly accelerates both training and inference. More specifically, MusicGen employs a token interleaving pattern to generate all codebooks in a single decoder pass, eliminating the need for cascading multiple models and making inference much faster.

5.2 Prompt-to-Prompt

Consider an audio sample A generated using a text prompt P. By injecting the attention maps obtained during A's generation into a new generation with a modified prompt P^*, we can perform a meaningful edit resulting in a new audio sample A^* that preserves the original's structure. To address specific editing operations, we employ three editing mechanisms akin to the ones introduced in [1]:

Replace: The user swaps tokens of the original prompt with others. For example replacing an acoustic with an electric guitar (Figure 5.1). We inject the attention maps of the source sample into the generation process with the modified prompt:

$$Edit(M_t, M_t^*, t) = \begin{cases} M_t^*, & \text{if } t < \tau \\ M_t, & otherwise \end{cases} \tag{5.1}$$

where τ is a timestamp parameter that determines until which step the injection is applied.

Refine: The user adds new tokens to the prompt. In this case, attention injection is applied only to the common tokens shared by both prompts. Formally, we utilize an alignment function A that takes a token index from the target prompt P^* and outputs the

corresponding token index in the original prompt P or $\emptyset$ if there isn't a match:

$$(Edit(M_t, M_t^*, t))_{i,j} = \begin{cases} (M_t^*)_{i,j}, & \text{if } A(j) = \emptyset \\ (M_t)_{i,A(j)}, & otherwise \end{cases} \tag{5.2}$$

It's worth recalling that index i corresponds to a value, while index j corresponds to a text token. Once again, we may set a timestamp τ to control the number of steps in which the injection is applied.

Reweight: Finally, the user may wish to strengthen or weaken the extent to which each token affects the result. To achieve this manipulation, we scale the attention map of the assigned token j^* with a parameter c ranging from -2 to 2, resulting in a stronger or weaker effect. The attention maps for the other tokens remain unchanged:

$$(Edit(M_t, M_t^*, t))_{i,j} = \begin{cases} c \cdot (M_t)_{i,j}, & \text{if } j = j* \\ (M_t)_{i,j}, & otherwise \end{cases} \tag{5.3}$$

In the original implementation of Prompt-to-Prompt, utilizing text-guided diffusion models [71], the output was decided early in the diffusion process. By restricting the number of injection steps τ, the authors managed to steer the generation process while maintaining flexibility in adapting the geometry to the new prompt. To align Prompt-to-Prompt with MUSICGEN's auto-regressive features, we apply the attention injection procedure at all timesteps. Since MUSICGEN treats audio generation as a sequence-to-sequence task, the notion of time doesn't correspond to the application of an iterative method like in the case of diffusion models, but rather to sampling new audio tokens. Thus, to guarantee that edits affect the entirety of the generated audio, this adjustment was deemed essential. As mentioned earlier, edits in the context of the original Prompt-to-Prompt paper, tailored for diffusion models, were applied for a set number of iterations of the diffusion process. This strategy aimed to strike a balance between generating novel samples and preserving the essential characteristics of the original. Acknowledging the auto-regressive nature of MUSICGEN, we investigate an alternative method: soft-blending. This technique merges the feature maps generated during the forward process with the injected ones, employing a weighted average for the output. Let X_t denote the feature map generated at time step t during the forward process, and let Y_t represent the injected feature map. The soft-blending technique combines these feature maps using a weighted average to produce the output feature map Z_t:

$$Z_t = aX_t + (1 - a)Y_t \tag{5.4}$$

where a is the blending parameter, with values between 0 and 1, and is expressed as:

$$a = \frac{i}{N}$$

where i represents the index of the current attention layer being considered and N represents the total number of attention layers in the decoder stack. This formulation ensures that the blending factor dynamically adjusts based on the position within the decoder

stack and replicates the original diffusion-based approach of prompt-to-prompt, where edits are applied for a set number of diffusion iterations.

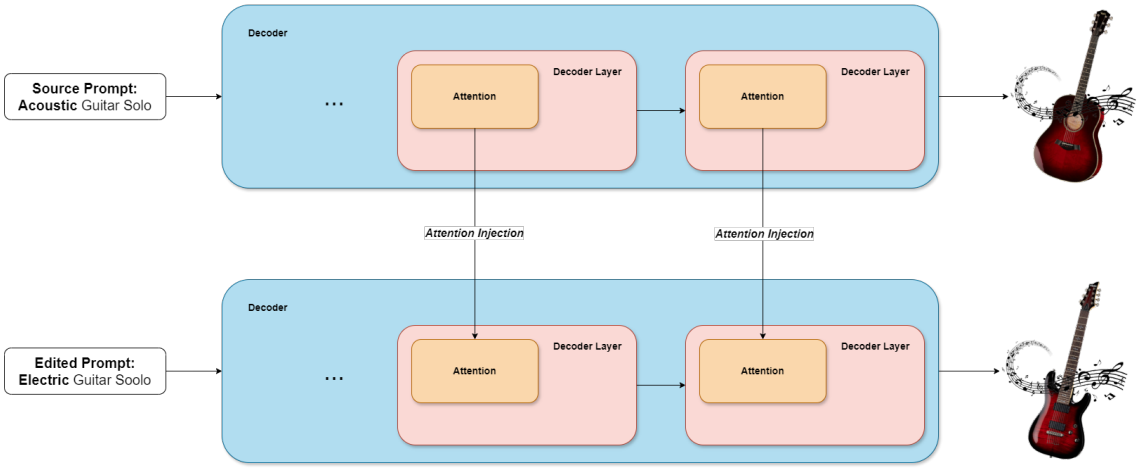


Figure 5.1. *Employing prompt-to-prompt throughout the decoder stack: The attention maps corresponding to the source prompt are injected into the forward process using the edited prompt at every decoder layer.*

Part III

Results

Chapter 6

Experimental Setup

6.1 Dataset construction

To evaluate our method, we initially create a dataset containing prompt pairs for each editing mechanism: Replace, Refine, and Reweight. Each pair consists of original and edited text prompts. Creating our dataset consists of two steps: (1) handpicking a small group of prompt pairs, followed by (2) leveraging ChatGPT 3.5 to generate additional pairs. This hybrid approach ensures a blend of manually crafted prompts and dynamically generated ones, providing a diverse range of inputs for evaluation. We consider different audio editing axes to organize our dataset effectively. Each axis represents a distinct aspect of audio content, offering unique opportunities for creative expression and artistic exploration:

- **Instrument Change**: Substituting one instrument or sound source with another, enhancing the audio by adding nuanced details, or recalibrating the emphasis on various sonic elements. This axis enables exploring diverse timbres, textures, and sonic characteristics within the audio composition.
- **Mood/Tonal Change**: Mood/tonal change involves altering, changing, or enhancing the emotional resonance and tonal nuances of the music. This axis encompasses modifications that evoke different emotional responses or shift the overall tonal coloration of the audio material.
- **Genre Shift**: Genre shift entails transitioning between different musical styles or genres. This axis facilitates the exploration of diverse stylistic conventions, rhythmic patterns, and instrumental arrangements across various musical genres. Genre shifts offer opportunities for creative experimentation and genre-blending.
- **Melodic Transformation**: Melodic transformation involves altering the melodic content of the music. This axis encompasses modifications to melodic contours, intervals, motifs, and themes, allowing for creative reinterpretations and variations of melodic material.
- **Harmonic Modification**: This axis encompasses changes in chord progressions, harmonic rhythm, harmonic density, and harmonic tension, allowing for harmonic

enrichment and exploration of tonal relationships. By exploring harmonic modifications, we can investigate how changes in harmonic progressions, chord voicings, and harmonic textures influence the harmonic character and emotional resonance of the music.

- **Form/Structure Variation**: Form/structure variation involves variations in the overall form or structure of the music. This axis encompasses changes in sectional arrangement, repetitions, transitions, and developmental processes, allowing for structural experimentation and narrative exploration.

Using MUSICGEN and Prompt-to-Prompt (as defined for autoregressive models), we generated 22 samples per edit category (Replace, Refine, and Reweight) with 5 random seeds per prompt pair. This process was repeated for Auffusion resulting in a total of 660 generated samples across both models.

6.2 Automated music coherence evaluation metrics

We employ multiple common evaluation metrics to assess the musical characteristics of the generated samples, including time-varying controllability, adherence to global text control, and overall audio realism:

- **Melody Accuracy**: Assesses the alignment of pitch classes (C, C#, ..., B; totaling 12) on a frame-by-frame basis between the source audio and the one derived from the application of Prompt-to-Prompt [33].
- **Dynamics Correlation**: Refers to Pearson's correlation between the source dynamics values on a frame-by-frame basis and the values derived from the application of Prompt-to-Prompt [33].
- **Rhythm F1 Score**: Adheres to the conventional approach to detecting beats and downbeats [34, 35], measuring the synchronization between the estimated timestamps of beats/downbeats derived from the source rhythm control and those generated from the application of Prompt-to-Prompt. Timestamps are obtained by applying an HMM postfilter [36] to the frame-wise probabilities of beats/downbeats (i.e., the rhythm control signal). Following [35], alignment between input and generated timestamps is considered if their difference is less than 70 milliseconds.
- **CLAP Score**: [37, 38] assesses text control adherence by calculating the cosine similarity between text and audio embeddings extracted from the CLAP model. CLAP is a dual-encoder foundation model with separate encoders for text and audio inputs. These encoders learn embedding spaces through a contrastive objective [72].

Chapter 7

Experimental Results

We perform an initial experiment where we systematically vary the impact -*"prompt strength"*- of injected attention maps in audio generation, gradually increasing the influence of textual cues on editing. We calculate the average cosine similarity between original and edited prompts in both Audio-to-Audio and Text-to-Audio contexts. Figure 7.1 shows that the edited audio maintains the qualities of the original, as indicated by the high mean Audio-to-Audio cosine similarity and the mean Text-to-Audio cosine similarity with the source prompt, which remains consistent regardless of the *"prompt strength"*. Additionally, the mean Text-to-Audio cosine similarity with the edited prompt, which remains stable, highlights that the edited audio remains in alignment with the edited prompt regardless of *"prompt strength"*.

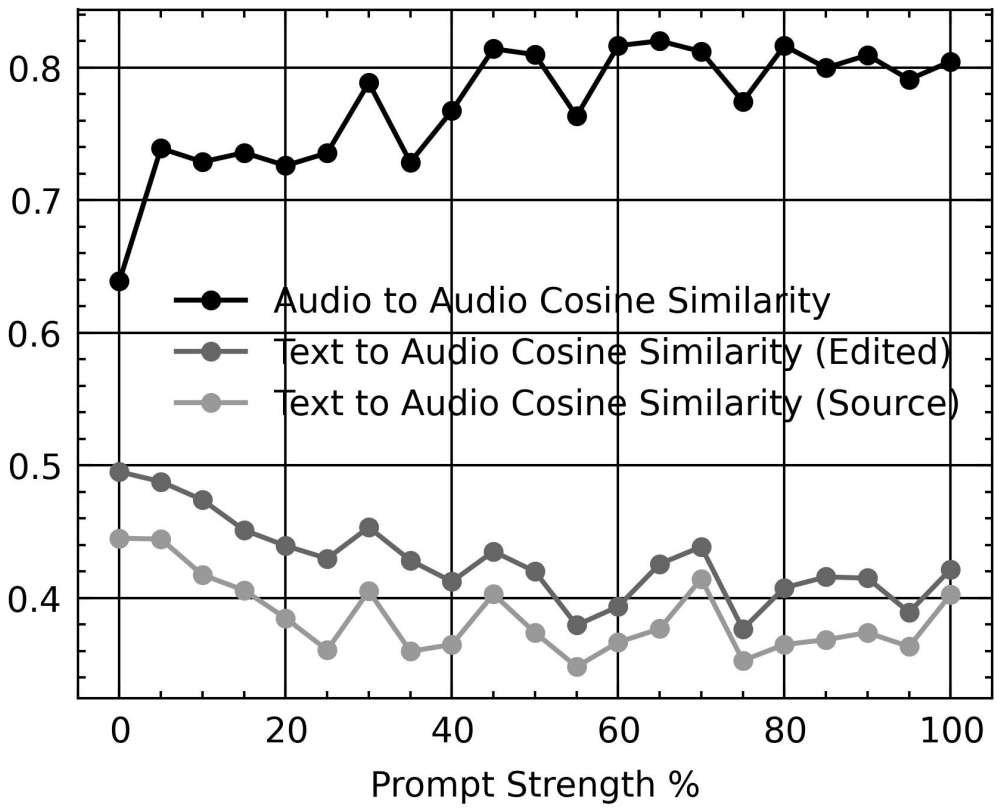


Figure 7.1. *Evaluation of audio and textual alignment with regards to "prompt strength".*

7.1 The effect of soft-blending

Additionally, we evaluate the effectiveness of soft blending cross-attention features quantitatively by computing Audio-to-Audio and Text-to-Audio cosine similarity metrics between the generated audio samples and the edited prompt. Audio-to-Audio and Text-to-Audio cosine similarity scores are averaged across the dataset, ensuring a comprehensive and objective evaluation. As indicated by Table 7.1, using soft-blending leads to higher mean Audio-to-Audio and Text-to-Audio Cosine Similarity scores. Notably, the standard deviation remains relatively the same, highlighting the reliability and consistency of the soft-blending method.

Configuration	T2A Similarity	A2A Similarity
Hard-blending	0.836 ∓ 0.087	0.400 ∓ 0.152
Soft-blending	0.849 ∓ 0.094	0.414 ∓ 0.157

Table 7.1. *Text-to-Audio & Audio-to-Audio Cosine Similarity with regards to blending strategy.*

7.2 Comparison of automated music metrics

Finally, we seek to investigate the effectiveness of Prompt-to-Prompt techniques in audio editing by examining both diffusion-based and auto-regressive models. This exploration aims to offer insights into the strengths and limitations of these models for creative audio manipulation. Our baseline strategy employs Auffusion [10], a diffusion-based approach. This method seamlessly integrates with the Prompt-to-Prompt methodology, providing a solid foundation for exploring prompt-guided audio editing. Additionally, we introduce an alternative avenue by incorporating a pre-trained frozen auto-regressive model, capitalizing on the advanced capabilities offered by the state-of-the-art MUSICGEN [23] model.

The metrics are averaged across the entire dataset and presented in Figure 7.2 while detailed results are showcased on Table 7.2. From Melody Accuracy to Dynamics Correlation, Rhythm F1 score, Audio-to-Audio cosine similarity, and Text-to-Audio cosine similarity, each metric provides a unique perspective on the capabilities of MUSICGEN and Auffusion in handling different audio editing tasks. Based on the provided data, MUSICGEN outperforms Auffusion across all evaluation metrics. It excels in capturing melody accuracy, exhibits superior similarity to both the original audio and target text prompt, and significantly outperforms Auffusion in dynamics correlation and rhythm F1 score. Our methodology, utilizing prompt-to-prompt in the auto-regressive model context for audio editing tasks, outperforms Auffusion with the MUSICGEN model. Notably, this marks the first successful use of prompt-to-prompt in the auto-regressive model audio editing context, highlighting its significance in achieving superior results.

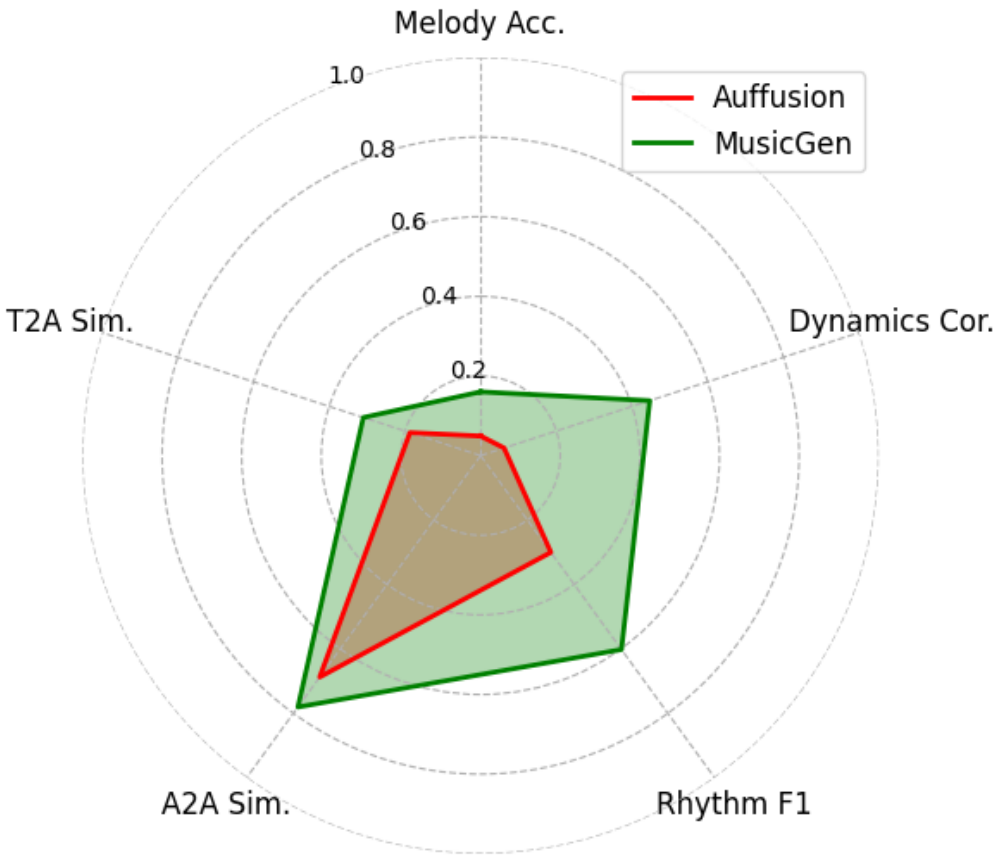
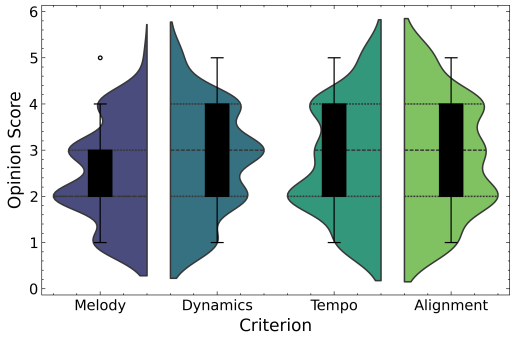
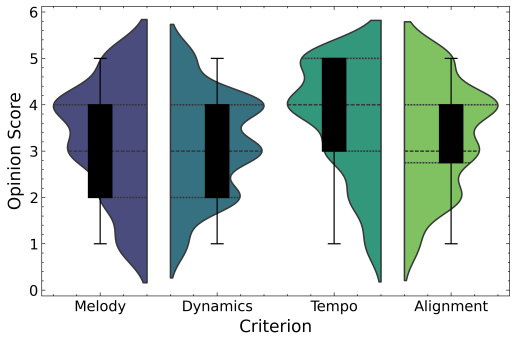


Figure 7.2. *Average evaluation metrics across editing mechanisms for both Auffusion and MUSICGEN.*

7.3 Human study



(a) *Distribution of MOS (Mean opinion score) in the case of Auffusion per evaluation axis.*



(b) *Distribution of MOS (Mean opinion score) in the case of MUSICGEN per evaluation axis.*

Figure 7.3. *Comparison of MOS distributions for Auffusion and MUSICGEN.*

To gauge how well our method preserves naturalness and adheres to the original audio content, we invited 24 evaluators and conducted a user study. The user study began with participants evaluating pairs of 10-second audio clips. Their task was to identify the clip within each pair that exhibited the highest degree of naturalness, characterized by a resemblance to familiar musical instruments as opposed to noise or artifacts. These clips, randomly sampled from our dataset, originated from the same text prompt but were generated by distinct models – Auffusion and MUSICGEN. Next, participants assessed the fidelity of 16 randomly sampled text-audio pairings in our dataset. Each pairing included both the original text prompt and audio, alongside an edited version. Participants assessed the extent to which the edited audio remains faithful to the original version, considering key elements like melody, tempo, and dynamics, as well as its textual alignment to the edited text prompt, rating these characteristics on a Likert [39] scale from 1 to 5. Based on the results of Table 7.3, our method was found to produce audio clips perceived as more natural by participants in our study. Mean Opinion Scores (MOS) for each criterion are depicted in Figure 7.3a and Figure 7.3b. To assess the statistical significance of our results we employ an unpaired t-test for the opinion score distributions. Table 7.4 summarizes the obtained p-values, indicating that the improvement obtained by utiliz-

ing the proposed music editing technique in conjunction with MUSICGEN significantly outperforms the baseline.

Edit	Model	Alignment	Dynamics	Melody	Tempo
Refine	Auffusion	2.76	2.57	2.41	2.81
Refine	MUSICGEN	**3.43**	**3.08**	**3.22**	**3.62**
Replace	Auffusion	2.91	2.80	2.44	2.74
Replace	MUSICGEN	**3.05**	**3.07**	**3.16**	**3.56**
Reweight	Auffusion	2.78	2.88	2.58	2.78
Reweight	MUSICGEN	**3.54**	**3.24**	**3.38**	**3.86**

Table 7.2. *Comparison of audio editing capabilities of MUSICGEN and Auffusion based on MOS (Mean opinion score) of faithfulness.*

Edit	Model	Naturalness
Refine	Auffusion	35.14%
Refine	MUSICGEN	**64.86%**
Replace	Auffusion	40.00%
Replace	MUSICGEN	**60.00%**
Reweight	Auffusion	17.95%
Reweight	MUSICGEN	**82.05%**

Table 7.3. *Comparison of audio editing capabilities of MUSICGEN and **Auffusion based** on MOS (Mean opinion score) of naturalness.*

Melody	Dynamics	Tempo	Alignment
2.66×10^{-12}	2.32×10^{-4}	2.85×10^{-15}	3.76×10^{-06}

Table 7.4. *p-values, obtained using an unpaired t-test, indicating the statistical significance for the opinion scores obtained from the human study.*

Chapter 8

Conclusion & Future Work

In conclusion, we explored using two models for audio editing: Auffusion and MUSICGEN. We began with Auffusion, leveraging its existing capabilities for prompt-based editing. Furthermore, we introduced an alternative approach by incorporating MUSICGEN, a pre-trained auto-regressive model known for its advanced capabilities. To align Prompt-to-Prompt with MUSICGEN's auto-regressive features, we applied the attention injection procedure at all timesteps. Additionally, we introduced soft-blending, a technique that merges the feature maps generated during the forward process with the injected ones, using a weighted average for the output. This novel method aims to enhance sample quality by replicating the original diffusion-based approach of prompt-to-prompt, where edits are applied for a set number of diffusion iterations. In our evaluation, MUSICGEN outperformed Auffusion across all evaluation metrics and editing categories. It showed superior accuracy in capturing melodies, better similarity to both the original audio and target text prompt, and notably exceeded Auffusion in dynamics correlation and rhythm F1 score. This study marks the first successful application of prompt-to-prompt in the auto-regressive model audio editing context. This work was submitted to ISMIR (International Society for Music Information Retrieval) 2024 conference.

Brooks et al. [73] employed Prompt-to-Prompt to generate a large dataset of image editing examples. Subsequently, they trained a conditional diffusion model on this dataset, enabling it to generalize to real images and user-written instructions. Inspired by their approach, a promising future direction entails creating an audio editing dataset with original prompts and corresponding audio, alongside edited prompts and audio, leveraging our proposed methodology. This dataset could then be employed to train a text-guided audio editing model.

We also plan to undertake thorough user studies, specifically focusing on diverse ethnic music and societal contexts. The anticipated edits will vary based on the user's background; for instance, imbuing a soundtrack with traditional music elements depends on the user's interpretation of "traditional." These studies will play a crucial role in assessing the effectiveness and inclusivity of the proposed audio editing techniques across different cultural settings.

We encountered a limitation arising from the cross-attention mechanism within MUSICGEN. To elaborate, within MUSICGEN's decoder, all cross-attention information between text and audio tokens is condensed into a single value. Expanding this mechanism

to provide a clearer depiction of the interaction between text and audio tokens would shed light on how MUSICGEN attends to text instructions and could result in improvements to audio editing quality.

Bibliography

[1] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch και Daniel Cohen-or. *Prompt-to-Prompt Image Editing with Cross-Attention Control*. *The Eleventh International Conference on Learning Representations*, 2023.

[2] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou και Dong Yu. *Diffsound: Discrete Diffusion Model for Text-to-Sound Generation*. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1720–1733, 2023.

[3] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin και Zhou Zhao. *Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models*. *Proceedings of the 40th International Conference on Machine Learning*Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato και Jonathan Scarlett, επιμελητές, τόμος 202 στο *Proceedings of Machine Learning Research*, σελίδες 13916–13932. PMLR, 2023.

[4] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail και Huaming Wang. *CLAP: Learning Audio Concepts from Natural Language Supervision*. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, σελίδες 1–5, 2023.

[5] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang και Mark D Plumbley. *AudioLDM: Text-to-Audio Generation with Latent Diffusion Models*. *Proceedings of the 40th International Conference on Machine Learning*Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato και Jonathan Scarlett, επιμελητές, τόμος 202 στο *Proceedings of Machine Learning Research*, σελίδες 21450–21474. PMLR, 2023.

[6] Haohe Liu, Qiao Tian, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Yuping Wang, Wenwu Wang, Yuxuan Wang και Mark D. Plumbley. *AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining*, 2023. arXiv:2308.05734 [cs, eess].

[7] Po Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze και Christoph Feichtenhofer. *Masked Autoencoders that Listen*. *Advances in Neural Information Processing Systems*S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho και A. Oh, επιμελητές, τόμος 35, σελίδες 28708–28720. Curran Associates, Inc., 2022.

[8] Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian και sheng zhao. *AUDIT: Audio Editing by Following Instructions with Latent Diffusion Models*. *Advances in Neural Information Processing Systems*A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt και S. Levine, επιμελητές, τόμος 36, σελίδες 71340–71357. Curran Associates, Inc., 2023.

[9] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish και Soujanya Poria. *Text-to-Audio Generation using Instruction Guided Latent Diffusion Model*. *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, σελίδα 3590–3598, New York, NY, USA, 2023. Association for Computing Machinery.

[10] Jinlong Xue, Yayue Deng, Yingming Gao και Ya Li. *Auffusion: Leveraging the Power of Diffusion and Large Language Models for Text-to-Audio Generation*, 2024. arXiv:2401.01044 [cs, eess].

[11] Jungil Kong, Jaehyeon Kim και Jaekyoung Bae. *HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis*. *Advances in Neural Information Processing Systems*H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan και H. Lin, επιμελητές, τόμος 33, σελίδες 17022–17033. Curran Associates, Inc., 2020.

[12] Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick και Shlomo Dubnov. *MusicLDM: Enhancing Novelty in text-to-music Generation Using Beat-Synchronous mixup Strategies*. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, σελίδες 1206–1210, 2024.

[13] Bing Han, Junyu Dai, Xuchen Song, Weituo Hao, Xinyan He, Dong Guo, Jitong Chen, Yuxuan Wang και Yanmin Qian. *InstructME: An Instruction Guided Music Edit And Remix Framework with Latent Diffusion Models*, 2023. arXiv:2308.14360 [cs, eess].

[14] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior και Koray Kavukcuoglu. *WaveNet: A Generative Model for Raw Audio*, 2016. arXiv:1609.03499 [cs].

[15] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman και Yossi Adi. *AudioGen: Textually Guided Audio Generation*. *The Eleventh International Conference on Learning Representations*, 2023.

[16] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford και Ilya Sutskever. *Jukebox: A Generative Model for Music*, 2020. arXiv:2005.00341 [cs, eess, stat].

[17] Aaronvan den Oord, Oriol Vinyals και koray kavukcuoglu. *Neural Discrete Representation Learning*. *Advances in Neural Information Processing Systems*I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan και R. Garnett, επιμελητές, τόμος 30. Curran Associates, Inc., 2017.

[18] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi και Neil Zeghidour. *AudioLM: a Language Modeling Approach to Audio Generation*, 2023. arXiv:2209.03143 [cs, eess].

[19] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund και Marco Tagliasacchi. *SoundStream: An End-to-End Neural Audio Codec*. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2022.

[20] Srihari Kankanahalli. *End-To-End Optimized Speech Coding with Deep Neural Networks*. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, σελίδες 2521–2525, 2018.

[21] Darius Petermann, Seungkwon Beack και Minje Kim. *Harp-Net: Hyper-Autoencoded Reconstruction Propagation for Scalable Neural Audio Coding*. *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, σελίδες 316–320, 2021.

[22] Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour και Christian Frank. *MusicLM: Generating Music From Text*, 2023. arXiv:2301.11325 [cs, eess].

[23] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi και Alexandre Defossez. *Simple and Controllable Music Generation*. *Advances in Neural Information Processing Systems*A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt και S. Levine, επιμελητές, τόμος 36, σελίδες 47704–47720. Curran Associates, Inc., 2023.

[24] Alexandre Défossez, Jade Copet, Gabriel Synnaeve και Yossi Adi. *High Fidelity Neural Audio Compression*. *Transactions on Machine Learning Research*, 2023. Featured Certification, Reproducibility Certification.

[25] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch και Daniel Cohen-Or. *NULL-Text Inversion for Editing Real Images Using Guided Diffusion Models*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 6038–6047, 2023.

[26] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein και Kfir Aberman. *DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 22500–22510, 2023.

[27] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman και Jun Yan Zhu. *Multi-Concept Customization of Text-to-Image Diffusion*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 1931–1941, 2023.

[28] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas και Feng Yang. *SVDiff: Compact Parameter Space for Diffusion Fine-Tuning*. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, σελίδες 7323–7334, 2023.

[29] Jason Lee, Kyunghyun Cho και Douwe Kiela. *Countering Language Drift via Visual Grounding*. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*Kentaro Inui, Jing Jiang, Vincent Ng και Xiaojun Wan, επιμελητές, σελίδες 4385–4395, Hong Kong, China, 2019. Association for Computational Linguistics.

[30] Manos Plitsis, Theodoros Kouzelis, Georgios Paraskevopoulos, Vassilis Katsouros και Yannis Panagakis. *Investigating Personalization Methods in Text to Music Generation*. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, σελίδες 1081–1085, 2024.

[31] Hila Manor και Tomer Michaeli. *Zero-Shot Unsupervised and Text-Based Audio Editing Using DDPM Inversion*, 2024. arXiv:2402.10009 [cs, eess].

[32] Inbar Huberman-Spiegelglas, Vladimir Kulikov και Tomer Michaeli. *An Edit Friendly DDPM Noise Space: Inversion and Manipulations*, 2024. arXiv:2304.06140 [cs].

[33] Shih Lun Wu, Chris Donahue, Shinji Watanabe και Nicholas J Bryan. *Music ControlNet: Multiple time-varying controls for music generation*. *arXiv preprint arXiv:2311.07069*, 2023.

[34] M. F. McKinney, D. Moelants, M. E. P. Davies και A. Klapuri. *Evaluation of Audio Beat Tracking and Music Tempo Extraction Algorithms*. *Journal of New Music Research*, 36(1):1–16, 2007.

[35] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis και C Colin Raffel. *MIR_EVAL: A Transparent Implementation of Common MIR Metrics*. *ISMIR*, τόμος 10, σελίδα 2014, 2014.

[36] Florian Krebs, Sebastian Böck και Gerhard Widmer. *An Efficient State-Space Model for Joint Tempo and Meter Tracking*. *ISMIR*, σελίδες 72–78, 2015.

[37] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick και Shlomo Dubnov. *Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation*. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, σελίδες 1–5, 2023.

[38] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick και Shlomo Dubnov. *HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection*. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, σελίδες 646–650, 2022.

[39] Ankur Joshi, Saket Kale, Satish Chandel και D Kumar Pal. *Likert scale: Explored and explained*. *British journal of applied science & technology*, 7(4):396–403, 2015.

[40] David Stutz. *Collection of LaTeX resources and examples*. https://github.com/davidstutz/latex-resources, 2020.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser και Illia Polosukhin. *Attention is all you need*. *Advances in neural information processing systems*, 30, 2017.

[42] Ilya Sutskever, Oriol Vinyals και Quoc V Le. *Sequence to Sequence Learning with Neural Networks*. *Advances in Neural Information Processing Systems*Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence και K.Q. Weinberger, επιμελητές, τόμος 27. Curran Associates, Inc., 2014.

[43] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku και Dustin Tran. *Image Transformer*. *Proceedings of the 35th International Conference on Machine Learning*Jennifer Dy και Andreas Krause, επιμελητές, τόμος 80 στο *Proceedings of Machine Learning Research*, σελίδες 4055–4064. PMLR, 2018.

[44] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov και Sergey Zagoruyko. *End-to-End Object Detection with Transformers*. *Computer Vision – ECCV 2020*Andrea Vedaldi, Horst Bischof, Thomas Brox και Jan Michael Frahm, επιμελητές, σελίδες 213–229, Cham, 2020. Springer International Publishing.

[45] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit και Neil Houlsby. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. *CoRR*, abs/2010.11929, 2020.

[46] Linhao Dong, Shuang Xu και Bo Xu. *Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition*. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, σελίδες 5884–5888, 2018.

[47] Anmol Gulati, James Qin, Chung Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu και Ruoming Pang. *Conformer: Convolution-augmented Transformer for Speech Recognition*. *Proc. Interspeech 2020*, σελίδες 5036–5040, 2020.

[48] Xie Chen, Yu Wu, Zhenghao Wang, Shujie Liu και Jinyu Li. *Developing Real-Time Streaming Transformer Transducer for Speech Recognition on Large-Scale Dataset*. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, σελίδες 5904–5908, 2021.

[49] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas και Alpha A Lee. *Molecular transformer: a model for*

uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.

[50] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma και Rob Fergus. *Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences*. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[51] Jonathan Ho και Tim Salimans. *Classifier-Free Diffusion Guidance*. 2022. Publisher: arXiv Version Number: 1.

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser και Björn Ommer. *High-resolution image synthesis with latent diffusion models*. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, σελίδες 10684–10695, 2022.

[53] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan και Surya Ganguli. *Deep Unsupervised Learning using Nonequilibrium Thermodynamics*. *Proceedings of the 32nd International Conference on Machine Learning*Francis Bach και David Blei, επιμελητές, τόμος 37 στο *Proceedings of Machine Learning Research*, σελίδες 2256–2265, Lille, France, 2015. PMLR.

[54] Prafulla Dhariwal και Alexander Nichol. *Diffusion Models Beat GANs on Image Synthesis*. *Advances in Neural Information Processing Systems*M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang και J. Wortman Vaughan, επιμελητές, τόμος 34, σελίδες 8780–8794. Curran Associates, Inc., 2021.

[55] Jonathan Ho, Ajay Jain και Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. *Advances in Neural Information Processing Systems*H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan και H. Lin, επιμελητές, τόμος 33, σελίδες 6840–6851. Curran Associates, Inc., 2020.

[56] Yang Song και Stefano Ermon. *Generative Modeling by Estimating Gradients of the Data Distribution*. *Advances in Neural Information Processing Systems*H. Wallach, H. Larochelle, A. Beygelzimer, F.d' Alché-Buc, E. Fox και R. Garnett, επιμελητές, τόμος 32. Curran Associates, Inc., 2019.

[57] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon και Ben Poole. *Score-Based Generative Modeling through Stochastic Differential Equations*. 2020. Publisher: arXiv Version Number: 2.

[58] Shitong Luo και Wei Hu. *Score-based point cloud denoising*. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, σελίδες 4583–4592, 2021.

[59] Emmanuel Asiedu Brempong, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer και Mohammad Norouzi. *Denoising pretraining for semantic segmentation*.

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, σελίδες 4175–4186, 2022.

[60] Bahjat Kawar, Gregory Vaksman και Michael Elad. *Stochastic image denoising by sampling from the posterior distribution*. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, σελίδες 1866–1875, 2021.

[61] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow και Rianne Van Den Berg. *Structured denoising diffusion models in discrete state-spaces*. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

[62] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré και Max Welling. *Argmax flows and multinomial diffusion: Learning categorical distributions*. *Advances in Neural Information Processing Systems*, 34:12454–12465, 2021.

[63] Yusuke Tashiro, Jiaming Song, Yang Song και Stefano Ermon. *Csdi: Conditional score-based diffusion models for probabilistic time series imputation*. *Advances in Neural Information Processing Systems*, 34:24804–24816, 2021.

[64] Omri Avrahami, Dani Lischinski και Ohad Fried. *Blended diffusion for text-driven editing of natural images*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, σελίδες 18208–18218, 2022.

[65] Jongmin Yoon, Sung Ju Hwang και Juho Lee. *Adversarial purification with score-based generative models*. *International Conference on Machine Learning*, σελίδες 12062–12072. PMLR, 2021.

[66] Jin Sub Lee, Jisun Kim και Philip M Kim. *Score-based generative modeling for de novo protein design*. *Nature Computational Science*, 3(5):382–392, 2023.

[67] Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng και Jianzhu Ma. *Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures*. *Advances in Neural Information Processing Systems*, 35:9754–9767, 2022.

[68] Prafulla Dhariwal και Alexander Nichol. *Diffusion models beat gans on image synthesis*. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[69] Olaf Ronneberger, Philipp Fischer και Thomas Brox. *U-net: Convolutional networks for biomedical image segmentation*. *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, σελίδες 234–241. Springer, 2015.

[70] Jiaming Song, Chenlin Meng και Stefano Ermon. *Denoising Diffusion Implicit Models*, 2022. arXiv:2010.02502 [cs].

[71] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim

Salimans, Jonathan Ho, David J Fleet και Mohammad Norouzi. *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding*. *Advances in Neural Information Processing Systems*S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho και A. Oh, επιμελητές, τόμος 35, σελίδες 36479–36494. Curran Associates, Inc., 2022.

[72] Aaron van den Oord, Yazhe Li και Oriol Vinyals. *Representation Learning with Contrastive Predictive Coding*, 2019. arXiv:1807.03748 [cs, stat].

[73] Tim Brooks, Aleksander Holynski και Alexei A Efros. *Instructpix2pix: Learning to follow image editing instructions*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, σελίδες 18392–18402, 2023.

List of Abbreviations

ANN	Artificial Neural Network
DPMs	Diffusion Probabilistic Models
SLP	Single-Layer Perceptron Network
IS	Inception Score
FID	Fréchet Inception Distance
DPM	Diffusion Probabilistic Model
RVQ	Residual Vector Quantization
T2A	Text-to-Audio
A2A	Audio-to-Audio
VQ-VAE	Vector Quantised-Variational AutoEncoder
CLAP	Contrastive Language-Audio Pretraining
LDM	Latent Diffusion Model
LOA	Language of Audio
LLM	Large Language Model
DDIM	Denoising Diffusion Implicit Model
DDPM	Denoising Diffusion Probabilistic Model
PC	Principal Component
BAM	Beat-Synchronous Audio Mixup
BLM	Beat-Synchronous Latent Mixup
MOS	Mean Opinion Score
ISMIR	International Society for Music Information Retrieval